universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Prediction of Competitive CYP450 2C19 Inhibitors Using 3D Pharmacophore Modeling“

verfasst von / submitted by

Dominik Labuz

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Pharmazie (Mag.pharm.)

Wien, 2019 / Vienna, 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 449
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Diplomstudium Pharmazie
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thierry Langer

Danksagung

Mein größter Dank gilt Prof. Thierry Langer, der mir es überhaupt ermöglicht hat, meine Diplomarbeit in seiner Arbeitsgruppe am Department für pharmazeutische Chemie zu verfassen. Trotz seines dichten Terminkalenders wurden Fragen meinerseits immer sofort beantwortet und zusätzlicher Input durch sein umfassendes Wissen zu meiner Arbeit bereitgestellt.

Danke natürlich auch der Arbeitsgruppe um Dr. Thomas Seidel, die mir einen Arbeitsplatz zur Verfügung gestellt hat und bei technischen, wie auch inhaltlichen Problemen zur Seite stand.

Einen besonderen Dank möchte ich auch meinem Studienkollegen Martin Pramesberger aussprechen, der mir vor allem am Anfang meiner wissenschaftlichen Arbeit aufgrund seiner bereits vorhandenen Expertise in diesem Gebiet mit vielen hilfreichen Tipps den Einstieg erleichtert hat.

Darüber hinaus muss ich mich von ganzem Herzen bei meinen Eltern bedanken. Sie haben mich in erster Linie dazu gebracht Pharmazie zu studieren und haben mich das ganze Studium über bis zum Ende hin mit bestärkenden Worten und Ratschlägen unterstützt.

Und nicht zuletzt muss ich auch meinen Freunden einen großen Dank aussprechen. Auch wenn sie mich vom Lernen meistens eher abgehalten als dazu motiviert haben, wurde meine Studentenzeit durch sie zu einem unvergesslichen Lebensabschnitt, an den ich mich mein Leben lang gerne zurückerinnern werde.

Danke!

Zusammenfassung

Die Cytochrom P450 Enzyme mit ihrer einzigartigen prosthetischen Häm Gruppe in deren aktivem Zentrum gehören zu den wichtigsten Abwehrsystemen im menschlichen Organismus. Sie sind für den Phase I Metabolismus von Arzneistoffen und für die Biosynthese von endogenen Stoffen verantwortlich. Leider können manche Wirkstoffe nicht nur als Substrat dieser Enzyme agieren, sondern auch deren Fähigkeit Fremdstoffe zu metabolisieren hemmen. Wenn dann ein zweiter Arzneistoff verabreicht wird und um dasselbe verstoffwechselnde Isoenzym konkurriert, können gefährliche Arzneistoff-Wechselwirkungen mit erhöhten Plasmaspiegeln aufgrund einer Akkumulation auftreten. Das Ziel der vorliegenden Arbeit war es Pharmakophormodelle zu erstellen, die die Vorhersage von kompetitiven CYP450 2C19 Inhibitoren ermöglichen. Trainingsdatensätze von Inhibitoren und Nicht-Inhibitoren wurden aus der ChEMBL Datenbank entnommen und mit Hilfe des Data-Mining Programms Knime (Version 3.7.2) verarbeitet. Die Software LigandScout (Version 4.4) ermöglichte die Generierung von ligandenbasierten Pharmakophormodellen aus den Trainingsdatensätzen und strukturbasierten Modellen mit der 3D-Information des Makromoleküls aus der Protein Data Bank (PDB). Die folgenden virtuellen Screenings und die Verbesserungen der Modelle wurden ebenfalls mit LigandScout durchgeführt. Das Ergebnis dieser wissenschaftlichen Arbeit sind 13 Pharmakophormodelle, die für die Vorhersage von CYP450 2C19 Inhibitoren verwendet werden können. Moleküle, die von diesen Modellen identifiziert werden, könnten in Zukunft sofort aussortiert werden, bevor sie weitere Entwicklungsschritte durchlaufen.

Abstract

The cytochrome P450 enzymes with their unique heme prosthetic group in their active site belong to the most important defense systems in the human organism. They are responsible for phase I metabolism of drugs and for the biosynthesis of endogenous compounds. Unfortunately, some drugs do not act only as substrates of these enzymes, but they may also inhibit their ability to dispose xenobiotics. When a second therapeutic agent is co-administered, and it competes for the same metabolizing isoenzyme, dangerous drug-drug interactions with elevated plasma levels due to accumulation may occur. The goal of this thesis was to create pharmacophore models that are able to predict reversible inhibitors of the isoenzyme CYP450 2C19. Training datasets of inhibitors and non-inhibitors were extracted from the ChEMBL database and processed with the help of the data mining program Knime (version 3.7.2). The software LigandScout (version 4.4) allowed to create ligand-based pharmacophore models out of the training dataset and structure-based models using the 3D information of the macromolecule derived from the Protein Data Bank (PDB). The following virtual screening experiments and model refinement was also performed in LigandScout. The result of this scientific work are 13 pharmacophore models, which can be used to predict potential CYP450 2C19 inhibitors. Molecules identified by these models can be sorted out instantly from a drug discovery project, before going through further development steps.

Table of contents

1. Introduction

1.1 Incentive

Our organism lives in an environment where a daily exposure to a vast amount of chemicals like drugs, intermediates, or pollutants is common. To be able to protect the body against toxicological effects of these compounds, evolution has emerged with protection mechanisms: The cytochrome P450 enzymes.[1] They play a remarkable role in the disposal of xenobiotics and the metabolization of drugs, but are also responsible for the production of endogenous substances, like bile- and arachidonic acids or steroids. Up to date more than 50 CYP450 enzymes have been characterized, though only six of them metabolize approximately 90% of available drugs.[2] Unfortunately, drugs can interact with the CYP450 system by either inducing or inhibiting their metabolism. In case of inhibition therapeutic doses may cause adverse side effects due to elevated plasma concentration, when a CYP450 inhibitor is added as second drug to the medication plan, on the other hand pro-drugs might not be converted into their active metabolites, so there will be no pharmacological effect at all. Because of this, many pharmaceutical products were withdrawn from the market, mainly due to too late discovered dangerous drug-drug interactions.[1,2] As population becomes older and older, polypharmacy is needed to treat the increasing number of diseases, so the risk of cytochrome P450 related drug-drug interactions through enzyme inhibition is growing and becoming an important therapeutic concern.[3]

1.2 Aim

The aim of this work was to create pharmacophore models that can be used to predict potential inhibitors of the enzyme cytochrome P450 2C19. Developing a new drug is an expensive and long-lasting process.[4] Many candidates fail during clinical trials due to unwanted drug-drug interactions and therefore cannot be brought to the market. In order to prevent such setbacks, CYP2C19 pharmacophore models may be either used in the very early stage of drug discovery to identify compounds that may cause interactions with this enzyme, or in the optimization of the chemical structure of a promising molecule. Thus, pharmaceutical companies might save money and time spent in research and development, because potential

CYP2C19 inhibitors hit by the pharmacophore models may be sorted out of the pipeline instantly or might be redesigned and accordingly modified before undergoing further development.[5,6]

2. Biological Background

2.1 An Introduction to The Cytochrome P450s

CYP450s were first described almost 60 years ago in spectrophotometric experiments with rat liver cells, where a specific absorption band at 450nm was measured. The pigment responsible for the absorption at 450nm was heme in his reduced form, bound to carbon monoxide.[7] Nowadays it is known that cytochrome P450s exist in prokaryotes and eukaryotes such as bacteria, plants and different animals.[8] In mammals, they are present as membrane-bound microsomal or as mitochondrial heme proteins with a coordination complex consisting of an iron-porphyrin ring as the prosthetic group, spread over all tissue in human body, but mainly occurring in liver and small intestine.[1]

2.1.1 Nomenclature

According to the current scientific knowledge, there are 57 CYP genes in our human body, divided into 18 families. CYP Proteins have been in the focus of researchers for a long time, resulting in the identification of more and more CYP genes, hence more and more different CYP enzymes.[9] For classification purposes, in 1987 Nebert et al. was the first to recommend a nomenclature for the P450s, which was steadily updated to the nomenclature system that is currently still in use. The letters “CYP” describe the gene, followed by an arabic numeral (e.g. CYP3), expressing the family number. Another letter represents the subfamily (e.g. CYP3A) and then again, a number designating the individual gene/isoenzyme (e.g. CYP3A4).[10,11] If the primary amino acid sequence within an enzyme family displays a similarity of over 40%, it counts as a family. In addition, if there is a resemblance of more than 55% within the family, it belongs to a subfamily (e.g. CYP3A4, CYP3A5).[8]

2.1.2 Occurrence and Structural Features

Human CYP450s are either membrane-bound proteins that can be found in all types of tissue in human body, mainly in hepatic and intestine cells, or proteins that are present in the inner membranes of mitochondria of cells involved in steroidogenesis, for instance adrenal cortex or ovaries. Depending on the location, this enzyme group appears in variable quantity and perform different tasks, such as drug metabolization, cholesterol biosynthesis and chemical transformation of endogenous substances.[1,7,18] **Figure 1** shows some localizations and substrates metabolized by diverse human cytochromes P450s.

CYP	Localization	Typical substrate
1A1	lung, liver,brain, GIT, lymphocytes, heart	PAH (polycycl. arom.hydrocarbons)
1A2	liver	aromatic amines, PAH, caffeine
1B1	skin, brain, heart, lung, placenta, liver, kidney, GIT, spleen	PAH
2A6	liver	coumarin, steroids
2B1/2	brain	morphine
2B6	liver, heart	nicotine
2C8	liver, kidney	retinoids, taxol
2C9/10	liver	tolbutamide, diclofenac
2C19	liver, heart	(S)-mephenytoin, omeprazole, diazepam
2D6	liver, brain, heart	antidepressives, β-blockers
2E1	liver, lung, brain, endothelium, heart, bone marrow	ethanol, nitrosamines, acetaminophen
2F	lung	coumarins
3A4/5	liver, GIT, kidney, lung, brain, endothelium, placenta, lymphocytes	different – Ca channel blockers, cyclosporin, acetaminophen, taxol, steroids
3A7	fetus, placenta, (liver)	various, similar to 3A4
4A9/11	kidney	fatty acids
4B1	lung, placenta,	?
4F2/3	kidney	arachidonic acid derivatives

GIT, gastrointestinal tract; PAH, polycyclic aromatic hydrocarbons.

Figure 1: Localization, Substrates and Inductors of CYP450s[8]

Regarding the structure, there are some common properties that every eukaryotic CYP450 shares with each other. They consist of twelve alpha helices marked by the letters A – L and four beta sheets containing all in all between 400 and 500 amino acids. Beside of that, there is a central heme prosthetic group, i.e. an oxo-iron protoporphyrin cation bound to the protein through a thiolate ligation of a cysteine, giving the CYP450s their unique spectral absorption band at 450nm. This moiety acts also as the catalytic core of the enzyme, where the transfer of protons, electrons and oxygen activation takes place. That active site is surrounded by helices C, D and I-L together with beta sheets one and two and represents the most conserved

structural part of the enzymes. The outer regions, consisting of Helix B, C, F, G and the N-and C- terminal groups are flexible and can form in dependence of the bound substrate different protein structures.[12,13]

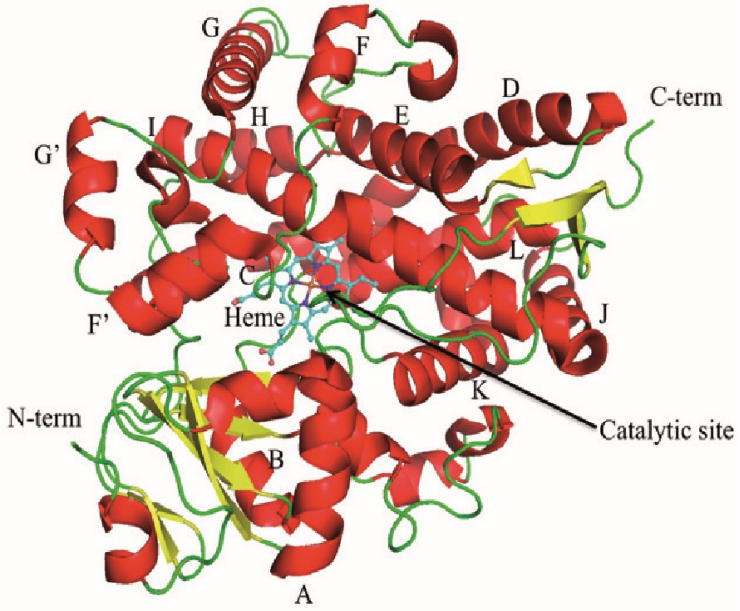


Figure 2: Structure of cytochrome P450 (CYP3A4 – PDB code 1TQN). Alpha helices are shown in red, Beta sheets are yellow and loops in green. Furthermore, there is the N-terminus, C-terminus and the catalytic site with the heme prosthetic group in the central cavity (carbon, oxygen, nitrogen and iron atoms in cyan, red, blue and orange).[14]

2.1.3 The Catalytic Mechanism

During phase one metabolism, P450s perform various oxidation or reduction mechanisms at physiological environment that normally require a higher temperature. This is possible, because chemical reactions, like hydroxylation, dealkylation or oxidation are catalysed by their active site. The cytochrome P450 are considered as oxygenases, to be more exact, monooxygenases. Their name results of their utilization of molecular oxygen, by inserting it into a substrate and therefore creating a water molecule by reducing its second oxygen with the help of two electrons. The process is called “the catalytic cycle” (**Figure 3**).[13,15]

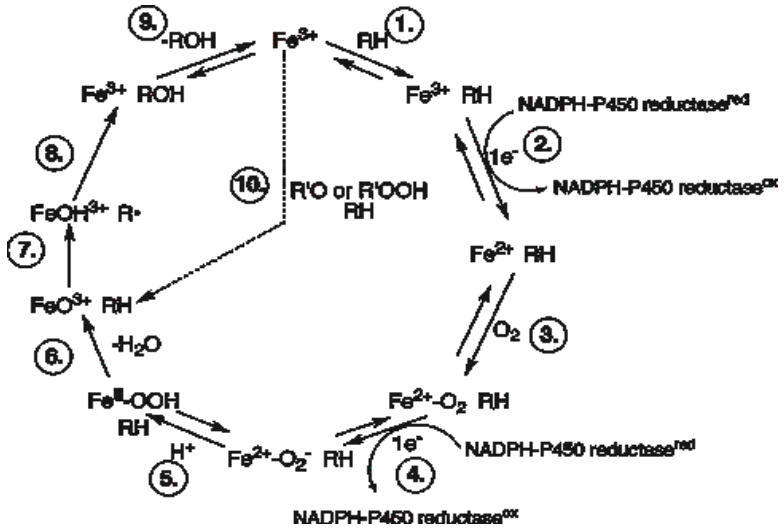


Figure 3: Steps of the CYP450 catalytic cycle.[16]

The cycle starts with a substrate binding close to the active site of the enzyme, causing a change of conformation, altering the redox potential and displacing a water molecule (1). NADPH-P450 reductase transfers one electron and reduces the active centre into a ferrous state ($Fe^{3+} \rightarrow Fe^{2+}$) (2). After this step, molecular oxygen comes into play and binds to the Fe^{2+} ion giving a superoxide complex (3). Another electron is added to the cycle by NADPH-P450 reductase or sometimes by cytochrome b_5, forming a peroxo group (4). After that, two protons attack the dioxygen bond, breaking it, resulting in a water molecule, which leaves a reactive FeO^{3+} product behind (5,6). The oxidation chemistry takes place now. FeO^{3+}, also called "perferryl oxygen complex", reacts with the substrate hydroxylating it (7,8). In the end the product is being released and the active site of the protein returns to its initial position with a water molecule taking its former place in the active site (9).

This is an example of a hydroxylation reaction. The reaction mechanism depends on the type of the CYP enzyme and on the substrate that is involved in the catalytic cycle. Reduction, dealkylation or carbon-carbon bond splitting are also possible, as CYP450s enzymes catalyse a wide range of chemical reactions.[13, 15-17]

2.2 Role of CYP450s in Medication

2.2.1 An Introduction into Drug Metabolism

Every therapeutic agent that is taken to prevent, to treat or to cure diseases share the same fate in the human body by undergoing a range of chemical modifications (for e.g. N- and O-dealkylation, aliphatic- and aromatic hydroxylation, N- and S-oxidation). A major responsibility for this process bears the cytochrome P450 system in the liver and in the gastrointestinal tract.[18] The main reason why these compounds enter the organism is related to their lipophilicity. This property allows them to be absorbed through for example the skin, the lungs or the gastrointestinal tract and to unfold their pharmacological effect. But on the other hand, it is the lipophilicity that makes foreign substances so hard to eliminate. They get reabsorbed, so the clearance slows down, leading to an accumulation with adverse side effects. Thus, lipophilic drugs require chemical modifications, also called "biotransformation" in order to be eliminated later on. Basically, this process can be divided into two steps (**Figure 4**): phase I and phase II, whereas phase I is not necessary, if the substrate has already hydrophilic groups. In phase I a hydrophilic functional group (-OH, -COOH, NH_2,) is inserted e.g. by oxidation in the active site of cytochromes P450s to the molecule making it more hydrophilic. In phase 2 the hydrophilic group is used for conjugation with very hydrophilic agents, like glucuronic acid, sulfates, amino acids or glutathione, making the molecule even much more polar. Enzymes involved in phase II are called transferases for example UDP-glucuronosyltransferases (UGTs), sulfotransferases (SULTs), glutathione S-transferases (GSTs) and methyltransferases. This step promotes the excretion of polar products through urine or bile.[19]

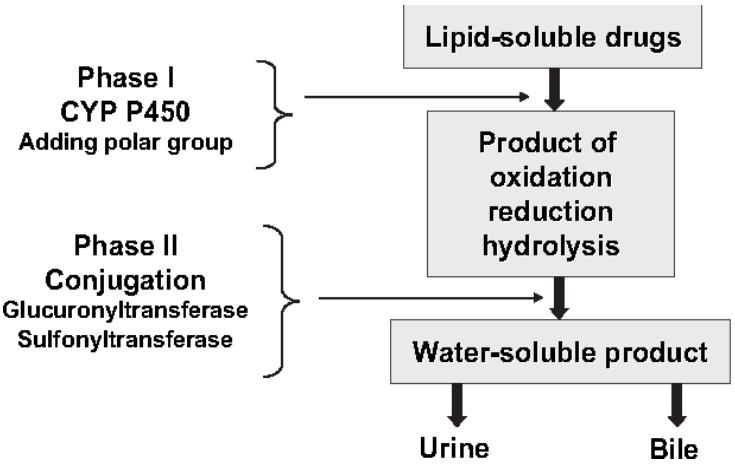


Figure 4: Biotransformation of lipid-soluble drugs, divided into phase I and phase II[20]

2.2.2 CYP Polymorphism

Although 18 cytochrome P450 families are identified today, only 3 of them (CYP1 – CYP3) play a role in the biotransformation of pharmaceutical drugs, hence are of clinical importance. Included are the isoenzymes CYP 1A2, 2B6, 2C9, 2C19, 2D6 and 3A4, which metabolize around 90% of all clinically relevant drugs. Even more, most of the compounds are not disposed by all these enzymes together, but only by one or a few of them.[21] This might have a great clinical impact, if the responsible enzymes are influenced by genetic polymorphism and expressed with a too high or very low catalytic activity (**Figure 5**). It changes the drugs plasma concentration or the plasma level of its active or toxic metabolites, causing adverse side effects or no therapeutic effects. One example is the conversion of morphine from codeine, where poor metabolizers with a CYP2D6 polymorphism will have no analgesic effect, while extensive biotransformation can lead to an opioid intoxication.[22] Depending on the metabolic capability of the CYP450s, patients are divided into four groups: poor, intermediate, extensive and ultra-rapid metabolizers, thus poor being the ones with the lowest enzyme- and ultra-rapid the ones with the highest enzyme activity.[23]

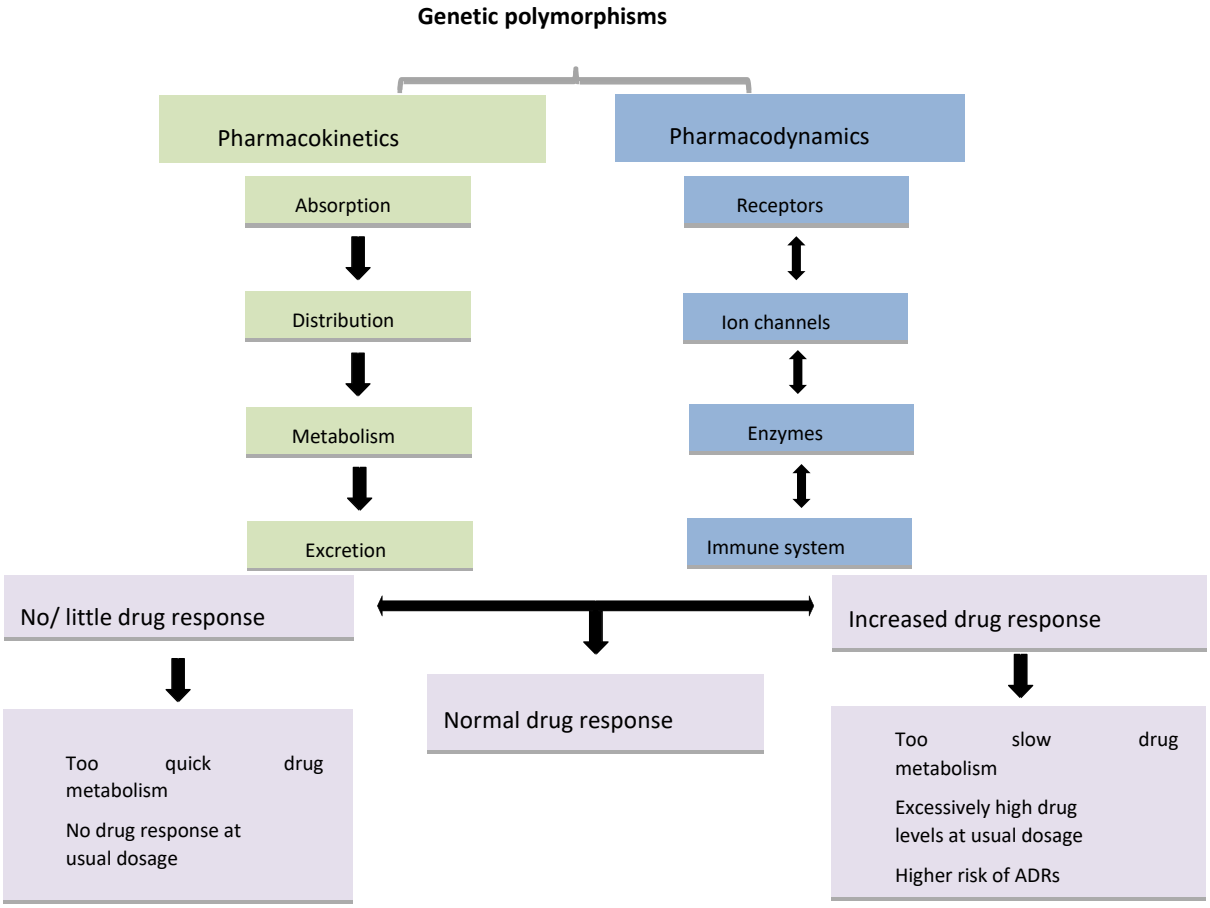


Figure 5: Genetic polymorphisms and its consequence on individual drug response.[23]

Conclusively, polymorphism is an important factor in the individual response to therapeutic agents due to different regulation of metabolizing enzymes by their encoding genes. This genetic variability pertains either a whole population, like Caucasians, where 5-10% are affected by CYP 2D6 polymorphism and have a lack of activity or just some individuals in not affected populations.[4]

2.2.3 CYP Inhibition

Many pharmaceutical compounds do not only act as a substrate of diverse cytochromes P450s, but in addition may also change the activity of these enzymes by induction or inhibition. As long as only one drug is taken with no other CYP substrates, CYP inhibition might occur, but simply without interaction, thus without any consequence. Unfortunately, nowadays polypharmacy is common. When two or more drugs are administered at the same time or time-period and they compete for the same metabolizing enzyme, CYP inhibition can lead to dangerous drug-drug interaction with elevated plasma levels due to accumulation. Adverse side effects occur, and in dependence of the therapeutic window they may be of serious nature.[24]

In general, the term inhibition refers to an enzyme, whose activity is decreased due to interaction with an enzyme inhibitor. The time of inhibition often correlates with the half-life of the molecules. Inhibition mechanisms can be divided into two groups (Pelkonen et al. 2008[26]): reversible or irreversible (mechanism-based inhibition), though earlier published papers (Lin et al. 1998[24], E.Tanaka. 1998[25]) suggest three categories: reversible-, quasi-irreversible- and irreversible inhibition. Because distinguishing between them is quite hard, these mechanisms are summed up as reversible and irreversible in this work. Furthermore, reversible inhibition can be split into the competitive type, the non-competitive, the uncompetitive and the mixed-type.[25-27] Going into detail, a reversible, competitive inhibition takes place, when drugs compete for the binding site in the catalytic centre of the enzyme. It is only dose-dependent and stops as soon as the inhibitor is excreted. Drug interactions are likely to result because of this type of inhibition. In case of non-competitive inhibition, the molecule responsible for the inhibition does not bind to the active site of the enzyme but to another site, leading to a decreased catalytical activity. In the event of the uncompetitive type, the inhibitor forms a complex with enzyme and substrate. The consequence is an

unproductive enzyme-substrate-inhibitor complex.[27] Last, but not least mixed type inhibition shows similarities between the competitive and the non-competitive type of inhibition, making it hard for classification. On the other side irreversible inhibition means a complete inactivation (suicide inhibition) of the enzyme by a covalent bound to the heme centre or by forming a complex with the binding site of the protein. This happens through chemical transformation of drugs into reactive intermediates during one of the steps of the catalytic cycle, therefore it is called mechanism-based inhibition. The result is a long-lasting, time depending blockade and the need of a re-synthesis of the CYP450.[26]

2.2.4 CYP Induction

Induction of drug metabolism enzymes during medication may have tremendous effects on the therapy results. On the one side, it leads to an accelerated elimination of pharmaceutical compounds, which ends up either in the loss or in in the absence of the therapeutic effect. On the other side the conversion to possibly toxic intermediates is sped up, so the balance between the formation of these metabolites and the excretion is of major relevance.[24] The event of induction is often considered as an “double-edged sword”, because it enables the organism to detoxicate xenobiotics at a higher rate and hence to defend itself more effective, but as stated above, it influences the drug therapy and may even activate intermediates acting as procarcinogens. The term induction refers to the process, where a molecule induces the activity of one or more enzymes by increasing their intracellular concentration. The reason for that is a stimulated synthesis of new protein due to higher gene expression or the slowed rate of degradation. In case of cytochrome P450 induction of new synthesis is much more common than a decreased rate of degradation. Induction is dose-dependent, the higher the dose of the inducer, the more transcriptional activity is shown.[27]

From a molecular point of view, the mechanism of induction is regulated by transcription factors, which furthermore are activated by ligands. Intracellular receptors, like aryl hydrogen receptor (AhR), pregnane X receptor (PXR) and constitutive androstane receptor (CAR) play a role as chemical sensors, bind specific ligands and mediate the transcription of genes, that are involved in drug metabolism and disposal.[26]

2.3 Cytochrome P450 2C19

2.3.1 An Introduction into The Significance of CYP450 2C19

CYP 2C19 contributes to the metabolization of approximately 7% of clinically used drugs. The main reason why this enzyme became a focus of research is its polymorphism, which was discovered in the early 1980s and is considered as one of the most important clinical polymorphisms.[21] CYP 2C19 is encoded by genes located on chromosome 10, contains 490 amino acids and is expressed mainly in the liver along with CYP 2C8 and 2C9, but shows also activity in the gut wall. These enzymes are also called monooxygenases.[28,29]

2.3.2 CYP 450 2C19 Structural features And Active Site Characteristics

While the crystal structures of CYP2C8 and CYP2C9 were already described in 2003 and 2004, it took eight more years to identify the structural characteristics of CYP 2C19.[28,30,31] The determining method was X-ray crystallography giving insight into the atomic architecture of CYP2C19 complexed with a chemical inhibitor called compound "0XV" (**Figure 6**). Similarly, to other cytochrome P450s, CYP 2C19 also displays twelve alpha helices and four beta sheets (although **Figure 7** depicts only three beta sheets). Furthermore, the structure shows two internal cavities: One is the substrate binding site above the heme, where also 0XV resides, and the second is an antechamber, probably serving as a substrate access channel. In addition, a solvent channel (black arrow in **Figure 7**) appears between helix F' and the N-terminal loop, close to helix A.[28]

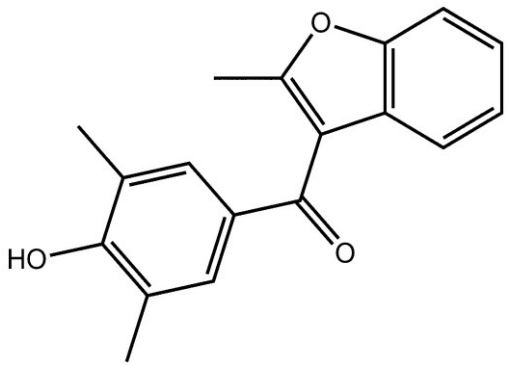


Figure 6: Inhibitor 0XV (2-methyl-1-benzofuran-3-yl)-(4-hydroxy-3,5-dimethylphenyl)[28]

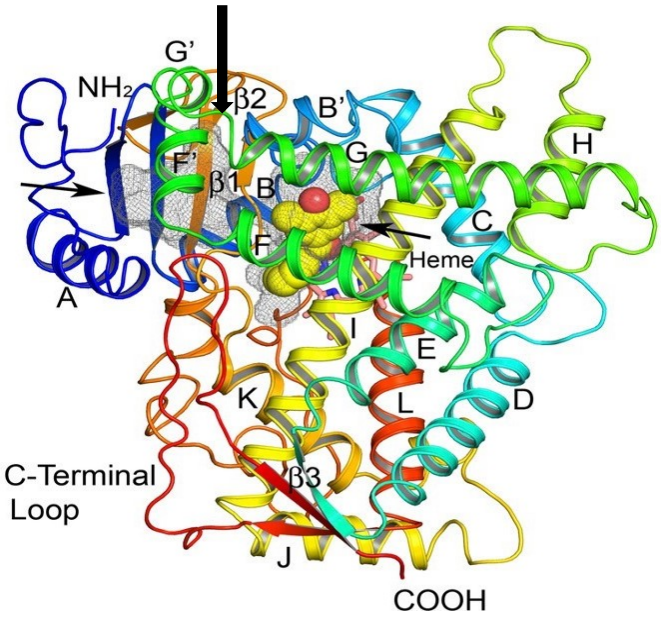


Figure 7: Two internal cavities shown as mesh surfaces. Within the central one inhibitor 0XV is displayed as yellow atomic spheres with the red one occupying the active site of the enzyme.[28]

Taking a closer look at the complex of the active site of CYP 2C19 with the inhibitor 0XV, there are some given interactions between these structures. A water molecule forms a hydrogen bond with the hydroxyl group of phenyl, whereas a second H_2O molecule, bound to the heme iron, is connected to the carbonyl moiety of 0XV and the carbonyl of Ala-297 through hydrogen bonds. Moreover, aromatic and hydrophobic interactions are observed between benzofuran and some amino acid residues: Phe-476 is likely to interact through aromatic effects with benzofuran, the methyl group on Thr-301 shows hydrophobic interactions, just as Cβ and Cγ of Glu-300, Ile-205 and Ile-362 on helix I. The methyl group of benzofuran forms a contact to Leu-366. Continuing with the phenolic moiety of 0XV, one of the methyl groups is exposed to Val-208, while the other interacts with Asp-293, Phe-114 however forms an aromatic interaction with the aromatic ring.[28] It was already demonstrated before, that Phe-114 and Phe-476 play a pivotal role in the binding affinity of substrates to CYP2C9 due to aromatic interactions, which seems to be the same for CYP 2C19.[32]

Beside of this, earlier research projects that used homology modeling, pharmacophore models, CoMFA, and docking studies have implicated that lipophilicity is the most important factor for CYP2C19 inhibition, since polar functional groups, except a carbonyl or sulphoxid

moiety between the hydrophobic groups, show a rather detrimental effect on the affinity. This assumption was later clearly proven by analysis of the crystal structure of CYP450 2C19 complexed with its inhibitor 0XV.[28,33,34]

2.3.3 CYP450 2C19 Polymorphism

CYP450 2C19 catalyses the oxidation of many clinically used drugs, for example anxiolytics, proton pump inhibitors, and anticonvulsants. More than 30 alleles are identified today, though only CYP2C19*2 and CYP2C19*3 being responsible for poor metabolism in 2%-6% of Caucasian and 15%-30% of Asian population.[20,23] People with the CYP 2C19*17, on the other hand, are considered as ultra-rapid metabolizer, mainly found in 18%-27% of Europeans and 10%-26% of Africans.[35] To find out whether a patient is a poor- or fast metabolizer, (S)-mephenytoin is administered, and later on, the clearance of its 4-hydroxy metabolite is measured.[29]

Polymorphism has a big clinical impact on the pharmacokinetic of some drugs, like omeprazole and clopidogrel. While poor metabolizer benefit from the higher AUC values of omeprazole and increased cure rates in the eradication of Helicobacter pylori, they lack the ability to transform clopidogrel into his active metabolite, which may result in cardiac disorders.[29,35] For fast metabolizer it is the other way around. They have decreased plasma concentration of omeprazole, which might lead to therapeutic failure. However, to confirm this, more studies need to be done with patients with the gain-of-function variant of CYP2C19, because in contrast to loss-of-function alleles, data concerning this issue is rather less consistent. In case of clopidogrel, there is a higher bleeding risk as an adverse side effect due to increased plasma levels of clopidogrel's active metabolite, but it results also in less cardiovascular events. Therefore, it is important to screen especially the CYP2C19 enzyme activity in patients who undergo therapy with CYP2C19 substrates to increase therapy efficacy and safety.[36,37]

All in all, it is hard to decide if drug and dose are appropriate when CYP2C19 polymorphism occurs. To solve this concern, the clinical Pharmacogenetics Implementation consortium of the Pharmacogenomics Research Network was established back in 2009. This association has set the goal to create drug/gene guidelines for personalized medicine. They shall help clinicians to translate genotype information from patients into safe and effective clinical decisions in order to optimize the pharmacotherapy with adequate drug and dose.[38]

2.3.4 CYP 450 2C19 Substrates, Inducers and Inhibitors

CYP 2C19 Substrates	CYP 2C19 Inhibitors	CYP 2C19 Inducers
esomeprazole	esomeprazole	carbamazepine
lansoprazole	lansoprazole	efavirenz
omeprazole	omeprazole	enzalutamidenorethindrone
pantoprazole	pantoprazole	pentobarbital
diazepam	chloramphenicol	prednisone
phenytoin	cimetidine	rifampicin
S-mephenytoin	felbamate	ritonavir
phenobarbitone	fluoxetine	St. John's Wort
amitriptyline	fluvoxamine	
carisoprodol	indomethacin	
citalopram	isoniazid	
chloramphenicol	ketoconazole	
clomipramine	modafinil	
clopidogrel	oral contraceptives	
cyclophosphamide	oxcarbazepine	
hexobarbital	probenicid	
imipramine	ticlopidine	
indomethacin	topiramate	
labetalol	voriconazole	
R-mephobarbital		
moclobemide		
nelfinavir		
nilutamide		
primidone		
progesterone		
proguanil		
propranolol		
teniposide		
R-warfarin→8-OH		
voriconazole		

Figure 8: Table with substrates, inhibitors and inducers of cytochrome P450 2C19[39]

In general, CYP450 2C19 has a selectivity for neutral or basic and moderately lipophilic compounds, which can also form hydrogen bonds.[8] As shown in **Figure 8**, 2C19 contributes to the metabolism of many drugs like proton pump inhibitors, several antidepressants, anticonvulsants, anti-infectives, and drugs with effect on the cardiovascular system that are used in daily clinical routine. Some of them show a wide therapeutic window for e.g. PPI's, hence interaction through CYP inhibition or polymorphism may not influence the safety. Others have narrow therapeutic ranges like antidepressants or antiepileptics, causing adverse events when being administered with an inhibitor. Furthermore, there are prodrugs like

clopidogrel, which need to be converted into an active metabolite to have a pharmacological effect at all.[29]

3. Computational Background

3.1 3D Pharmacophore Modeling

Pharmacophore modeling is a technique applied in computational chemistry and has become a widely used tool for rational molecular design. It is used in the early stages of drug discovery and development for prediction of drug properties in order to minimize costs and future risks. Going back in history, Paul Ehrlich in the early 1900s has often been cited to be the first one to introduce the concept of pharmacophore, defined as molecules with certain chemical features, which are truly essential for the drugs biological activity and that molecules with similar features are able to carry out the same biological effect.[40] However, according to J. H. Van Drie (2007), it is a mistake that Paul Ehrlich was given all the credit for the invention of pharmacophores, because this term is nowhere stated in Ehrlich's original articles. The first person, who reported this concept was Monty Kier from the Battele Institute in Ohio, who published papers between 1967 and 1971 about pharmacophores.[41]

However, in 1998, the International Union of Pure and Applied Chemistry (IUPAC) established a new definition of the term pharmacophore, reads as follows:

"*A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response. A pharmacophore does not represent a real molecule or a real association of functional groups, but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds towards their target structure. The pharmacophore can be considered as the largest common denominator shared by a set of active molecules*. *This definition discards a misuse often found in the medicinal chemistry literature which consists of naming as pharmacophores simple chemical functionalities such as guanidines, sulfonamides or dihydroimidazoles (formerly imidazolines), or typical structural skeletons such as flavones, phenothiazines, prostaglandins or steroids.*"[42]

A pharmacophore model is only of high quality if it shows all the necessary types of functional groups of a molecule that are responsible for the ligand-target interaction, summed up as features like hydrogen-bond donors, hydrogen-bond acceptors, hydrophobic, aromatic, positively and negatively charged groups (**Figure 9**). To reflect the real nature of these bonds, the model is displayed in a three-dimensional way, which also describes the probable positions of the chemical interactions in space with respect to steric arrangements. This is important to ensure a high predictive power, when it comes to application of pharmacophore models, for instance in virtual screening of databases, in the discovery of new compounds, in profiling, or in the simple understanding of interactions with receptor binding sites. In 3D pharmacophore modeling, there are two different approaches: ligand-based- and structure-based modeling.[43-45]

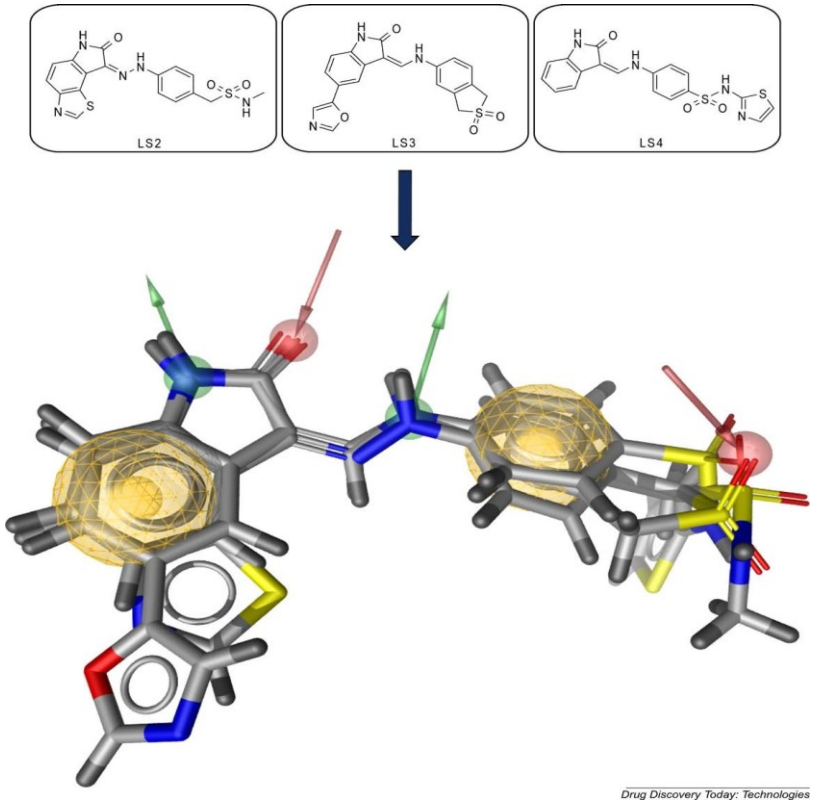


Figure 9: An example of a pharmacophore model derived from 3 ligands in their binding conformation. Yellow spheres describe aromatic interactions, while green arrows show hydrogen bond donors and the red arrows the acceptors. Blue stands for nitrogen, red for oxygen and yellow for sulfur atoms.[43]

3.1.1 Ligand-Based Modeling

Ligand-based pharmacophore modeling is used when there is no knowledge about the target's protein structure, but active ligands that trigger a proved biological effect on the target, are available. In this type of modeling it is necessary in the first step to cover all possible conformations for each flexible ligand. The second step involves the alignment of these ligands by their chemical features, like H-bond donors/acceptors or lipophilic properties in a three-dimensional space. The aim is to create a 3D pharmacophore model that identifies the chemical interactions with the macromolecule by geometrically overlapping the binding conformations with the for the activity required chemical features.[46,47] The interactions of course must not necessarily resemble the truth, they are rather a hypothesis, built on the belief, that ligands, which are responsible for a type of activity, share similar chemical features that carry out the biological effect.[43] The main problem in ligand-based modeling is conformation flexibility, because the active one is usually unknown. Hence all possible conformations must be generated and displayed. One method to solve this concern is the pre-enumerating concept, where the ligands conformation for each compound is constructed in an earlier stage and saved in a database. The second method calculates the flexibilities of the ligands during the process of pharmacophore modeling, which of course requires a better computer, but no large database memory (on-the-fly method). Typical programs, that are involved in these steps are Catalyst from Accelrys Inc, DiscoTech and Gasp, both from Tripos Inc. They differ in the algorithms used for alignment and ligand flexibility modeling.[46,47] Ligand-based pharmacophore modeling can be done using for example the following software packages: Catalyst, MOE, Phase or LigandScout.[48]

3.1.2 Structure-Based Modeling

Structure-based pharmacophore modeling, on the other hand, can be used in case of the availability of a 3D structure of the protein or a complex consisting of ligand and receptor, e. g. resolved by x-ray crystallography. Due to the higher and higher availability of 3D structures of target proteins derived from the Protein Data Bank (PDB) this concept has become a focus of researches in this area.[48] An advantage over ligand-based modeling is the insight into the active site of the macromolecule, where the essential chemical features for correspondence with ligands can be determined. The shape of the small molecule is taken and the key

interaction points with the protein are being identified. Afterwards, they are transformed into pharmacophoric features and used as the foundation to build the pharmacophore model. Typical software that involve structure-based pharmacophore modeling are for example LigandScout, Pocket v.2 and GBPM.[46]

3.1.3 Pharmacophore-Based Virtual Screening

Pharmacophore-based virtual screening is a process of querying a chemical library with the aid of ligand-based or structure-based pharmacophores for identifying potential drug candidates. Nowadays this technique is an indispensable tool in in silico drug discovery, helping to distinguish between presumably active and inactive compounds within large molecular databases.[49] Molecules that are retrieved by the pharmacophore template from virtual libraries are so called “hits”. They get detected by similar geometrically overlapped chemical features they share with the pharmacophore model, resulting in a hit list, from where the most promising hits are tested in biological assays to confirm their activity.[50] Following computational steps take place in virtual screening: First, same as in ligand-based modeling, the handling of conformational flexibility of the molecules is necessary. It can be dealt again with either the pre-enumerating method or with the on-the-fly method. A further action is pattern identification, which determines if the given pharmacophoric features are present in molecules from the database, filtering out all the ones that do not fit to the query pharmacophore model. This is a simple but effective method which reduces the amount of compounds that need to be screened and consequently saves time.[46] The next approach is the examination of the 3D overlay of the pharmacophoric features with the library molecules, thus a close inspection of possible conformations is needed. Various algorithms of computer software like for example from MOE, LigandScout, Phase or Catalyst take care of this time-consuming, for the quality of the screening important step and decide whether a molecule is rejected or determined as a hit.

Figure 10 provides an overview of the workflows of virtual screening ranging from the creation of a pharmacophore model to hit-list analysis.[43]

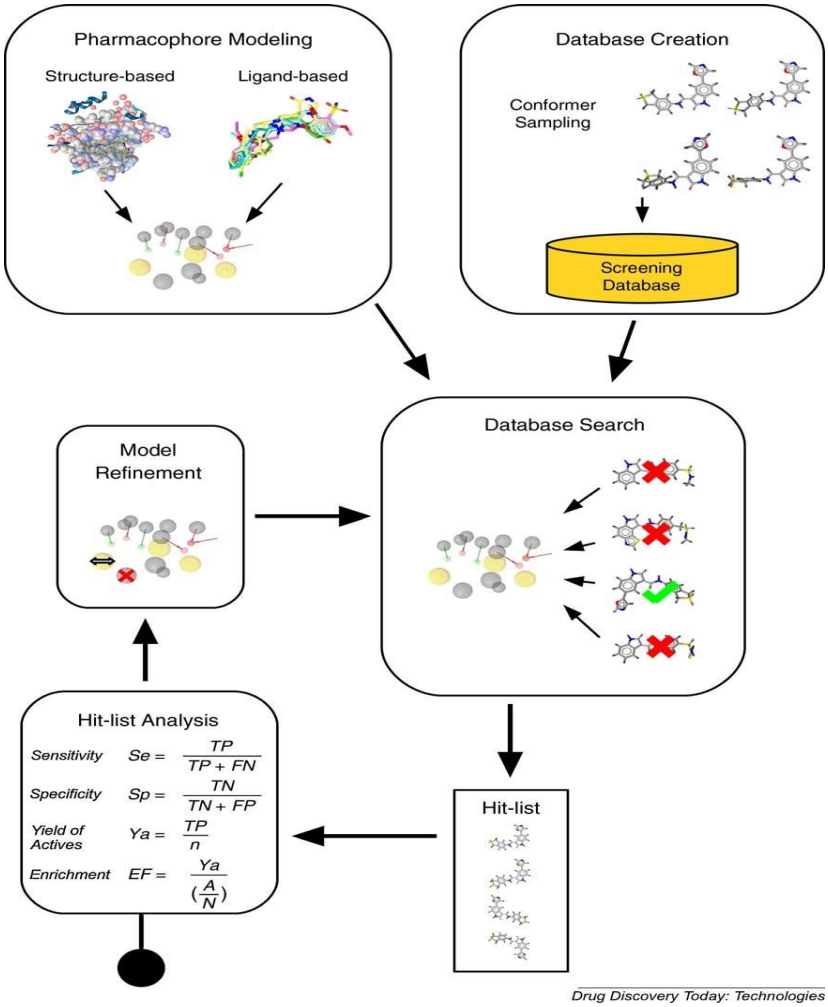


Figure 10: Basic workflows of 3D pharmacophore-based virtual screening.[43]

The validation of the hit-list afterwards serves as a quality assurance for the generated pharmacophore model. In case of unsatisfying outcomes, the model may undergo a refinement in order to retrieve better results. Typical parameters utilized in validation are sensitivity, specificity, yield of actives and the enrichment factor, which are described below.[51]

Sensitivity is the number of true positives divided by the sum of true positives and false negatives, showing the percentage of truly active molecules. If Sensitivity was 1, every active compound would be detected during virtual screening, while 0 would mean not a single hit at all.[51]

$$Se = \frac{TP}{TP+FN}$$

Specificity on the other hand reflects the fraction of the truly inactives being correctly rejected. A specificity of 0 is the case when all inactives are selected as actives, whereas 1 stands of all inactives being declined, thus an ideal pharmacophore.[51]

$$Sp = \frac{TN}{TN+FP}$$

Yield of Actives is the amount of actives in a hit list. It is determined by the number of true actives divided by the number of molecules in the search list.[51]

$$Ya = \frac{TP}{n}$$

The Enrichment factor is a measure for the performance of a pharmacophore model and indicates how many more hits are retrieved compared to a random search. It is calculated by taking the Yield of Actives and dividing it by the sum of the division of actives in the database (A) by the total number of molecules. To obtain an enrichment factor of high value, it is important to have way more inactives than actives, because the more the actives and inactives are equally distributed, the harder it is to discriminate if the pharmacophore model is responsible for the hit detection or if it is just coincidence. The formula for the enrichment reads as follows:[51]

$$E = \frac{Ya}{A/N}$$

Moreover, there is a graphical technique to evaluate and to compare the accuracy of pharmacophore models - the receiver operating characteristic plot (ROC plot). A ROC plot consists of the ROC curve, of two axes, one encoding Sensitivity, the other one 1-Specificity and of a diagonal rising from the lower left to the upper right corner, which represents a random search of the compound library. After virtual screening and hit analysis, Sensitivity and 1-Specificity are plotted against each other and a ROC curve is created. In an ideal case, the curve steeply rises into the direction of the upper left corner (Sensitivity) and then continuous vertically to the right upper end (1-Specificity), curving above the diagonal. The result of such an event is a large area under the curve (AUC), measuring the good performance of the pharmacophore model. In contrast, when the AUC appears to be similar to the one from the diagonal or even smaller, the model is assessed to be of low quality.[52]

4. Materials

4.1 CheMBL

The ChEMBL database, also known as ChEMBLdb (https://www.ebi.ac.uk/chembl/) is a freely accessible database, which contains as of 2017 more than 1.6 million bioactive molecules, including known, approved drugs or just investigational compounds and 14 million bioactivity types, manually extracted from published peer reviewed medicinal chemistry literature. It is operated by the European Bioinformatics Institute (EBI), located at the Wellcome Trust Genome Campus, Hinxton, UK. The purpose behind ChEMBL was to provide scientists an easy access to standardized and high-quality information about chemical structures and their properties, as before datasets, that might have been of interest, had to be collected, harmonized and clustered by researchers themselves, which has been an expensive and time intensive matter. The issue is that the bioactivity data published in journals frequently differs from literature to literature in for e.g. bioactivity units or abbreviations, ending up in inhomogeneous information and in addition, compounds are often depicted only as pictures, so they are unsearchable. Also, in contrast to for example protein structure or gene expression data, there is no obligation for authors to upload small molecule data in open access databases, so the demand for such resources especially from medicinal chemists was rising, resulting in the development of different open source databases aside from ChEMBL, like PubChem BioAssay or ChemBank.[53,54]

But while these databases tend to focus only on few thematic contents, ChEMBL covers all the properties that are necessary to know in the process of evaluating a potential drug candidate. It captures not only the relationship between the ligand and its biological target in form of the experimental measured binding affinity, but also if the ligand possesses drug-like properties, as required physico-chemical parameters, proper absorption, distribution, metabolism, excretion or the right toxicity endpoints, derived from in vitro and/or in vivo experiments, of course standardized to common units and values by scientists.[55,56]

In conclusion, ChEMBL has changed the field of computational chemistry, because scientists working in this field no longer depend only on inhouse data or data they retrieved from literature by themselves, but they can freely access integrative information provided by the

drug discovery community from all over the world, giving everyone the opportunity to get the best out of his research in this area.[56]

4.2 Knime

In the past years the number of available information in life science has risen to an amount that makes the use of data processing software inevitable. With Knime, also called "Konstanz Information Miner", researchers in pharmacogenomics, metabolomics, chem- and bioinformatics that deal daily with many different molecules, assays and protein structures were offered a free data mining program, ideal for analysing and further processing obtained information. The software is a Java-based modular data mining platform and was founded in 2004 at the University of Konstanz in Germany.[57]

Basically, Knime's interface consists of a large workbench, where nodes from the node repository are dragged into by the drag & drop function. Nodes can be understood as single workflow units, which are able to perform different tasks with available data. These nodes again can be connected to each other with a graph through their input/output ports to establish a workflow. In the end this workflow is displayed as a graphically represented pipeline, that allows scientists easily to track their own workflow or helps other analysts to retrace the processing steps. As stated above, the user can choose from different nodes with different functions, for example nodes to read a database, nodes to process and filter data or nodes to visualize the results of a workflow. In most cases a typical workflow in Knime starts with a data reading node, which reads a file or extracts it from a database. The extracted data is stored internal and often represented in table format. To further modify the information, this table is passed on the next node in the pipeline, often a node responsible for the manipulation of the information. A main point of Knime is also the data mining aspect, allowing the user to cluster their input, create decision trees or the simple reduction of data. In the end, the user is able to see the result in form of different visualizations like tables, heat maps or figures.[58,59]

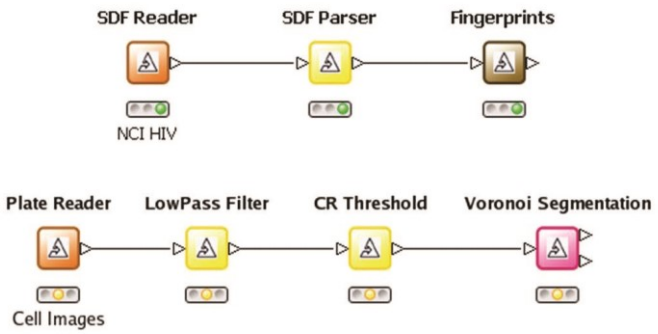


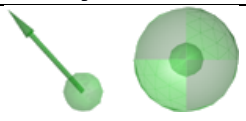
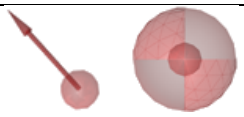
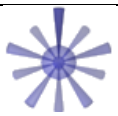
Figure 11: Nodes, linked to each other forming a workflow[59]

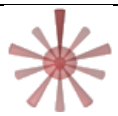
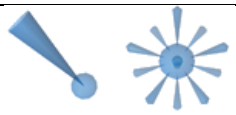
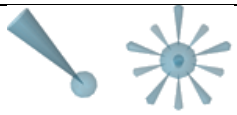
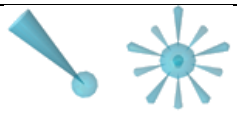
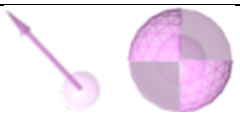
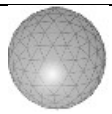
In this scientific study version 3.7.2 of Knime and its LigandScout Expert KNIME Extension, offered by Inte:Ligand GmbH were used to handle the data retrieved from ChEMBL.

4.3 LigandScout

LigandScout, developed by Inte:Ligand GmbH (http://www.inteligand.com/), located in Vienna, Austria, is a commercial scientific software for molecular design, which allows the user to create three-dimensional pharmacophore models from ligand-based and structure-based approaches and to screen these pharmacophores against virtual libraries. The initial release of LigandScout was in 2005, since then many new versions were published.[60] For this diploma thesis, the version LigandScout 4.4 was used. Along with other commercial drug discovery programs, like Moe, Catalyst or Phase, it is one of the most important applications for pharmacophore modelling and virtual screening.[61]

Table 1: Pharmacophoric features with spheres in LigandScout 4.4 (from the LigandScout 4.4 User Manual[62])

Symbol	Feature (shortcut)
	Hydrogen bond donor (HBD)
	Hydrogen bond acceptor (HBA)
	Positive ionizable area (PI)

	Negative ionizable area (NI)
	Hydrophobic interaction (H)
	Aromatic ring (AR)
	Iron binding location (FEB)
	Zinc binding location (ZNB)
	Magnesium binding location (MGB)
	Manganese binding location (MNB)
	Halogen bond donor (XBD)
	Excluded volume

4.4 PDB

The Protein Data Bank (PDB, http://www.rcsb.org/pdb/) is a freely accessible, international depository for three- dimensional structural data of biological macromolecules. It was founded at Brookhaven National Laboratories (BNL) in 1971 in New York, USA. In October 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) took responsibility over the management of the PDB with the goal to provide a high-quality resource for biological, biochemical and computational research. Thanks to the improved technology in crystallography, spectroscopy and electron-microscopy, more and more structures of proteins were elucidated and the PDB has grown to the world's biggest protein database, holding from

7 structures in the beginning to over 140.000 biomolecular structures with related information in March 2019, increasing on a weekly basis.[63,64] In 2003 the worldwide PDB (wwPDB) was established to maintain an uniform content and format of the PDB, since structural data of proteins were submitted from all over the world.[65]

Regarding the PDB data deposition, each macromolecular structure is assigned with a PDB-ID, a four-letter code, for example 1ke6. The first number is always an integer, whereas the others may be alphabetical or numerical. This code has two purposes: First it serves as a reference to the structure on the one hand and, second, it allows the researcher to retrieve the structural data from the PDB database. Prior to deposition, data needs to be acquired. Therefore, coordinates of atoms, NMR restraints etc. are submitted either via e-mail or via ADIT (AutoDep Input Tool), a tool developed by RCSB PDB. It helps to make sure, that data submitted by scientists is managed correctly, does not miss any required information and complies with the rules of the MMCIF dictionary, another tool that contains terms for crystallographic experiments and macromolecular structure.[66]

Deposited structure information already assigned by a PDB-ID needs to go through a validation process. It ensures that the models fit the experimental data and guarantees therefore their quality. Properties as covalent bond distances, angles, stereochemistry, atom nomenclature and close contacts between atoms are checked, the sequence is compared and the distances between water and all polar atoms are calculated. In case of mistakes they are reported to the depositor and corrected in correspondence with the author. Finally, the structure becomes available to all users of the PDB.[66]

As for the PDB website it offers users a broad range of visualizations of protein 3D structures with zoom in functions to be better able to understand biological processes or to inspect drug-target interactions. Furthermore, there is an excellent developed search function with the possibility to set different search refinements. A wildtype search option allows users to look for biological targets that do not contain mutations, since some PDB entries include genetically modified types. As soon as the desired target is found, a summary of important information, such as release date, determining method, or the reference paper are shown. Moreover, structures, experimental properties or the sequence file may also be downloaded in different file formats, compressed or non-compressed.[67]

5. Results and Discussion

5.1 Data Processing

At first, all compounds from the ChEMBL database which have cytochrome P450 2C19 as target had to be extracted with the ChEMBL-DB-Extractor. For this and all further steps involved in modifying the data, the program Knime 3.7.2 and especially its nodes from the LigandScout Expert KNIME Extensions were used.

Table 2: Used KNIME nodes

KNIME-nodes	Description
ChEMBL-DB-Extractor*	Fetches molecules from the ChEMBL database that show activity towards the selected target
Library Filter*	Node for filtering Molecule Libraries
Table Merger*	Merges two fitting tables together
HiLite Row Splitter	Seperates hilited from non-hilited rows in the input data according to their hilite status at the time of node's execution
Duplicate Remover*	Calculates an Extended Connectivity Fingerprint for each inpout molecule and only keeps the first encountered instance with the same fingerprint
ICon-Node*	Generates conformations using the Inte:Ligand Conformer Generator
SDF-Writer*	Writes molecules into MDL Structure-Data Files (SDF)
SDF-Reader*	Reads molecules from MDL Structure-Data Files (SDF) for subsequent processing
PMZ/PML Pharmacophore-Reader*	Creates a list of pharmacophore-models
LDB-Writer*	Writes molecules into LigandScout Local Screening Database (LDB) files
Database List*	Creates references to LigandScout screening databases (ldb)
Activity Profiling*	Screens ldb databases against multiple LigandScout pharmacophores

* **from LigandScout Expert KNIME Extensions**

The ChEMBL DB Extractor Node fetched 26936 molecules with the following properties: Target name, compound ChEMBL ID, bioactivity type, activity comment, activity operator, activity units, assay description, activity value, molecular formula, SMILES and the reference paper for the activity assay.

Then the molecules were primary filtered by their bioactivity type, giving Inhibition, IC_{50}, K_i, Potency and AC_{50} as the bioactivity types of interest. Afterwards it was necessary to take a closer look at the activity values, the activity comments and the assay descriptions to provide a data set with the highest possible quality. Since the original data was found to exhibit a lot of missing activity values/units, inconclusive activity results or doubtful assays as sources, in addition to a lot of duplicates, many molecules had to be removed. The remaining compounds were filtered by their below depicted activity values, resulting in four different subsets.

Table 3: Threshold- IC_{50}

Activity Class	**Activity Threshold- IC_{50} (nM)**
Very actives	0.0-99.9 nM
Actives	100-1999.9 nM
Less actives	2000.0-9999.9 nM
Inactives	>10000.0 nM

Table 4: Threshold- K_i

Activity Class	**Activity Threshold- K_i (nM)**
Very actives	0.0-99.9 nM
Actives	100-1999.9 nM
Less actives	2000.0-9999.9 nM
Inactives	>10000.0 nM

Table 5: Threshold- Potency

Activity Class	**Activity Threshold- Potency (nM)**
Very actives	0.0-99.9 nM
Actives	100.0-1999.9 nM
Less actives	2000.0-9999.9 nM
Inactives	>10000.0 nM

Table 6: Threshold-INH

Activity Class	**Activity Threshold- INH (nM)**
Very actives	0.0-99.9 nM
Actives	100.0-1999.9 nM
Less actives	2000.0-9999.9 nM
Inactives	>10000.0 nM

Sets that did not belong to the same bioactivity type, but fell into the same activity class, got merged.

At this step it is important to point out, that particularly the very active compounds, which were responsible for the generation of the pharmacophore models later had to be highly standardized and from confident assays. Hence, only the very active compounds of the bioactivity types IC_{50}, K_i, INH and Potency were considered and imported into LigandScout.

Molecules of the type "Inhibition", which were also clearly inhibitors of the enzyme 2C19, were tested with higher concentrations (mostly >1-100µM) and their activity units were stated as % inhibition. Moreover, the bioassays were performed with substrates involving coenzymes and lastly, irreversible inhibition was also investigated in many of these molecules, so a comparison between these and the inhibitors from the other bioactivity types was hardly possible, thus they had to be discarded.

The assays for AC_{50} just described activity outcomes, but often it was not clarified what kind of mechanism of action had been tested. Thus, only compounds with the activity comment "Inactive" have been taken and categorized as inactive molecules. So far, the distribution of the compounds looked as follows:

Table 7: Distribution of the filtered compounds

	Very actives	Actives	Less actives	Inactives	Total
Compounds	147	818	1185	9939	12089

Now the set of very actives was split into three subsets according to the table below. The purpose behind the division was to test, whether the pharmacophore model built of the most potent inhibitors would also be the most selective one.

Table 8: Subgroups activity thresholds

Activity Class	**Activity Threshold (nM)**
Strong very actives	0.0-9.9 nM
Moderate very actives	10.0-49.9 nM
Weak very actives	50-99.9 nM

In order to transfer these compounds from Knime into LigandScout, it was necessary to save them as SDF and LDB Files. The SDF-File (very active molecules) was imported into the ligand-based view of LigandScout, while the two LDB Files, which consisted of all the very actives and inactives served as our virtual library to screen the pharmacophore models against. Conformers had to be created for both, our ligands and our library molecules, using the iCon Conformer Generator applying the following settings (**Figure 12**):

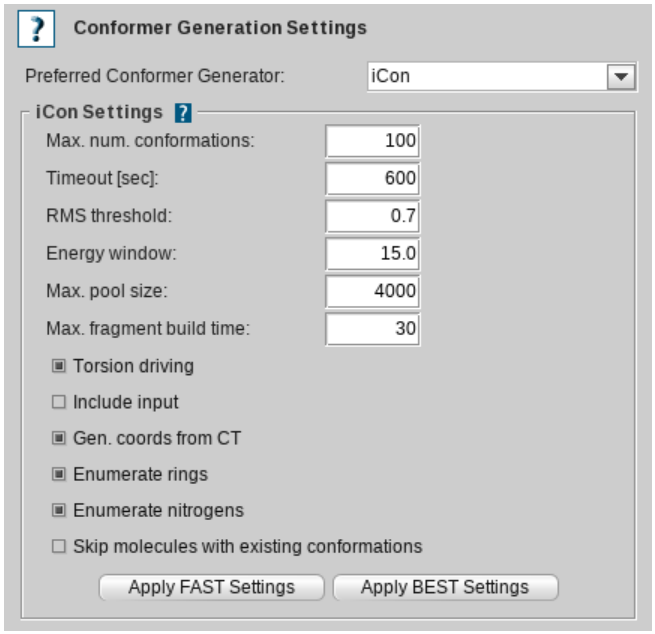


Figure 12: Applied settings for conformer generation

Then the molecules in the ligand-based view were clustered by their pharmacophore feature similarity with the LigandScout's Pharmacophore RDF-Code Similarity algorithm (**Figure 13**).

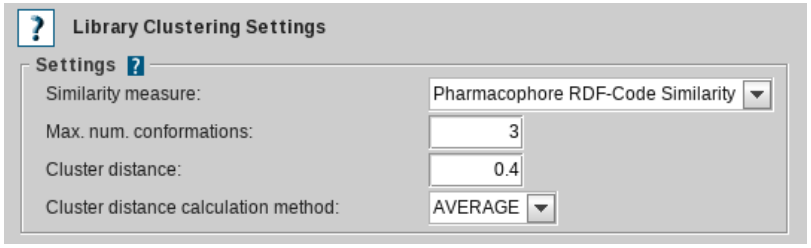


Figure 13: Applied settings for ligand clustering

Molecules that belonged to the same cluster have been used to create shared feature pharmacophore models, which then were used for virtual screening. Clusters, on the other hand, which consisted only of one compound had to be filtered out, since building a pharmacophore model requires at least two different molecules. This process was done for all

three subclasses of the very active molecules. **Table 9** shows the number of clusters created out of the number of available molecules per subset and which of these clusters contained more than one chemical structure and were used for pharmacophore creation.

Table 9: Overview of the generated clusters

Strong very actives **Number of molecules = 12**		**Moderate very actives** **Number of molecules = 68**		**Weak very actives** **Number of molecules = 67**	
Number of clusters = 10		**Number of clusters = 35**		**Number of clusters = 42**	
Cluster ID	**Compound name**	**Cluster ID**	**Compound name**	**Cluster ID**	**Compound name**
10	CHEMBL 1355644	25	CHEMBL 358216	31	CHEMBL 1603852
10	CHEMBL 1356336	25	CHEMBL 358880	31	CHEMBL 1513990
10	CHEMBL 1436169	25	CHEMBL 151170	32	CHEMBL 1327203
		26	CHEMBL 1480806	32	CHEMBL 1232474
		26	CHEMBL 566899	33	CHEMBL 1316956
		26	CHEMBL 1554721	33	CHEMBL 1396261
		26	CHEMBL 585861	33	CHEMBL 1592464
		26	CHEMBL 1408427	33	CHEMBL 1318495
		27	CHEMBL 2018917	33	CHEMBL 1554577
		27	CHEMBL 2018921	33	CHEMBL 1404598
		27	CHEMBL 2019024	34	CHEMBL 2018904
		28	CHEMBL 564782	34	CHEMBL 2018907
		28	CHEMBL 551784	34	CHEMBL 2018908
		28	CHEMBL 559580	34	CHEMBL 2018909
		28	CHEMBL 550170	35	CHEMBL 488400
		28	CHEMBL 556891	35	CHEMBL 502335
		28	CHEMBL 556892	35	CHEMBL 296580
		28	CHEMBL 563716	35	CHEMBL 2018920
		28	CHEMBL 559974	36	CHEMBL 345544
		29	CHEMBL 1320943	36	CHEMBL 1402420
		29	CHEMBL 1602632	37	CHEMBL 1395138
		30	CHEMBL 1221	37	CHEMBL 2219652
		30	CHEMBL 1446716	38	CHEMBL 489738
		30	CHEMBL 91	38	CHEMBL 1559163
		31	CHEMBL 1321655	38	CHEMBL 1370903
		31	CHEMBL 833	39	CHEMBL 3623290
		32	CHEMBL 262968	39	CHEMBL 1330763
		32	CHEMBL 113227	40	CHEMBL 7162
		32	CHEMBL 112592	40	CHEMBL 1412163
		33	CHEMBL 151354	41	CHEMBL 1320406
		33	CHEMBL 151280	41	CHEMBL 1612050
		33	CHEMBL 151974	42	CHEMBL 551782
		33	CHEMBL 1602926	42	CHEMBL 559581
		34	CHEMBL 3605543	42	CHEMBL 564015

		34	CHEMBL 1314367	42	CHEMBL 552156
		34	CHEMBL 1590156	42	CHEMBL 1355088
		34	CHEMBL 1329102	42	CHEMBL 1354190
		34	CHEMBL 1515449		
		34	CHEMBL 1356060		
		35	CHEMBL 231306		
		35	CHEMBL 3898132		
		35	CHEMBL 3891461		
		35	CHEMBL 2207500		

The settings for the pharmacophore generation and virtual screening were as followed (**Figure 14** and **Figure 15**):

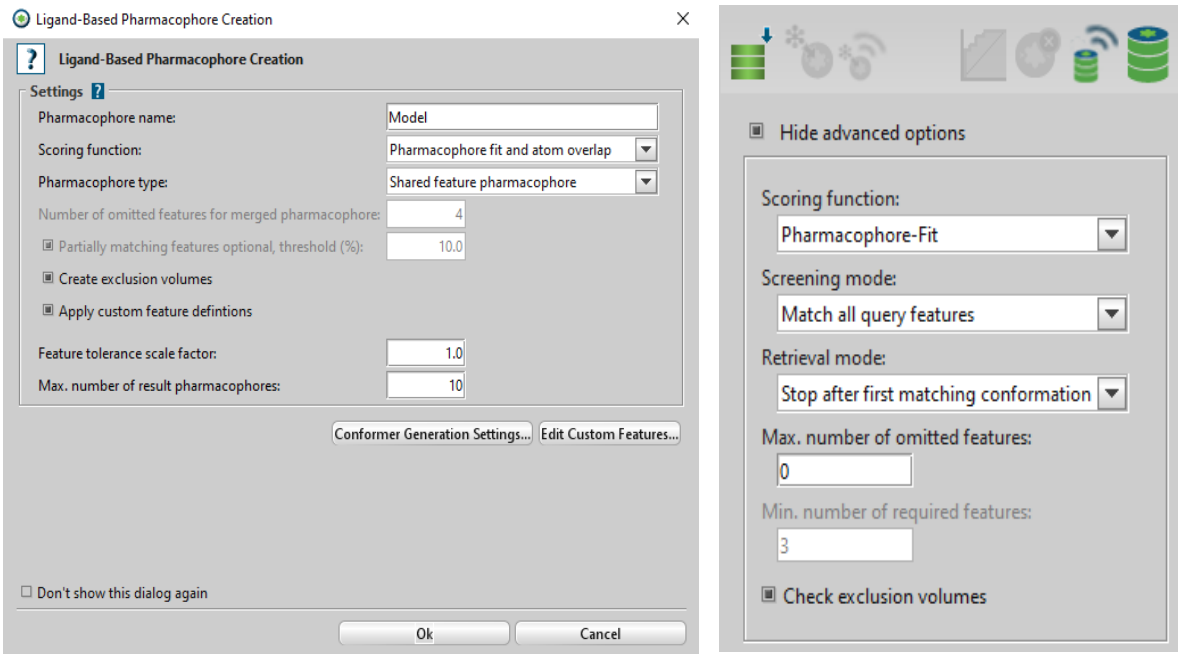


Figure 14 & Figure 15: Settings for pharmacophore generation and virtual screening

All the available clusters were tested by virtual screening. Since many models were not even able to detect the compounds they had been created of, there are only eight pharmacophore models described below in the ligand-based approach section. These were the models with promising results from the first screening on, which could be further tuned to achieve a better performance.

5.2 Ligand-Based Modeling

5.2.1 Pharmacophore Model 1.1

For this model, cluster 10 of the first subset was used, containing the compounds CHEMBL 1436169, CHEMBL 1346336 and CHEMBL 1355644. The initial pharmacophore model consisted of three H-bond acceptors, one H-bond donor, four aromatic rings and three hydrophobic interactions.

Table 10: Model 1.1 before refinement

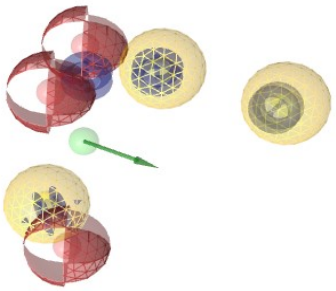
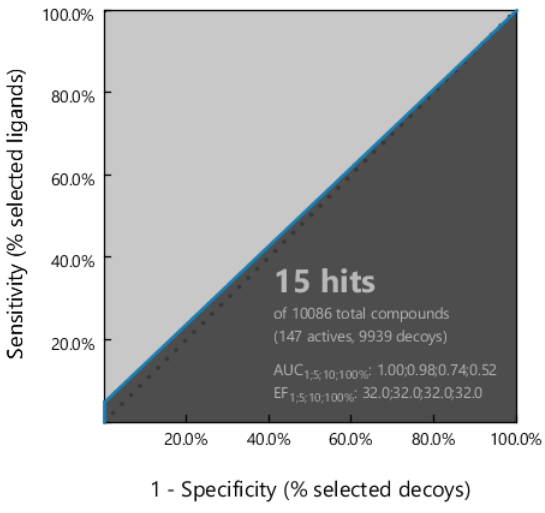
Model name	Depiction of the model
Model 1.1	

Table 11: Screening results of model 1.1 before refinement



True Positives	7

False Positives	8
Sensitivity (100%)	4.76%
Specificity (100%)	99.91%

Table 11 provides an overview of the screening results. The database included 10086 compounds, 147 were actives and 9939 decoys. 15 molecules were retrieved, 7 of them true positives, while 8 of them were false positives, giving a sensitivity of 4,76% and a specificity of 99.91%. EF and AUC are always depicted on the screening results pictures.

Model 1.1 had more false positives than true positives, so there was a need for refinement, which led to ten true positives and only three false positives. The exclusion spheres were responsible for the removal of many false positives, since a lot of them were bigger molecules which might had have the correct alignment of the pharmacophoric features but may have simply not fitted the binding site of the enzyme in the biological assay. Furthermore, one aromatic feature was removed, leading to a higher sensitivity. In the end the features included three H-bond acceptors, one H-bond donor, two aromatic features and two hydrophobic properties.

Table 12: Model 1.1 after refinement

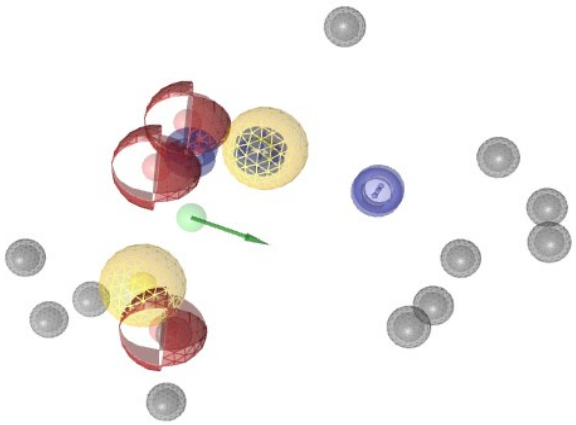
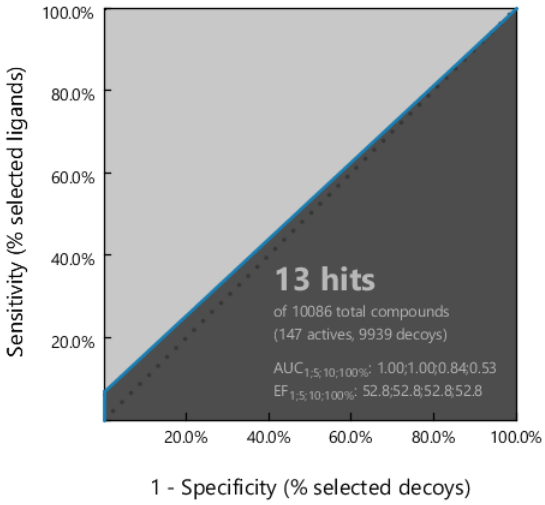
Model Name	Depiction of the Model
Model 1.1	

Table 13: Screening results of model 1.1 after refinement



True Positives	10
False Positives	3
Sensitivity (100%)	6.80%
Specificity (100%)	99.96%

5.2.2 Pharmacophore Model 2.1

Model 2.1 has been built starting with compounds CHEMBL 358216, CHEMBL 358880 and CHEMBL 151170 belonging to cluster 25 of the moderate very actives group. The following features were found on the pharmacophore model: two H-bond acceptors, two aromatic rings and two hydrophobic features.

In virtual screening experiment, this model obtained 62 hits: 13 true positives and 49 false negatives, making model tuning necessary. Same as in Model 1.1, addition of exclusion spheres contributed to higher selectivity by removing almost all false positives, which again were bigger molecules. The improved model also retrieved 13 true positives, since the features had not been changed, but only 2 false positives, making it more specific than before.

Table 14: Model 2.1 before refinement

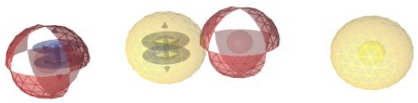
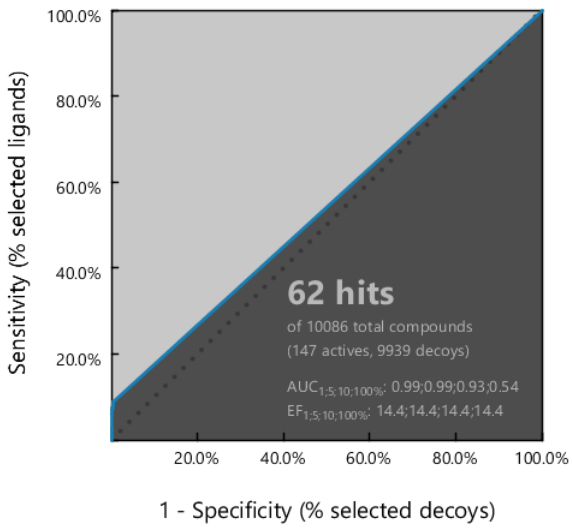
Model name	Depiction of the model
Model 2.1	

Table 15: Screening results of model 2.1 before refinement



True Positives	13
False Positives	49
Sensitivity (100%)	8.84%
Specificity (100%)	99.50%

Table 16: Model 2.1 after refinement

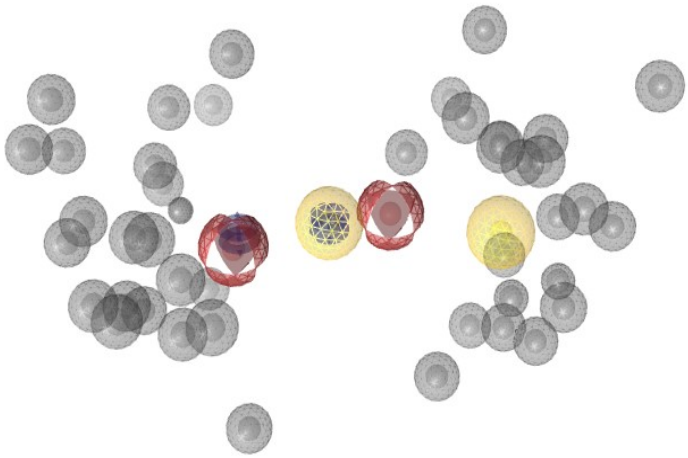
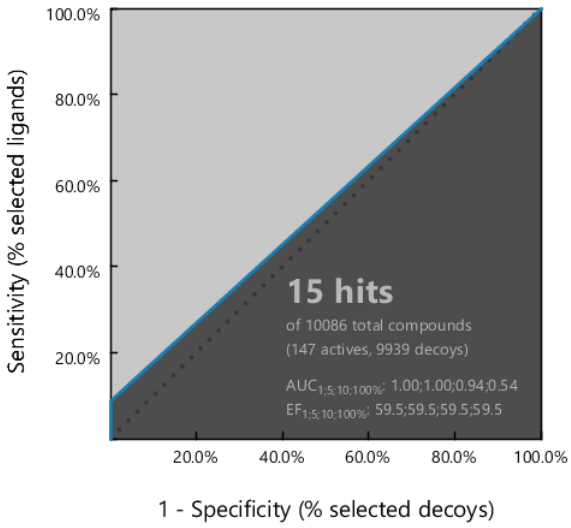
Model name	Depiction of the model
Model 2.1	

Table 17: Screening results of model 2.1 after refinement



True Positives	13
False Positives	2
Sensitivity (100%)	8.84%
Specificity (100%)	99.98%

5.2.3 Pharmacophore Model 2.2

Model 2.2 was built starting of compounds CHEMBL 2018917, CHEMBL 2018921 and CHEMBL 2019024 belonging to cluster 27, second subset. The following features were present in this model: two H-bond acceptors, one H-bond donor, three aromatic rings and two hydrophobic interactions. Also in this case, addition of exclusion volume spheres to minimize the rate of false positives with a really good result: After refinement the model retrieved 15 true positives and only 7 false positives, compared to 17 true positives and 47 false positives beforehand.

Table 18: Model 2.2 before refinement

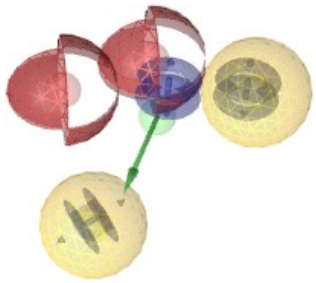
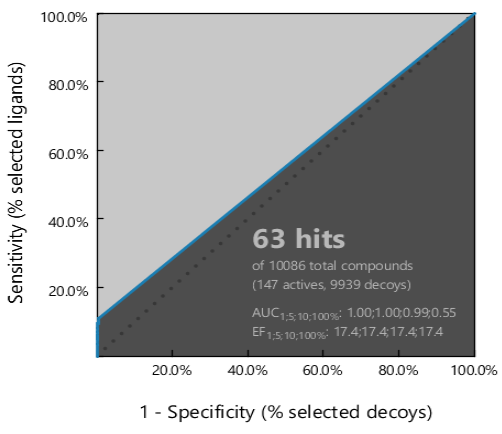
Model name	Depiction of the model
Model 2.2	

Table 19: Screening results of model 2.2 before refinement



True Positives	16
False Positives	47
Sensitivity (100%)	10.88%
Specificity (100%)	99.53%

Table 20: model 2.2 after refinement

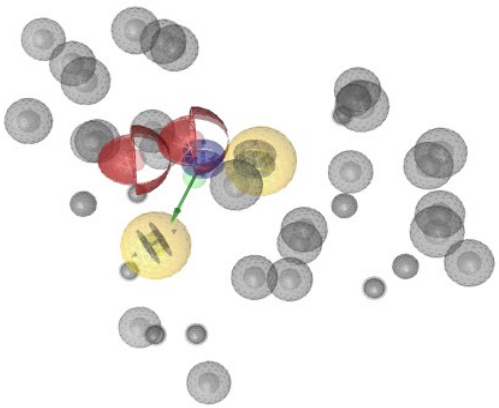
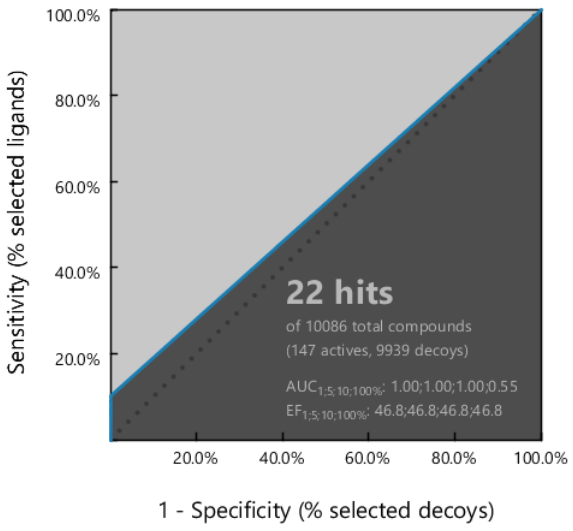
Model name	Depiction of the model
Model 2.2	

Table 21: Screening results of model 2.2 after refinement



True Positives	15
False Positives	7
Sensitivity (100%)	10.20%
Specificity (100%)	99.93%

5.2.4 Pharmacophore Model 2.3

The substances CHEMBL 1221, CHEMBL 1446716, CHEMBL 91 and CHEMBL 1321655 of cluster 30 were used to create this model, which was characterized by one H-bond acceptor, two aromatic rings, five hydrophobic features and three halogen-bond donors. As **table 23** shows, this model turned out to be very specific, but by retrieving only four true positive hits, not quite sensitive. Hence, some features had to be deleted, ending up with one H-bond acceptor, two aromatic rings, four hydrophobic interactions and two halogen-bond donors. Thus, the refined model retrieved four more true positives, increasing the sensitivity from 2.72% to 5.44%.

Table 22: Model 2.3 before refinement

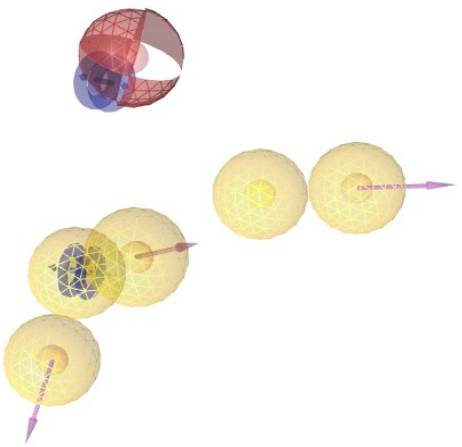
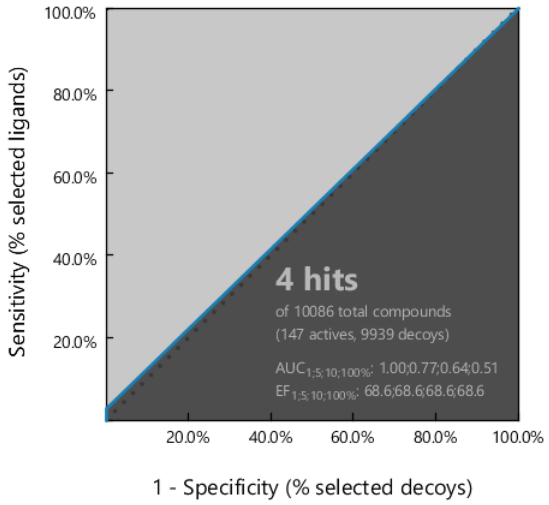
Model name	Depiction of the model
Model 2.3	

Table 22: Screening results of model 2.3 before refinement



True Positives	4
False Positives	0
Sensitivity (100%)	2.72%
Specificity (100%)	100%

Table 24: Model 2.3 after refinement

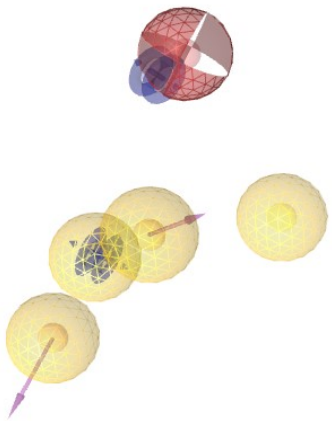
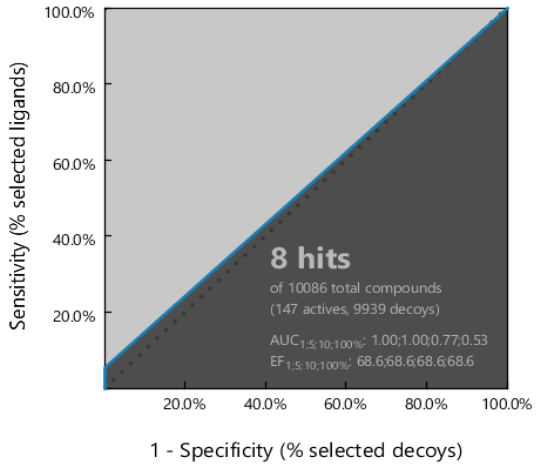
Model name	Depiction of the model
Model 2.3	

Table 25: Screening results of model 2.3 after refinement



True Positives	8
False Positives	0
Sensitivity (100%)	5.44%
Specificity (100%)	100.00%

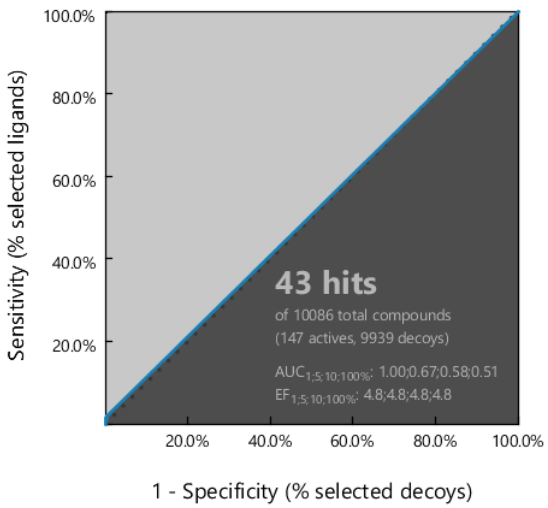
5.2.5 Pharmacophore Model 2.4

Model 2.4 contained the molecules CHEMBL 1321655, CHEMBL 833 of cluster 31 and consisted of only four features: two aromatic rings, one hydrophobic feature and one iron-binding location. On the first sight, particularly after the first screening, it did not look like a promising model, because the ratio between true positives and false positives was much in favour towards the false positives. Since there has been no iron-binding feature in the models before, this model differed from the others and therefore, might have been able to retrieve different very active molecules, as the models before. After adding a lot of exclusion spheres, the original 3 true positives remained the same, but the rate of false positive could be reduced from 40 to 3. The features underwent no refinement, as changing feature types or removing one led to a much more unspecific pharmacophore model.

Table 26: Model 2.4 before refinement

Model name	Depiction of the model
Model 2.4	

Table 27: Screening results of model 2.4 before refinement



True Positives	3
False Positives	40
Sensitivity (100%)	2.04%
Specificity (100%)	99.60%

Table 28: Model 2.4 after refinement

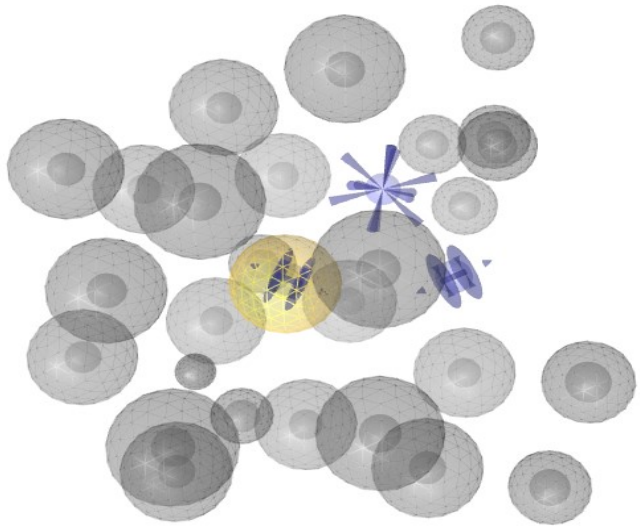
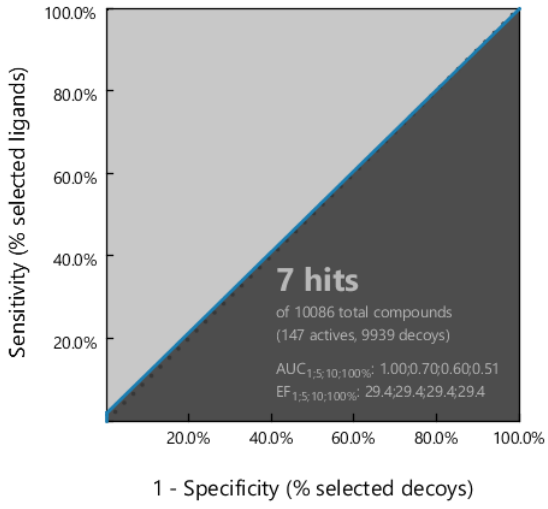
Model name	Depiction of the model
Model 2.4	

Table 29: Screening results of model 2.4 after refinement



True Positives	3
False Positives	4
Sensitivity (100%)	2.04%
Specificity (100%)	99.96%

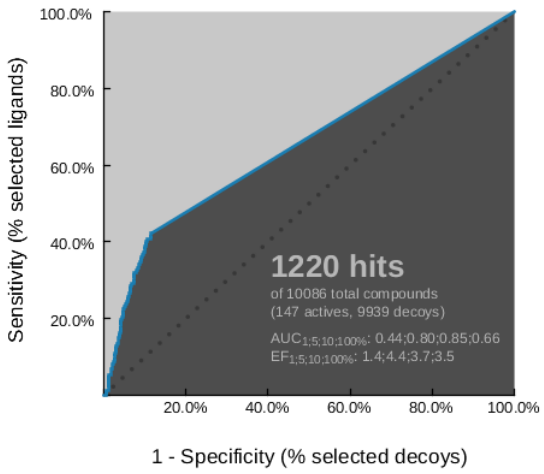
5.2.6 Pharmacophore Model 2.5

For pharmacophore model 2.5 cluster 33 with the compounds CHEMBL 151354, CHEMBL 151280, CHEMBL 151974 and CHEMBL 1602926 served as a template. As depicted in **Table 30**, this model displayed only three hydrophobic features, resulting in an already anticipated, very unspecific screening test. A refinement had to be carried out by including an aromatic ring and one H-bond acceptor. These additional two features have made the model selective towards true positives, obtaining seven true positives and five false positives lastly.

Table 30: Model 2.5 before refinement

Model name	Depiction of the model
Model 2.5	

Table 31: Screening results of model 2.5 before refinement



True Positives	62

False Positives	1158
Sensitivity (100%)	42.17%
Specificity (100%)	89.56%

Table 32: Model 2.5 after refinement

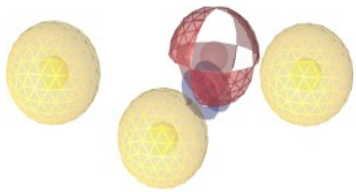
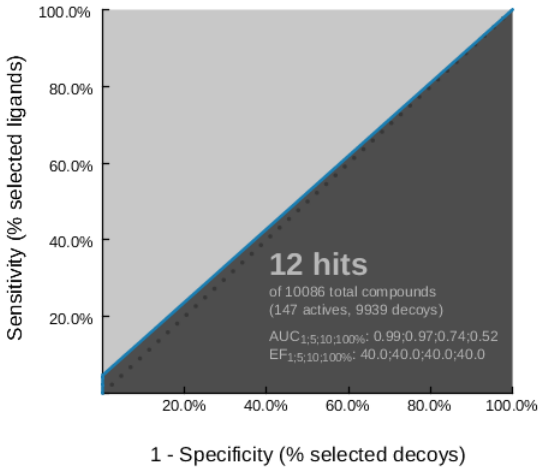
Model name	Depiction of the model
Model 2.5	

Table 33: Screening results of model 2.5 after refinement



True Positives	7
False Positives	5
Sensitivity (100%)	4.76%
Specificity (100%)	99.95%

5.2.7 Pharmacophore Model 3.1

Model 3.1 was the first model built of molecules from the third subset. CHEMBL 1316956, CHEMBL 1396261, CHEMBL 1592464, CHEMBL 1318495 and CHEMBL 1554577 of cluster 33 have been used to generate this model, which was characterized by two H-bond acceptors, one H-bond donor, three aromatic rings and one hydrophobic feature. The first virtual screening retrieved 11 true positives, but also 16 false positives, thus a refinement was required. Since many of the false positives were big molecules, some exclusion spheres were appended to limit the size of the possibly obtained compounds. The second screening proved this concept and retrieved the original 11 true positive, but only 1 false positive.

Table 34: Model 3.1 before refinement

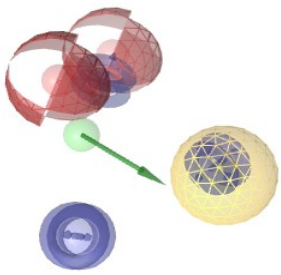
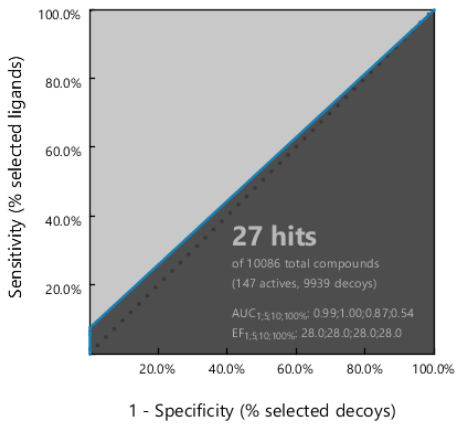
Model name	Depiction of the model
Model 3.1	

Table 35: Screening results of model 3.1 before refinement



True Positives	11
False Positives	16
Sensitivity (100%)	7.48%
Specificity (100%)	99.84%

Table 36: Model 3.1 after refinement

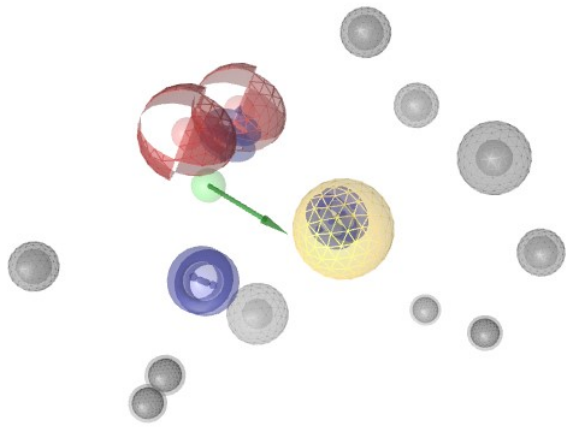
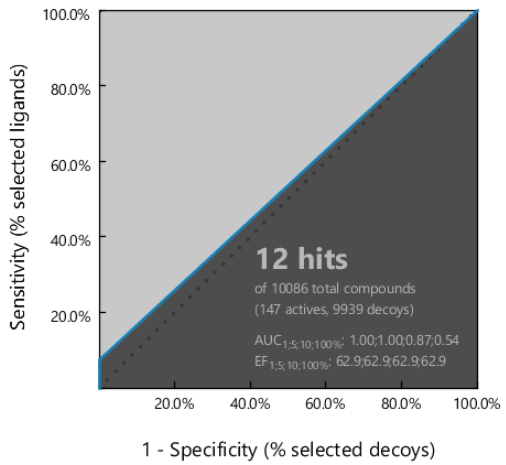
Model name	Depiction of the model
Model 3.1	

Table 37: Screening results of model 3.1 after refinement



True Positives	11
False Positives	1
Sensitivity (100%)	7.48%
Specificity (100%)	99.99%

5.2.8 Pharmacophore Model 3.2 & Activity Profiling

Pharmacophore model 3.2 has been generated of CHEMBL 2018904, CHEMBL 2018907, CHEMBL 2018908 and CHEMBL 2018909 belonging to cluster 34. As **Table 38** demonstrates, model 3.2 had three H-bond acceptors, one H-bond donor, two aromatic rings and two hydrophobic regions. It retrieved ten true positives and nine false positives, thus a refinement was appropriate to improve the model's performance. By adding exclusion spheres, the rate of false positives could be reduced from nine to three, changing the specificity from 99.91% to 99.97%. The number of actives has also been reduced from ten to seven, fortunately the three hits which got lost here were already covered by the other models.

Table 38: Model 3.2 before refinement

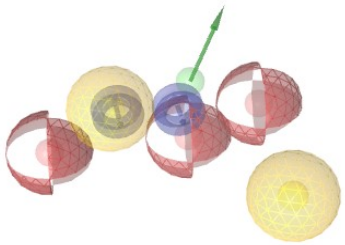
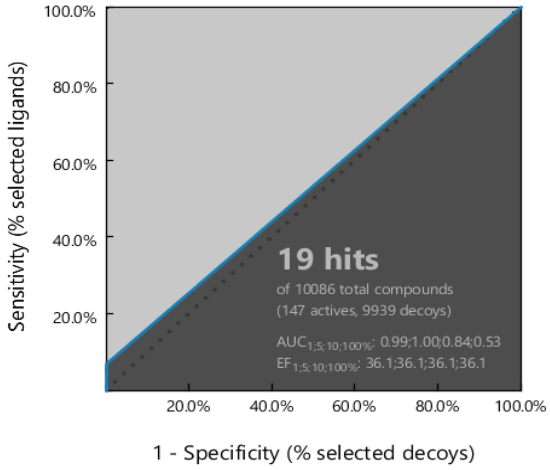
Model name	Depiction of the model
Model 3.2	

Table 39: Screening results of model 3.2 before refinement



True Positives	10
False Positives	9
Sensitivity (100%)	6.80%
Specificity (100%)	99.91%

Table 40: Model 3.2 after refinement

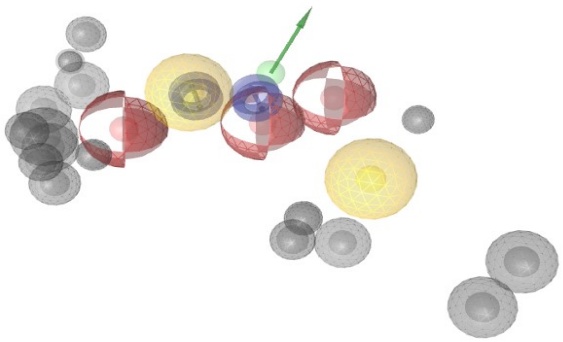
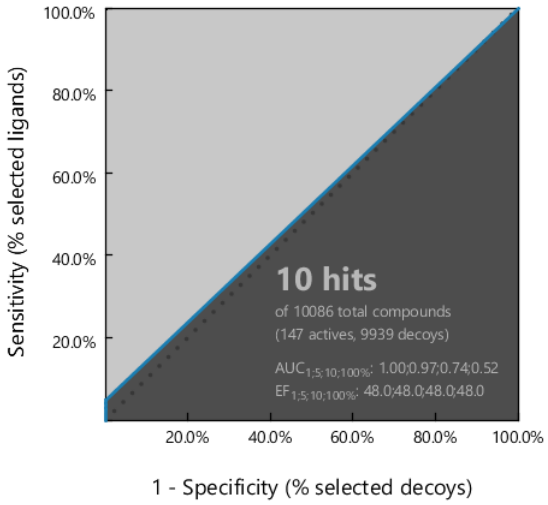
Model name	Depiction of the model
Model 3.2	

Table 41: Screening results of model 3.2 after refinement



True Positives	7
False Positives	3
Sensitivity (100%)	4.76%
Specificity (100%)	99.97%

At this point it was important to test how many of the very active compounds were already covered by the above described eight pharmacophore models. The activity profiling node from LigandScout Expert KNIME Extensions was used to perform sceening and to visualize the results as a heat map (**Figure 16)**.

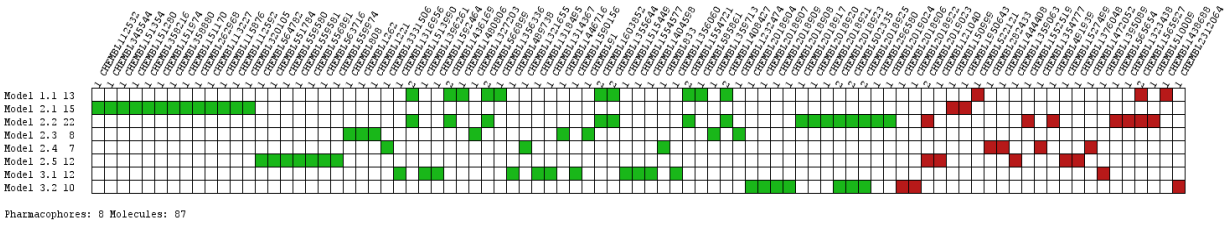


Figure 16: Activity profiling of pharmacophore models 1.1 – 3.2 after refinement

As demonstrated in **Figure 16**, 64 true actives out of 147 and 23 decoys out of 9939 have been already retrieved by the pharmacophore models 1.1 – 3.2. Since 64 was not even the half of 147 (43,54%), further models had to be generated.

In the next step, all 83 active compounds, that have not been obtained by any of the models until then, had been separated from the already retrieved molecules and clustered using the LigandScout RDF-Code Similarity clustering algorithm (**Figure 14**) into 62 clusters.

13 clusters consisted of more than 2 molecules, hence 13 new pharmacophore models have been created. Only six of them proved to be useful in the following virtual screenings and were therefore used to create additional models. These are described below as Pharmacophore models 4.1 – 4.6.

Table 42: Overview of the clusters used for the further 6 pharmacophore models

Very actives – not covered **Number of molecules = 83**	
Number of clusters = 62	
Cluster ID	**Compound name**
50	CHEMBL 2079587
50	CHEMBL 2036210
51	CHEMBL 3898132
51	CHEMBL 3891461
52	CHEMBL 1602632
52	CHEMBL 1446716
53	CHEMBL 1437447
53	CHEMBL 1315820
54	CHEMBL 1565525
54	CHEMBL 1370903
60	CHEMBL 551782
60	CHEMBL 550170
60	CHEMBL 556892
60	CHEMBL 564015
60	CHEMBL 552156
60	CHEMBL 3605543
60	CHEMBL 1559163

5.2.9 Pharmacophore Model 4.1

Model 4.1 derived from the molecules CHEMBL 2079587 and CHEMBL 2036210 of cluster 50. This model contained the following features: Six H-bond acceptors, one negative ionizable feature, three aromatic rings and three hydrophobic interactions. The high number of features made this model very selective, in virtual screening it recognized two true positives and no

decoys at all. The refinement of this model proved itself not satisfying, as removing features led to the retrieval of a lot more decoys, while the rate of true positives did not rise in a meaningful way, so the model remained with its initial properties.

Table 33: Model 4.1

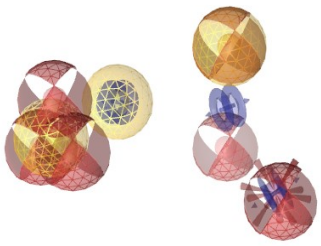
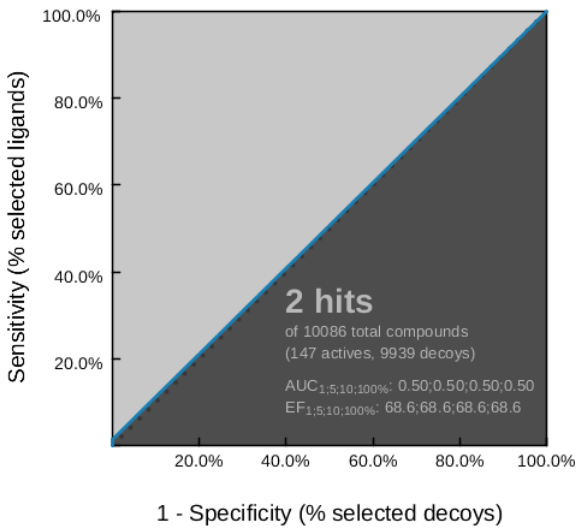
Model name	Depiction of the model
Model 4.1	

Table 44: Screening results of model 4.1



True Positives	2
False Positives	0

Sensitivity (100%)	1.36%
Specificity (100%)	100.00%

5.2.10 Pharmacophore Model 4.2

The second model was made of the structures CHEMBL 3898132 and CHEMBL 3891461 belonging to Cluster 51. As Model 4.1, it consisted of a large number of features: Four H-bond acceptor, one H-bond donor, four aromatic rings, two hydrophobic features, one halogen bond donor and finally one iron-binding location. As expected, model 4.2 retrieved in virtual screening only two hits and no decoys. Fortunately, the model's performance could be improved by deleting some features, so after refinement, it retrieved three true positives and one inactive, while displaying the following features: Two H-bond acceptors, three aromatic rings and one iron-binding location.

Table 45: Model 4.2 before refinement

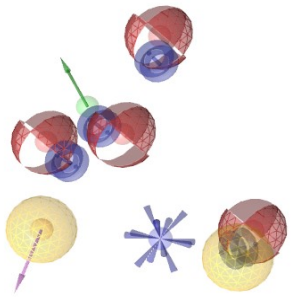
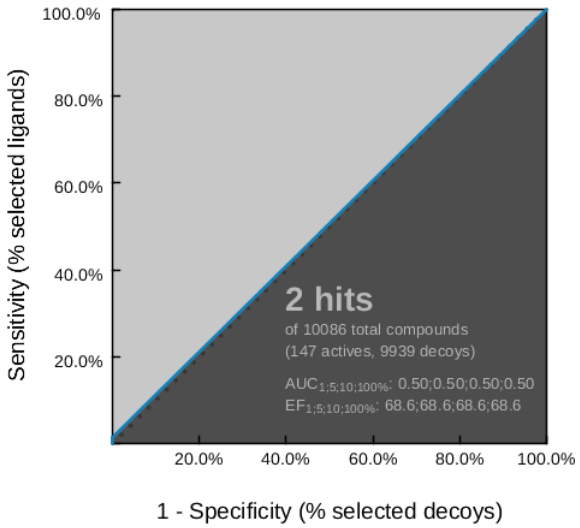
Model name	Depiction of the model
Model 4.2	

Table 46: Screening results of model 4.2 before refinement



True Positives	2
False Positives	0
Sensitivity (100%)	1.36%
Specificity (100%)	100.00%

Table 47: Model 4.2 after refinement

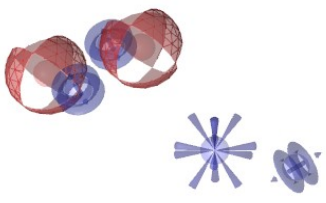
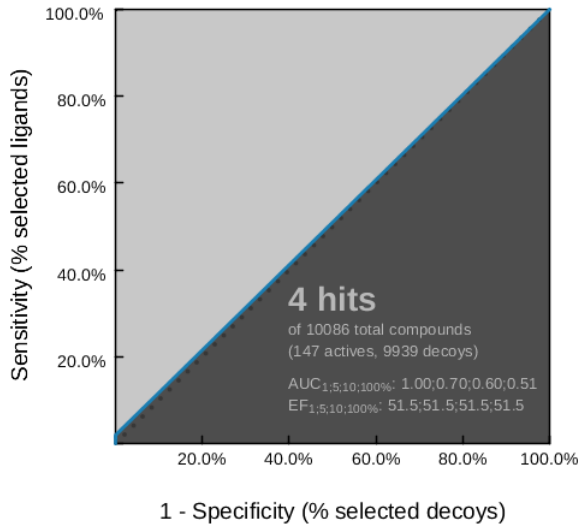
Model name	Depiction of the model
Model 4.2	

Table 48: Screening results of model 4.2 after refinement



True Positives	3
False Positives	1
Sensitivity (100%)	2.04%
Specificity (100%)	99.99%

5.2.11 Pharmacophore Model 4.3

For generation of this model, cluster 53 containing the compounds CHEMBL 1602632 and CHEMBL 1446716 served as the input. The result of the shared pharmacophore generation was a model with one H-bond acceptor and one H-bond donor with the same coordinates, one aromatic ring and three hydrophobic interactions. This model obtained three active hits and six decoys. A refinement was possible by increasing the excluded volume spheres, so in the end the rate of decoys could be minimized from six to three, while the number of true positives stayed the same.

Table 49: Model 4.3 before refinement

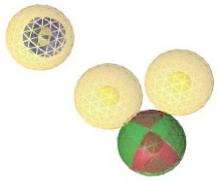
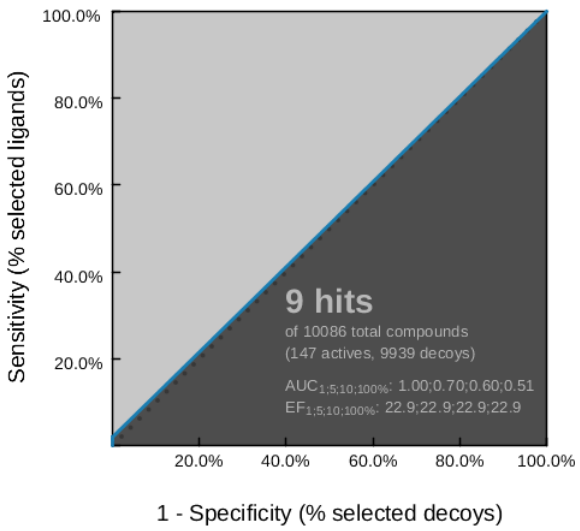
Model name	Depiction of the model
Model 4.3	

Table 50: Screening results of model 4.3 before refinement



True Positives	3
False Positives	6
Sensitivity (100%)	2.04%
Specificity (100%)	99.94%

Table 41: Model 4.3 after refinement

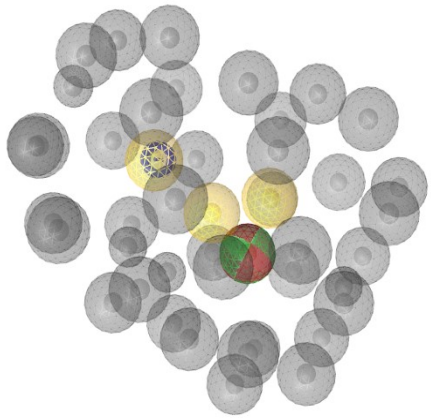
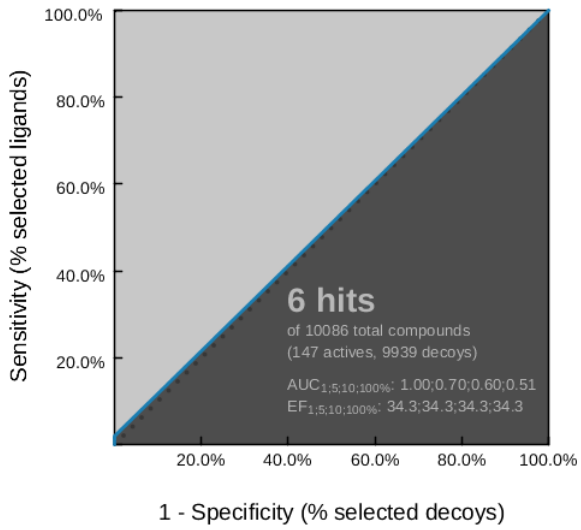
Model name	Depiction of the model
Model 4.3	

Table 52: Screening results of model 4.3 after refinement



True Positives	3
False Positives	3
Sensitivity (100%)	2.04%
Specificity (100%)	99.97%

5.2.12 Pharmacophore Model 4.4

The compounds CHEMBL 1437447 and CHEMBL 1315820 of cluster 53 have been used to generate this pharmacophore model. Five H-bond acceptors, three aromatic rings and one hydrophobic interaction described this model's properties. With two true positives and two decoys in virtual screening the result was acceptable. An attempt to improve the model failed, since removing features only increased the rate of decoys, while true positives which were obtained then, had been covered already by other models.

Table 53: Model 4.4

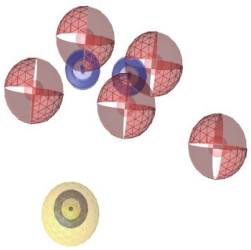
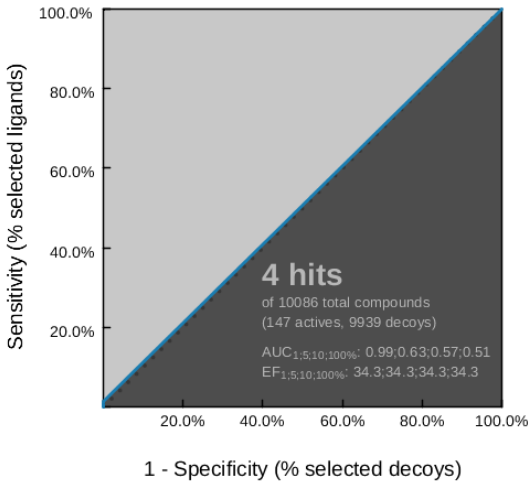
Model name	Depiction of the model
Model 4.4	

Table 54: Screening results of model 4.4



True Positives	2
False Positives	2
Sensitivity (100%)	1.36%
Specificity (100%)	99.98%

5.2.13 Pharmacophore Model 4.5

Model 4.5 consisted of the chemical structures CHEMBL 1565525 and CHEMBL 1370903 of cluster 54. It contained two H-bond acceptors, two aromatic rings and two hydrophobic interactions. Model 4.5 led to the detection of 3 very actives and 16 decoys. Regarding the decoys, most of these molecules had large structures, so they could be removed by adding exclusion volume spheres. This way, the decoys were reduced from 16 to 7, making the model more specific. During refinement, the feature count and feature types underwent no change due to a loss of specificity by really not meaningful improving sensitivity.

Table 55: Model 4.5 before refinement

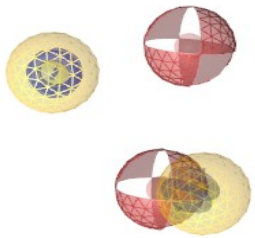
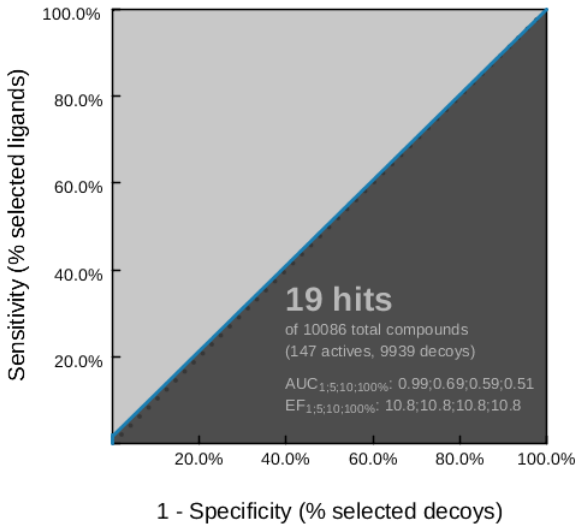
Model name	Depiction of the model
Model 4.5	

Table 56: Screening results of model 4.5 before refinement



True Positives	3
False Positives	16
Sensitivity (100%)	2.04%
Specificity (100%)	99.84%

Table 57: Model 4.5 after refinement

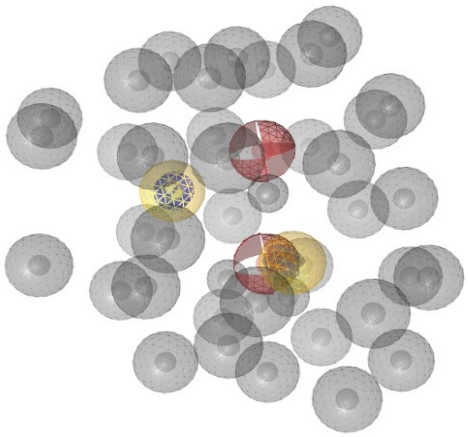
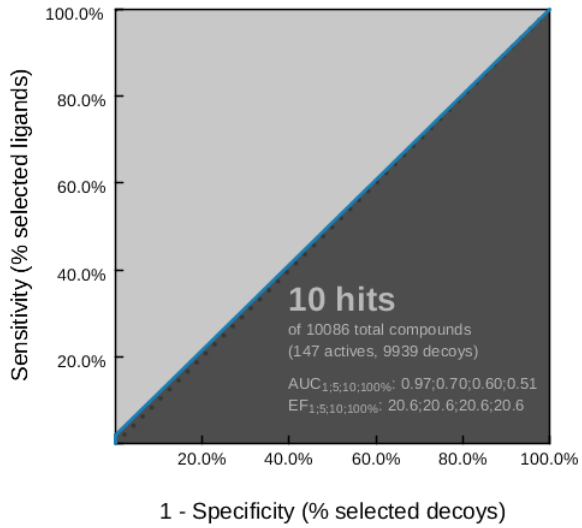
Model name	Depiction of the model
Model 4.5	

Table 58: Screening results of model 4.5 after refinement



True Positives	3
False Positives	7
Sensitivity (100%)	2.04%
Specificity (100%)	99.93%

5.2.14 Pharmacophore Model 4.6 & Activity Profiling

For pharmacophore model 4.6 cluster 60 was used, which contained the molecules CHEMBL 551782, CHEMBL 550170, CHEMBL 556892, CHEMBL 564015, CHEMBL 552156, CHEMBL 3605543 and CHEMBL 1559163. With only two H-bond acceptors, one aromatic ring and two hydrophobic interactions this model was found to be too unspecific, which became evident from the high false positive retrieval rate. A model refinement was therefore done. As depicted in **Table 61**, only one additional halogen-bond donor was responsible for the vast improvement of this model. The new pharmacophore performed outstanding in virtual screening, leading to ten true actives and not a single decoy.

Table 59: Model 4.6 before refinement

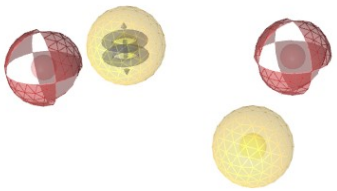
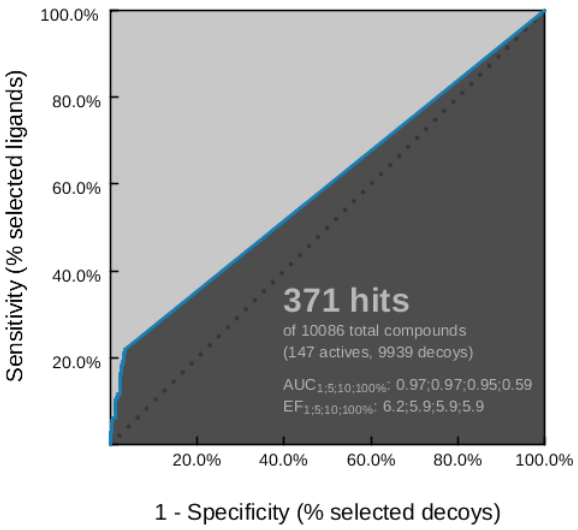
Model name	Depiction of the model
Model 4.6	

Table 60: Screening results of model 4.6 before refinement

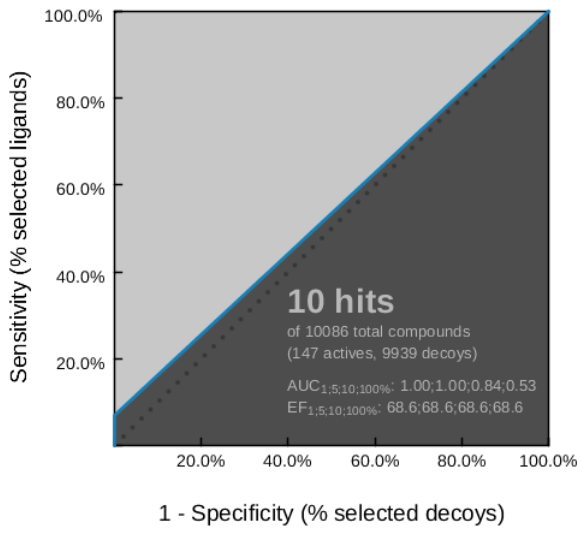


True Positives	32
False Positives	339
Sensitivity (100%)	21.77%
Specificity (100%)	96.70%

Table 61: Model 4.6 after refinement

Model name	Depiction of the model
Model 4.6	

Table 62: Screening results of model 4.6 after refinement



True Positives	10
False Positives	0
Sensitivity (100%)	6.80%
Specificity (100%)	100.00%

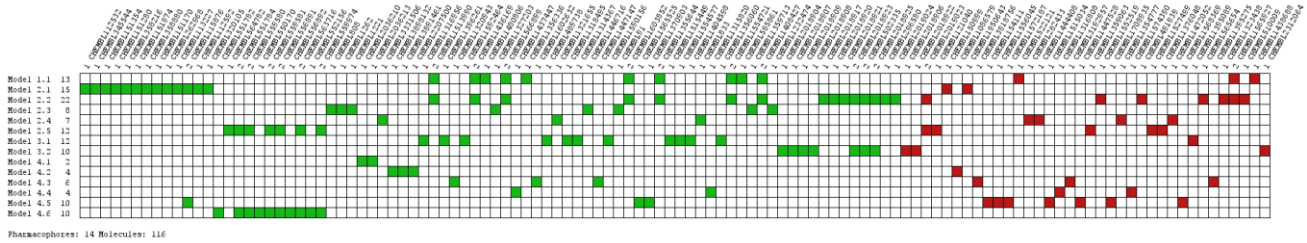


Figure 17: Activity profiling of the 14 pharmacophore models after refinement

The second activity profiling experiment shows that the new 6 pharmacophore models (4.1 – 4.6) covered 16 additional true actives and 13 additional decoys, while all pharmacophore models together detected 80 of 147 (54,42%) very actives and only 36 of 9939 (0,36%) inactives. Further attempts to retrieve more true positives with new pharmacophore models were not fruitful due to a high rate of false positives and only a few new true active molecules. Since there was a strong hit overlap between model 2.5 and model 4.6**,** the decision was made to remove model 2.5, causing a loss of one true positive and four decoys, that seemed to be meaningful to provide a high selectivity of the models. The final hit rate at that moment was as follows: 79 of 147 (53,74%) true actives and 32 of 9939 (0,32%) non-actives have been detected by the 13 pharmacophore models.

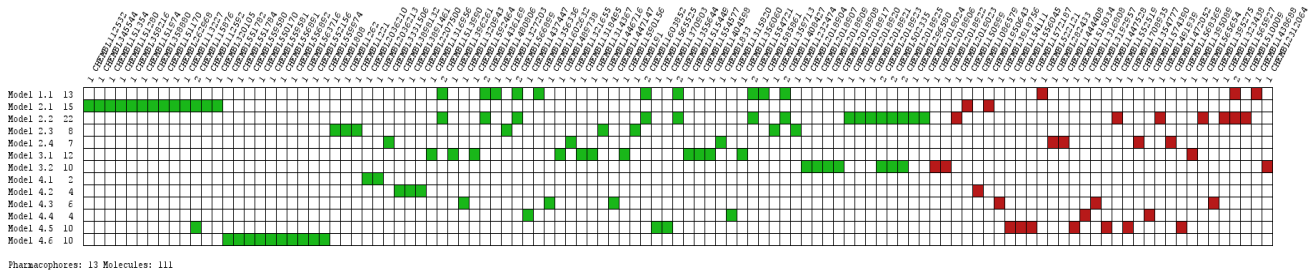
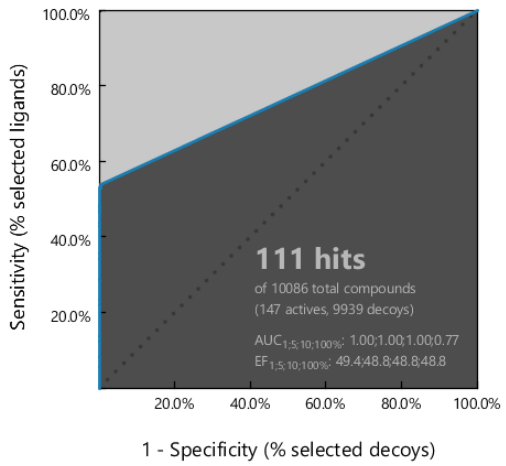


Figure 18: Activity profiling of the 13 pharmacophore models without model 2.5

Figure 19 gives an overview of the compounds that had been obtained by the pharmacophore models. In contrast, **Figure 20** represents those very active ChEMBL structures, which could not be found by model 1.1 – 4.6.

At this step, all pharmacophore models were imported into the LigandScout screening view and tested against the virtual library. **Table 63** demonstrates the detailed screening results and the corresponding ROC Plot of the parallel screen.

Table 63: Screening results of models 1.1 – 4.6 in parallel screening



True Positives	79
False Positives	32
Sensitivity (100%)	53.74%
Specificity (100%)	99.68%

One last important aspect to point out in the process of ligand-based pharmacophore modeling pertains the refinement of all pharmacophore models. The model tuning took place in form of removing or adding features or by changing either their position or/and feature tolerance. These were the key steps to increase the number of hits in order to reach a higher sensitivity. On the other hand, if a refinement had the purpose to reduce the number of decoys, exclusion volumes or just an additional feature were enough in most cases.

Conclusively, the way from the initial to the final model required numerous slight changes of features, careful inspections of the excluded volume spheres in order to not remove true actives, and a large number of repeated virtual screening experiments. Therefore, in the model descriptions above, only the original and the final models are given, since protocolling each individual step for each pharmacophore model would have taken too long and would have gone beyond the scope of this diploma thesis.

Figure 19: Screenshot of the retrieved 79 very active molecules

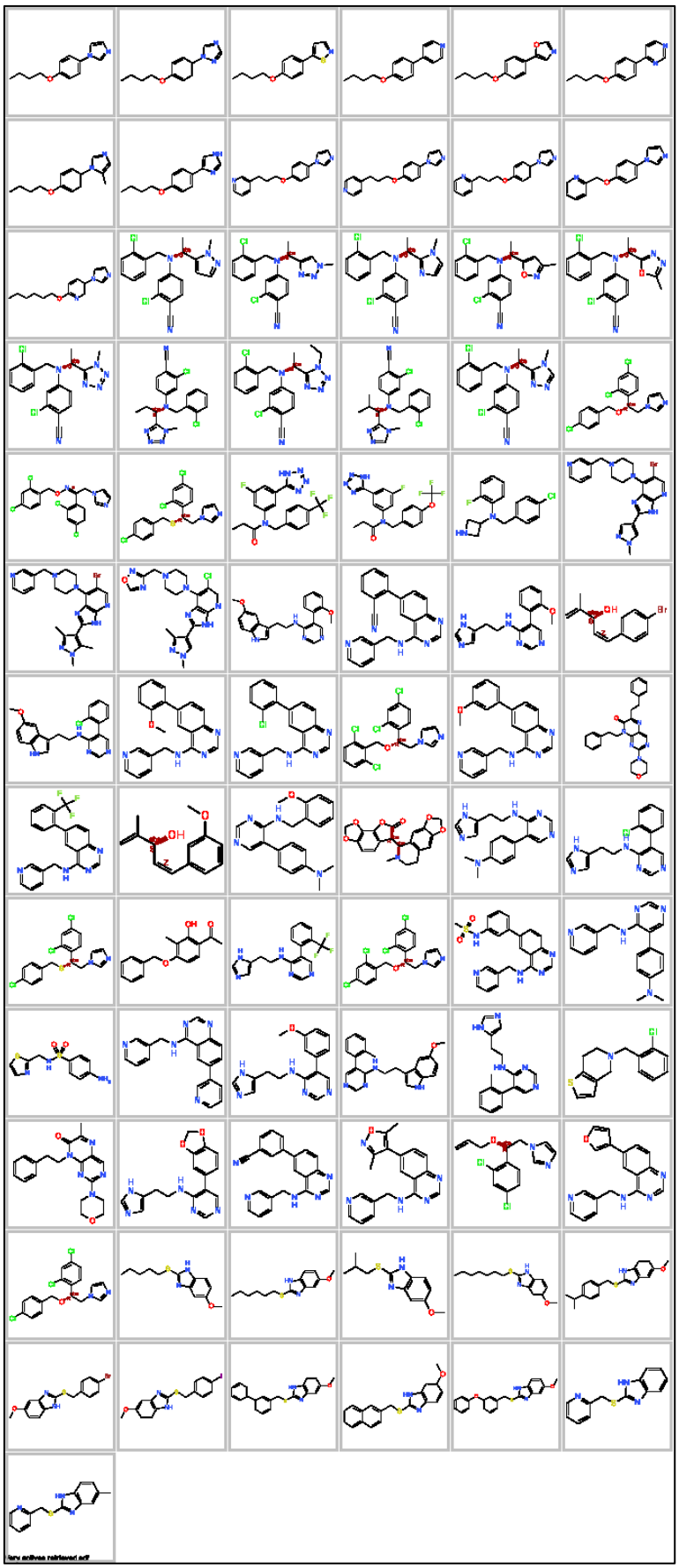
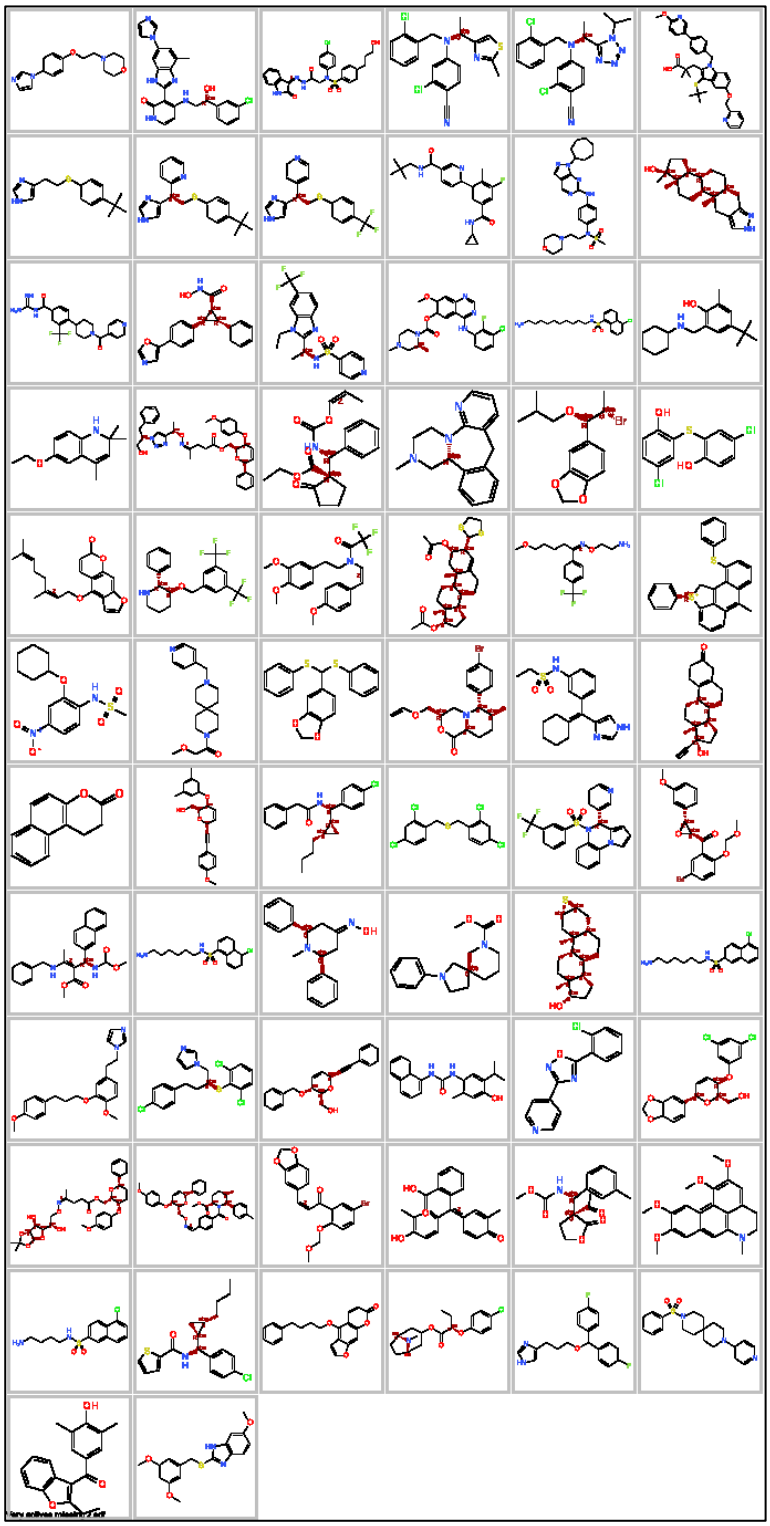


Figure 20: Screenshot of the missing 68 very active molecules

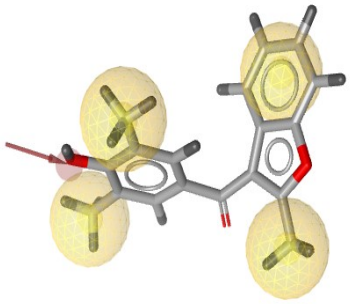


5.3 Structure-Based Modeling

For the structure-based approach experimental data of human microsomal cytochrome P450 CYP 2C19 was required. Here the Protein Data Bank served as the depository of choice, where the search for the cytochrome led to one X-ray structure with a 2.87 Å resolution, derived from Escherichia Coli, in a complex with its co-crystallized inhibitor "0XV502" (Ki=33nM), assigned with the PDB 4-letter code 4GQS (according to Reynald et al. 2012)[28].

The structural data of the identified macromolecule-inhibitor complex was imported into LigandScout, a pharmacophore model of this competitive inhibitor was created and used to query the virtual library. With only one H-bond acceptor and four hydrophobic features the pharmacophore model derived was found unlikely to be selective, as the following virtual screening finally proved: **Table 64** provides the screening results of this structure-based model. **Figure 22** and **Figure 23** demonstrate the given pharmacophoric features and the key interactions between the molecule's functional groups and the amino acids in the binding site of this enzyme. As described in section 2.3.2, the hydroxyl group of phenyl formed a H-bond with a water molecule, while the other moieties interacted with the amino acids in a hydrophobic or aromatic way.

Figure 21 & 22: Pharmacophoric features of the inhibitor "0XV502" and the amino key interactions



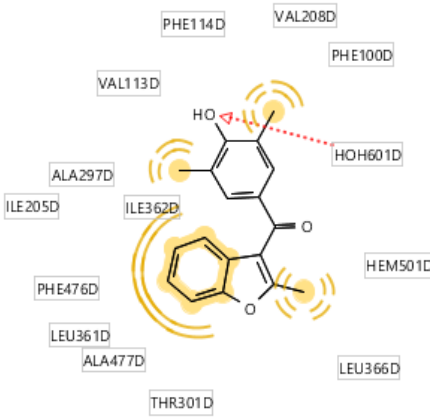
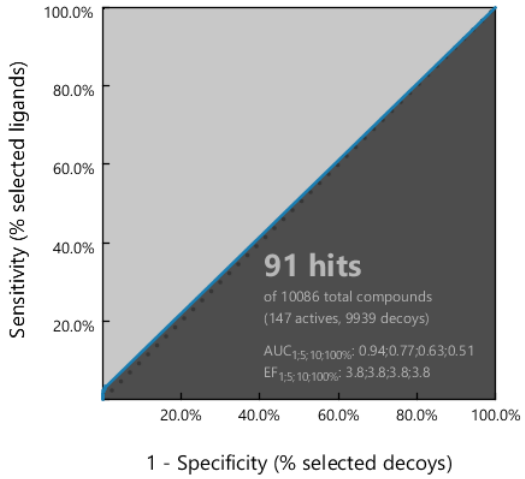


Table 54: Screening results of the structure-based pharmacophore model



True Positives	5
False Positives	86
Sensitivity (100%)	3.40%
Specificity (100%)	99.14%

5.4 Discussion

CYP450 2C19 contributes to the hepatic elimination of many clinically used drugs, so its inhibition through drug-drug interaction is always an issue that should be avoided.[3] For this reason, it is an essential step in drug development to screen new investigational drug molecules for their potential to mediate CYP inhibition.[6] This present study provides 13 pharmacophore models, which have the purpose to identify possible 2C19 inhibitors in the very early stage of pharmaceutical research.

Two approaches have been followed to create predictable pharmacophores: ligand-based and structure-based modeling. Because there was only one PDB entry of a CYP2C19 enzyme-inhibitor complex, our structure-based pharmacophore model derived from only one inhibitor. Due to its simple four hydrophobic and one H-bond acceptor features, it was an unselective pharmacophore model of low-quality (**Table 64**) with no applicability, so the further focus of this discussion remains on the ligand-based pharmacophore models.

The thorough selection of the most potent reversible inhibitors in the low nanomolar range (1-99.9nM) from ChEMBL was the most important step, as these molecules served on the one hand as the foundation of our ligand-based pharmacophore models and on the other hand, as our active virtual library for model evaluation. The division of these compounds into three different activity subsets should lead to pharmacophore models with a descending selectivity, when generated from structures of a subset exhibiting weaker activities. This was not the case, since model 1.1 (7 true actives : 8 decoys), which was built of the strongest inhibitors (1-9.9nM) did not necessarily perform better than for instance model 3.2 (10 true actives : 9 decoys), that was built of less potent inhibitors (50-99.9nM). This can be explained if one takes a closer look at the pharmacophoric features of LigandScout and their depiction as spheres. By having two molecules, just a small difference in their structure, for example a slight discrepancy in the location of a H-bond acceptor, may result in a much higher affinity to the receptor of one of them. Given that the spheres in LigandScout are having distance constraints of 1-6 Ångström, they might bias this little difference and consequently the affinity of the molecules and moreover the selectivity of the models.[62]

Nonetheless, the aim of the ligand-based approach in the first line was to cover all 147 very active structures. But despite of countless refinements, the models were able to hit only 79 (53,74%) of the active virtual library. The reason for this circumstance was that many of the retrieved molecules could be divided into subgroups with analogous structures and therefore similar pharmacophoric features, while most of the missing compounds shared not enough structural similarities with the obtained ones, hence could not be detected by the pharmacophore models. It also resembles the non-specific nature of the CYP450's, which metabolize a wide range of substrates with different structures to protect the organism at its best.[16]

The structural properties of the 147 very active molecules are perfectly displayed in **Figure 19** and **Figure 20**. **Figure 19** (**page 65**) represents a depiction of our 79 retrieved molecules, where a clear relation between compounds, which could be assigned to a subgroup, was seen. In contrast, most of the 68 molecules depicted in **Figure 20** were just too different regarding their size, their functional groups, and hence their pharmacophoric features.
An attempt to create additional selective pharmacophore models of these inhibitors had nevertheless been made, however, the following problem was encountered:

Clustering these molecules resulted in too many clusters that consisted of only one molecule. So only those with two or more molecules could be used for the generation of new pharmacophore models. Since there were no specific structural similarities between most of these structures, they were reduced to their unspecific features like hydrophobic interactions. This resulted in very unspecific virtual screening results with no possible refinement.

Figure 20 (**page 66**) shows also molecules, which might have fitted well into our pharmacophore models and have been retrieved by the initial pre-refinement model in the first hand. This can be explained by the used tuning process: Many of our pharmacophore models were refined by adding exclusion volumes to exclude decoys, which had the correct alignment of necessary features, but had too large molecule structures or exhibited steric clashes. Unfortunately, this led also to the exclusion of some true positives with for example a substituent in an unfavourable position. It often had to be decided, whether the models should rather be selective or unselective with a higher rate of true-, but also false positives. In the end the decision was made in favour of the selective ones, as only 32 of 9939 (0,32%) decoys were recognized as actives. This result implicates the applicability of the models in real world settings, where always a much smaller number of true actives is embedded into a high number of decoys, making specificity to a property of major importance in virtual screening.[68]

Furthermore, it is advisable to apply the models not as individual- but as complementary units to achieve the most satisfying screening results, regardless of the few overlaps, which are almost inevitable by using 13 pharmacophores. Of course, a 100% certainly discrimination between inhibitors and non-inhibitors will never be possible due to the already mentioned substrate diversity of CYP2C19. But despite of this fact, the models together were anyway able to cover more than 50% of the most potent molecules of the ChEMBL database proving their sensitivity and specificity in virtual screening.

In conclusion, the goal to create pharmacophore models that may be used to identify reversible inhibitors of CYP2C19 can be considered as achieved and in the future these 13 pharmacophores may be applied for this purpose in the early stage of drug discovery.

6. Abbreviations

Å	Ångström
AC50	Half maximal activity concentration
ADIT	AutoDep Input Tool
AhR	Aryl hydrogen receptor
AR	Aromatic ring
AUC	Area under the curve
BNL	Brookhaven National Laboratories
CAR	Constitutive androstane receptor
ChEMBLdb	ChEMBL database
CoMFA	Comparative Molecular Field Analysis
CYP450	Cytochrome P450
E	Enrichtment factor
EMBL	European Molecular Biology Laboratory
FEB	Iron binding location
FN	False negatives
FP	False positives
GBPM	GRID-based pharmacophore model
GST	Glutathione S-transferases
H	Hydrophobic interaction
H-bond	Hydrogen Bond
HBA	Hydrogen bond acceptor
HBD	Hydrogen bond donor
IC50	Half maximal inhibitory concentration
ID	Identifier
INH	Inhibition
IUPAC	International Union of Pure and Applied Chemistry
Ki	Inhibitory constant
KNIME	Konstanz Information Miner
LDB	LigandScout Database
MOE	Molecular Operating Environment
MGB	Magnesium binding location
MNB	Manganese binding location
NADPH	Nicotinamide adenine dinucleotide phosphate
NI	Negative Ionizable area
Nm	Nanometer
NMR	Nuclear magnetic resonance
PDB	Protein Data Bank
PI	Positive ionizable area
PPI	Proton pump inhibitor
PXR	Pregnane X receptor
RCSB	Research Collaboratory for Structural Bioinformatics
RDF	Resource Description Framework

ROC	Receiving operator characteristic
SDF	Structur-Data File
Se	Sensitivity
SMILES	Simplified Molecular Input Line Entry Specification
Sp	Specificity
SULT	Sulfotransferases
TN	True negatives
TP	True positives
UGT	UDP-glucuronosyltransferases
WwPDB	Worldwide Protein Data Bank
XBD	Halogen bond donor
Ya	Yield of actives
ZNB	Zinc binding location

7. References

1 Manikandan, P., & Nagini, S. (2017). Cytochrome P450 Structure, Function and Clinical Significance: A Review. *Current Drug Targets*, *19*(1), 38–54

2 Lynch, T., & Price, A. (2007). The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *American Family Physician*, *76*(3), 391–396

3 Doan, J., Zakrzewski-Jakubiak, H., Roy, J., Turgeon, J., & Tannenbaum, C. (2013). Prevalence and Risk of Potential Cytochrome P450–Mediated Drug-Drug Interactions in Older Hospitalized Patients with Polypharmacy. *Annals of Pharmacotherapy*, *47*(3), 324–332

4 Lin, J. H., & Lu, A. Y. (1997). Role of pharmacokinetics and metabolism in drug discovery and development. *Pharmacological Reviews*, *49*(4), 403–449

5 Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T., & Vianello, R. (2005). MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *Journal of Medicinal Chemistry*, *48*(22), 6970–6979.

6 de Groot, M. J. (2006). Designing better drugs: predicting cytochrome P450 metabolism. *Drug Discovery Today*, *11*(13–14), 601–606

7 Hasler, J. A., Estabrook, R., Murray, M., Pikuleva, I., Waterman, M., Capdevila, J., … Beaune, P. (1999). Human cytochromes P450. *Molecular Aspects of Medicine*, *20*(1–2), 1–137

8 Anzenbacher, P., & Anzenbacherová, E. (2001). Review: Cellular and Molecular Life Sciences Cytochromes P450 and metabolism of xenobiotics. *Cellular and Molecular Life Sciences*, *58*, 737–747

9 Sim, S. C., & Ingelman-Sundberg, M. (2010). The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. *Human Genomics*, *4*(4), 278–281

10 NEBERT, D. W., ADESNIK, M., COON, M. J., ESTABROOK, R. W., GONZALEZ, F. J., GUENGERICH, F. P., … WATERMAN, M. R. (2009). The P450 Gene Superfamily: Recommended Nomenclature. *Dna*, *6*(1), 1–11

11 Nelson DR, Koymans L, Kamataki T, et al. P450 superfamily: Update on new sequences, gene mapping, accession numbers and nomenclature. Pharmacogenetics 1996; 6(1): 1-42

12 Johnson, E. F., & Stout, C. D. (2013). Structural diversity of eukaryotic membrane cytochrome P450s. *Journal of Biological Chemistry*, *288*(24), 17082–17092.

13 Werck-Reichhart, D., & Feyereisen, R. (2000). Cytochromes P450: a success story. *Genome Biology*, 1(6):reviews3003.1–3003.9

14 Nair, P. C., McKinnon, R. A., & Miners, J. O. (2016). Cytochrome P450 structure–function: insights from molecular dynamics simulations. *Drug Metabolism Reviews*, *48*(3), 434–452

15 Meunier, B., Visser, P. De, & Shaik, S. (2004). *Mechanism of Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes*. *Chemical Reviews, 2004, 104(9), 3947-3980*

16 Isin, E. M., & Guengerich, F. P. (2007). Complex reactions catalyzed by cytochrome P450 enzymes. *Biochimica et Biophysica Acta - General Subjects*, Vol. 1770, pp. 314–329

17 Peter, G. F. (2007). Mechanisms of cytochrome P450 substrate oxidation: MiniReview. *Journal of Biochemical and Molecular Toxicology*, *21*(4), 163–168

18 Chang, G. W. M., & Kam, P. C. A. (1999). The physiological and pharmacological roles of cytochrome P450 isoenzymes. *Anaesthesia*, *54*(1), 42–50.

19 Parkinson, Andrew & Ogilvie, Brian. (2007). Biotransformation of Xenobiotics. Chapter 6, 133-224.

20 Wijnen, P. A. H. M., Op Den Buijsch, R. A. M., Drent, M., Kuipers, P. M. J. C., Neef, C., Bast, A., … Koek, G. H. (2007). Review article: The prevalence and clinical relevance of cytochrome P450 polymorphisms. *Alimentary Pharmacology and Therapeutics*, *26*(SUPPL. 2), 211–219.

21 Zanger, U. M., & Schwab, M. (2013). Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology and Therapeutics*, *138*(1), 103–141.

22 Meyer, U. A. (2000). Pharmacogenetics and adverse drug reactions. *The Lancet, 356*, 1667–1671.

23 Ahmed, S., Zhou, Z., Zhou, J., & Chen, S. Q. (2016). Pharmacogenomics of Drug Metabolizing Enzymes and Transporters: Relevance to Precision Medicine. *Genomics, Proteomics and Bioinformatics*, *14*(5), 298–313.

24 Lin,J H, & Lu,A Y. (1998). Inhibition and induction of cytochrome P450 and the clinical implications. *Clinical Pharmacokinetics*, *35*(5), 361–390.

25 Tanaka, E. (1998). Clinically important pharmacokinetic drug-drug interactions: Role of cytochrome P450 enzymes. *Journal of Clinical Pharmacy and Therapeutics*, *23*(6), 403–416.

26 Pelkonen, O., Turpeinen, M., Hakkola, J., Honkakoski, P., Hukkanen, J., & Raunio, H. (2008). Inhibition and induction of human cytochrome P450 enzymes: Current status. *Archives of Toxicology*, *82*(10), 667–715.

27 Hollenberg, P. F. (2002). Characteristics and common properties of inhibitors, inducers, and activators of CYP enzymes. *Drug Metabolism Reviews*, *34*(1–2), 17–35.

28 Reynald, R. L., Sansen, S., Stout, C. D., & Johnson, E. F. (2012). Structural Characterization of Human Cytochrome P450 2C19. *Journal of Biological Chemistry*, *287*(53), 44581–44591.

29 Desta, Z., Zhao, X., Shin, J., & Flockhart, D. A. (2002). Clinical significance of the Cytochrome P450 2C19 Genetic Polymorphism. *Drug Disposition, 41*(12), 913–958.

30 Williams, P. A., Cosme, J., Ward, A., Angove, H. C., Vinković, D. M., & Jhoti, H. (2003). Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature*, *424*(6947), 464–468.

31 Schoch, G. A., Yano, J. K., Wester, M. R., Griffin, K. J., Stout, C. D., & Johnson, E. F. (2004). Structure of human microsomal cytochrome P450 2C8: Evidence for a peripheral fatty acid binding site. *Journal of Biological Chemistry*, *279*(10), 9497–9503

32 Suzuki, H., Kneller, M. B., Rock, D. A., Jones, J. P., Trager, W. F., & Rettie, A. E. (2004). Active-site characteristics of CYP2C19 and CYP2C9 probed with hydantoin and barbiturate inhibitors. *Archives of Biochemistry and Biophysics*, *429*(1), 1–15.

33 Foti, R. S., Rock, D. A., Han, X., Flowers, R. A., Wienkers, L. C., & Wahlstrom, J. L. (2012). Ligand-based design of a potent and selective inhibitor of cytochrome P450 2C19. *Journal of Medicinal Chemistry*, *55*(3), 1205–1214.

34 Locuson, C. W., Suzuki, H., Rettie, A. E., & Jones, J. P. (2004). Charge and substituent effects on affinity and metabolism of benzbromarone-based CYP2C19 inhibitors. *Journal of Medicinal Chemistry*, *47*(27), 6768–6776.

35 Hirota, T., Eguchi, S., & Ieiri, I. (2012). Impact of Genetic Polymorphisms in CYP2C9 and CYP2C19 on the Pharmacokinetics of Clinically Used Drugs. *Drug Metabolism and Pharmacokinetics*, *28*(1), 28–37.

36 Li, Y., Tang, H. L., Hu, Y. F., & Xie, H. G. (2012). The gain-of-function variant allele CYP2C19*17: A double-edged sword between thrombosis and bleeding in clopidogrel-treated patients. *Journal of Thrombosis and Haemostasis*, *10*(2), 199–206.

37 Baldwin, R. M., Ohlsson, S., Pedersen, R. S., Mwinyi, J., Ingelman-Sundberg, M., Eliasson, E., & Bertilsson, L. (2008). Increased omeprazole metabolism in carriers of the CYP2C19*17 allele; a pharmacokinetic study in healthy volunteers. *British Journal of Clinical Pharmacology*, *65*(5), 767–774.

38 Saito, Y., Stamp, L. K., Caudle, K. E., Hershfield, M. S., McDonagh, E. M., Callaghan, J. T., … Lee, M. T. M. (2016). CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clinical Pharmacology and Therapeutics*, *99*(1), 36–37.

39 Flockhart DA. Drug Interactions: Cytochrome P450 Drug Interaction Table. Indiana University School of Medicine (2007). "https://drug-interactions.medicine.iu.edu" Accessed [05/2019].

40 Voet, A., Qing, X., Lee, X. Y., De Raeymaecker, J., Tame, J., Zhang, K., & De Maeyer, M. (2014). Pharmacophore modeling: advances, limitations, and current utility in drug discovery. *Journal of Receptor, Ligand and Channel Research*, 81.

41 Drie, J. H. Van. (2007). Monty Kier and the Origin of the Pharmacophore Concept. *Internet Electronic Journal of Molecular Design - IEJMD*, *6*, 271–279.

42 Wermuth, C. G., Ganellin, C. R., Lindberg, P. & Mitscher, L. a. Glossary for Chemists of Terms Used in Medicinal Chemistry. *Pure Appl. Chem.* **70**, 1129–1143 (1998).

43 Seidel, T., Ibis, G., Bendix, F., & Wolber, G. (2010). Strategies for 3D pharmacophore-based virtual screening. *Drug Discovery Today: Technologies*, *7*(4).

44 Kapetanovic, I. M. (2008). Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. *Chemico-Biological Interactions*, *171*(2), 165–176.

45 Guner, Osman. (2005). The impact of pharmacophore modeling in drug design. *8(7):567-72.*

46 Yang, S. Y. (2010). Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. *Drug Discovery Today*, *15*(11–12), 444–450.

47 Langer, T., & Wolber, G. (2004). Pharmacophore definition and 3D searches. *Drug Discovery Today: Technologies*, *1*(3), 203–207.

48 Wolber, G., Seidel, T., Bendix, F., & Langer, T. (2008). Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today*, *13*(1–2), 23–29.

49 Langer, T. & Hoffmann, R. D. Pharmacophore modelling: applications in drug discovery. *Expert Opin. Drug Discov.* **1**, 261–267 (2006).

50 Kim, K.-H., Kim, N. D. & Seong, B.-L. Pharmacophore-based virtual screening: a review of recent applications. *Expert Opin. Drug Discov.* **5**, 205–222 (2010).

51 Kirchmair, J., Ristic, S., Eder, K., Markt, P., Wolber, G., Laggner, C., & Langer, T. (2007). Fast and efficient in silico 3D screening: Toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *Journal of Chemical Information and Modeling*, *47*(6), 2182–2196.

52 Triballeau, N., Acher, F., Brabet, I., Pin, J. P., & Bertrand, H. O. (2005). Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *Journal of Medicinal Chemistry*, *48*(7), 2534–2547.

53 Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., … Overington, J. P. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(D1), 1100–1107.

54 Gaulton, A., Hersey, A., Nowotka, M. L., Patricia Bento, A., Chambers, J., Mendez, D., … Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, *45*(D1), D945–D954.

55 Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., … Overington, J. P. (2015). ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, *43*(W1), W612–W620.

56 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.

57 P. Mazanetz, M., J. Marmon, R., B. T. Reisser, C. & Morao, I. Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. *Curr. Top. Med. Chem.* **12**, 1965–1979 (2013).

58 Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., … Wiswedel, B. (2009). KNIME - the Konstanz information miner. *ACM SIGKDD Explorations Newsletter*, *11*(1).

59 Sieb, C., Meinl, T., & Berthold, M. (2007). Parallel and distributed data pipelining with KNIME. *The Mediterranean Journal of Computers and Networks*, *3*(2), 43–51.

60 Wolber, G., & Langer, T. (2005). LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *Journal of Chemical Information and Modeling*, *45*(1), 160–169.

61 Langer, T. (2010). Pharmacophores in drug research. *Molecular Informatics*, *29*(6–7), 470–475.

62 LigandScout User Manual 4.4 – Inte:Ligand GmbH (Available by Inte:Ligand GmbH, Vienna, Austria)

63 Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank www.rcsb.org. *Nucleic Acids Research*, *28*(1), 235–242.

64 Fine, J., & Chopra, G. (2019). Lemon: a framework for rapidly mining structural information from the Protein Data Bank. *Bioinformatics*, (March), 1–3.

65 Berman, H., Henrick, K., Nakamura, H., & Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, *35*(SUPPL. 1), 2006–2008.

66 Berman, H. M., Bluhm, W. F., Philip, E., Marvin, J., Weissig, H., & John, D. (2002). The Protein Data Bank. *Acta Cryst D58*, 899–907.

67 Rose, P. W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., … Burley, S. K. (2017). The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*, *45*(D1), D271–D281.

68 Réau, M., Langenfeld, F., Zagury, J. F., Lagarde, N., & Montes, M. (2018). Decoys selection in benchmarking datasets: Overview and perspectives. *Frontiers in Pharmacology*, *9*(JAN).

8. Appendix

Hit list of the 13 ligand-based models sorted by pharmacophore-fit score (compounds 1-37)

	Mark	Name	#	Matching Features	Pharmacophore-Fit Score	Best Match	Active/Decoy	Bioactivity Type	Activity Value	Activity Units
1	☐	CHEMBL2036213	28		138.25	Model 4.1	active	IC50	51.00	nM
2	☐	CHEMBL2036210	27		137.27	Model 4.1	active	IC50	38.00	nM
3	☐	CHEMBL808	24		98.68	Model 2.3	active	IC50	40.00	nM
4	☐	CHEMBL1355644	56		98.53	Model 1.1	active	Potency	3.20	nM
5	☐	CHEMBL1436169	38		98.53	Model 1.1	active	Potency	7.90	nM
6	☐	CHEMBL566899	41		98.50	Model 1.1	active	Potency	12.60	nM
7	☐	CHEMBL1554721	63		98.47	Model 1.1	active	Potency	12.60	nM
8	☐	CHEMBL91	52		98.47	Model 2.3	active	Potency	25.10	nM
9	☐	CHEMBL1574390	100		97.86	Model 4.4	decoy	AC50		
10	☐	CHEMBL1315820	61		97.75	Model 4.4	active	Potency	25.10	nM
11	☐	CHEMBL1356336	43		97.57	Model 1.1	active	Potency	7.90	nM
12	☐	CHEMBL1446716	49		97.55	Model 2.3	active	Potency	10.00	nM
13	☐	CHEMBL1327203	40		97.53	Model 2.3	active	Potency	79.40	nM
14	☐	CHEMBL1316808	94		97.52	Model 4.4	decoy	AC50		
15	☐	CHEMBL1437447	42		97.49	Model 4.4	active	Potency	63.10	nM
16	☐	CHEMBL585861	64		97.42	Model 1.1	active	Potency	20.00	nM
17	☐	CHEMBL1221	26		97.40	Model 2.3	active	IC50	16.00	nM
18	☐	CHEMBL1513990	34		97.28	Model 1.1	active	Potency	79.40	nM
19	☐	CHEMBL1480806	39		97.10	Model 1.1	active	Potency	10.00	nM
20	☐	CHEMBL1603852	53		97.06	Model 1.1	active	Potency	50.10	nM
21	☐	CHEMBL1565927	108		97.04	Model 1.1	active	Potency	12589.30	nM
22	☐	CHEMBL522121	90		96.88	Model 1.1	decoy	AC50		
23	☐	CHEMBL1408427	66		96.60	Model 1.1	active	Potency	25.10	nM
24	☐	CHEMBL1359713	65		96.36	Model 2.3	active	Potency	10.00	nM
25	☐	CHEMBL1438688	110		96.19	Model 1.1	decoy	Potency	10000.00	nM
26	☐	CHEMBL1232474	67		96.08	Model 2.3	active	Potency	63.10	nM
27	☐	CHEMBL1262	25		94.10	Model 2.3	active	IC50	40.00	nM
28	☐	CHEMBL2018920	73		88.83	Model 2.2	active	Ki	70.00	nM
29	☐	CHEMBL2018921	74		88.82	Model 2.2	active	Ki	40.00	nM
30	☐	CHEMBL2018917	72		88.82	Model 2.2	active	Ki	40.00	nM
31	☐	CHEMBL2019024	79		88.69	Model 2.2	active	Ki	20.00	nM
32	☐	CHEMBL2018907	69		88.60	Model 3.2	active	Ki	70.00	nM
33	☐	CHEMBL502335	76		88.21	Model 2.2	active	Ki	60.00	nM
34	☐	CHEMBL2018923	75		87.87	Model 2.2	active	Ki	70.00	nM
35	☐	CHEMBL2018908	71		87.77	Model 3.2	active	Ki	50.00	nM
36	☐	CHEMBL2018925	77		87.72	Model 2.2	active	Ki	50.00	nM
37	☐	CHEMBL2018909	70		87.54	Model 3.2	active	Ki	60.00	nM

Hit list of the 13 ligand-based models sorted by pharmacophore-fit score (compounds 38-74)

	Mark	Name	#	Matching Features	Pharmacophore-Fit Score	Best Match	Active/Decoy	Bioactivity Type	Activity Value	Activity Units
38	☐	CHEMBL2018922	81		87.41	Model 3.2	decoy	Ki	50000.00	nM
39	☐	CHEMBL2018904	68		87.36	Model 3.2	active	Ki	60.00	nM
40	☐	CHEMBL296580	78		87.32	Model 2.2	active	Ki	50.00	nM
41	☐	CHEMBL2018906	80		86.98	Model 3.2	decoy	Ki	42500.00	nM
42	☐	CHEMBL2019023	82		85.87	Model 2.2	decoy	Ki	10900.00	nM
43	☐	CHEMBL2312064	111		85.64	Model 3.2	decoy	INH	20.00	uM
44	☐	CHEMBL1323438	107		85.39	Model 2.2	decoy	Potency	15848.90	nM
45	☐	CHEMBL510009	109		85.37	Model 2.2	decoy	Potency	15848.90	nM
46	☐	CHEMBL1552519	97		85.32	Model 2.2	decoy	AC50		
47	☐	CHEMBL565654	105		85.27	Model 2.2	decoy	Potency	10000.00	nM
48	☐	CHEMBL491939	101		85.26	Model 2.2	decoy	AC50		
49	☐	CHEMBL1554577	58		78.70	Model 3.1	active	Potency	50.10	nM
50	☐	CHEMBL1356060	62		78.18	Model 3.1	active	Potency	39.80	nM
51	☐	CHEMBL1316956	33		77.61	Model 3.1	active	Potency	79.40	nM
52	☐	CHEMBL590156	51		77.57	Model 3.1	active	Potency	10.00	nM
53	☐	CHEMBL1314367	48		77.56	Model 3.1	active	Potency	12.60	nM
54	☐	CHEMBL1396261	35		77.56	Model 3.1	active	Potency	79.40	nM
55	☐	CHEMBL1404598	59		77.46	Model 3.1	active	Potency	50.10	nM
56	☐	CHEMBL1592464	37		77.45	Model 3.1	active	Potency	63.10	nM
57	☐	CHEMBL1395089	104		77.25	Model 3.1	decoy	Potency	10000.00	nM
58	☐	CHEMBL489738	45		76.84	Model 3.1	active	Potency	63.10	nM
59	☐	CHEMBL1515449	57		76.63	Model 3.1	active	Potency	10.00	nM
60	☐	CHEMBL1318495	47		76.61	Model 3.1	active	Potency	79.40	nM
61	☐	CHEMBL112532	1		68.32	Model 2.1	active	IC50	7.00	nM
62	☐	CHEMBL358880	7		68.31	Model 2.1	active	IC50	46.00	nM
63	☐	CHEMBL151170	8		68.18	Model 2.1	active	IC50	46.00	nM
64	☐	CHEMBL345544	2		68.12	Model 2.1	active	IC50	68.00	nM
65	☐	CHEMBL358216	5		68.02	Model 2.1	active	IC50	46.00	nM
66	☐	CHEMBL150999	83		68.01	Model 2.1	decoy	IC50	18700.00	nM
67	☐	CHEMBL1370903	55		67.84	Model 4.5	active	Potency	79.40	nM
68	☐	CHEMBL3891461	31		67.84	Model 4.2	active	IC50	30.00	nM
69	☐	CHEMBL1708935	98		67.80	Model 4.5	decoy	AC50		
70	☐	CHEMBL563716	21		67.63	Model 4.6	active	IC50	33.00	nM
71	☐	CHEMBL556891	19		67.54	Model 4.6	active	IC50	33.00	nM
72	☐	CHEMBL551784	15		67.51	Model 4.6	active	IC50	40.00	nM
73	☐	CHEMBL559581	18		67.45	Model 4.6	active	IC50	91.00	nM
74	☐	CHEMBL556892	20		67.44	Model 4.6	active	IC50	33.00	nM

Hit list of the 13 ligand-based models sorted by pharmacophore-fit score (compounds 75-111)

	Mark	Name	#	Matching Features	Pharmacophore-Fit Score	Best Match	Active/Decoy	Bioactivity Type	Activity Value	Activity Units
75	☐	CHEMBL551782	14		67.44	Model 4.6	active	IC50	67.00	nM
76	☐	CHEMBL113227	10		67.43	Model 2.1	active	IC50	46.00	nM
77	☐	CHEMBL1602632	44		67.38	Model 4.3	active	Potency	12.60	nM
78	☐	CHEMBL1320943	36		67.38	Model 4.3	active	Potency	20.00	nM
79	☐	CHEMBL2207500	32		67.34	Model 4.2	active	IC50	30.00	nM
80	☐	CHEMBL151280	4		67.24	Model 2.1	active	IC50	46.00	nM
81	☐	CHEMBL1568369	103		67.21	Model 4.5	decoy	Potency	3162.30	nM
82	☐	CHEMBL3192957	95		67.21	Model 4.3	decoy	AC50		
83	☐	CHEMBL262968	9		67.19	Model 2.1	active	IC50	46.00	nM
84	☐	CHEMBL1950643	85		67.14	Model 2.1	decoy	IC50	30000.00	nM
85	☐	CHEMBL1086579	84		67.11	Model 4.2	decoy	IC50	10000.00	nM
86	☐	CHEMBL552156	22		67.11	Model 4.6	active	IC50	63.00	nM
87	☐	CHEMBL151354	3		67.11	Model 2.1	active	IC50	46.00	nM
88	☐	CHEMBL550170	17		67.10	Model 4.6	active	IC50	33.00	nM
89	☐	CHEMBL559974	23		67.09	Model 4.6	active	IC50	33.00	nM
90	☐	CHEMBL154111	87		67.06	Model 4.5	decoy	IC50		
91	☐	CHEMBL1515034	93		67.04	Model 4.5	decoy	AC50		
92	☐	CHEMBL1556045	88		67.01	Model 4.5	decoy	AC50		
93	☐	CHEMBL112592	12		66.97	Model 2.1	active	IC50	46.00	nM
94	☐	CHEMBL1447528	96		66.92	Model 4.5	decoy	AC50		
95	☐	CHEMBL1565525	54		66.85	Model 4.5	active	Potency	31.60	nM
96	☐	CHEMBL1395275	106		66.80	Model 4.3	decoy	Potency	12589.30	nM
97	☐	CHEMBL151974	6		66.78	Model 2.1	active	IC50	46.00	nM
98	☐	CHEMBL3918756	86		66.73	Model 4.3	decoy	IC50	53000.00	nM
99	☐	CHEMBL1572187	89		66.69	Model 4.5	decoy	AC50		
100	☐	CHEMBL1447147	50		66.69	Model 4.3	active	Potency	50.10	nM
101	☐	CHEMBL559580	16		66.53	Model 4.6	active	IC50	33.00	nM
102	☐	CHEMBL3898132	30		66.50	Model 4.2	active	IC50	30.00	nM
103	☐	CHEMBL115876	11		66.34	Model 2.1	active	IC50	46.00	nM
104	☐	CHEMBL320105	13		65.88	Model 2.1	active	IC50	46.00	nM
105	☐	CHEMBL1321655	46		48.57	Model 2.4	active	Potency	20.00	nM
106	☐	CHEMBL833	60		48.56	Model 2.4	active	Potency	39.80	nM
107	☐	CHEMBL282433	91		48.48	Model 2.4	decoy	AC50		
108	☐	CHEMBL1472052	102		47.34	Model 2.4	decoy	AC50		
109	☐	CHEMBL1444408	92		47.31	Model 2.4	decoy	AC50		
110	☐	CHEMBL3331506	29		46.52	Model 2.4	active	IC50	70.00	nM
111	☐	CHEMBL1354777	99		46.48	Model 2.4	decoy	AC50		