universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Prediction of Competitive CYP450 2C9 Inhibitors Using 3D-Pharmacophore Models“

verfasst von / submitted by

Martin Pramesberger

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
Mitbetreut von / Co-Supervisor:	Dipl.-Ing. Dr. Thomas Seidel

Acknowledgements

First of all, I have to thank Thierry Langer, because my scientific work in his research group in the Department of Pharmaceutical Chemistry would not have been possible without him. As the supervisor of the present work, a large part of the success is devoted to his commitment. Not only Professor Langer's technical knowledge and experience in the field of molecular modeling, but also his enthusiasm and visions made the cooperation a pleasure.

My next special thanks go to Sharon Bryant, who supported me throughout the entire research work. With her great experience and competence in the field of pharmacophore modeling, she was a significant part of the project and was my primary contact person for all arising questions. Her passion for pharmacophore modeling and her motivational nature inspired me from the very beginning and drove me forward throughout the whole research.

Many thanks also go to Thomas Seidel, who not only offered me a working place in Thierry Langer's lab, but also helped me with technical questions and the provision of various computer programs. His impressive expertise in computational science was of great help with various problems that came up during my time in his group.

At this point I would also like to express my thanks to Oliver Wieder, who not only was always prepared to provide me with professional advice and assistance, but also made the months in the lab a great time with amusing conversations and lots of fun. Of course, my thanks go not only to him, but also to all the other members of Thierry Langer's group.

Finally, but most of all, I would like to thank my family and friends, who supported me not only during the time of my diploma thesis but rather during my entire pharmacy study. Thank you very much for motivating me throughout all the years.

Abstract

The isoenzymes of the cytochrome P450 superfamily represent the most important class of enzymes during phase 1 metabolism, responsible for transforming molecules into more hydrophilic forms and thus easier to excrete. In addition to the biotransformation of endogenous substances, they are of great importance for metabolism of exogenous compounds, the so-called xenobiotics, such as drugs. However, drugs are not only substrates of CYP450 enzymes, many can also inhibit these enzymes, which leads to strongly elevated plasma levels of affected CYP450 substrates and thus increased side effects and toxicity can occur. The aim of this thesis was to generate high quality and selective 3D-pharmacophore models, using both structure-based and ligand-based approaches, to be used to predict competitive inhibitors of CYP450 2C9.

First, compounds experimentally assessed at CYP450 2C9 were extracted from the ChEMBL database. Those with biological assay annotations were filtered, cleaned, processed and assembled into training and test datasets to be used not only for predictive model development, but also for testing the performance of the different pharmacophore models generated in this work. The program LigandScout (4.4 Expert) was used for processing and filtering compound libraries, 3D-pharmacophore modeling, virtual screening and activity profiling. In addition, the Inte:Ligand Expert Extensions (v. 1.5.1) provided in the KNIME Analytics Platform (version 3.6.1) was used for creating pipelining workflows to facilitate the processing of large amounts of data. For the structure-based approach, the Protein Data Bank (PDB) was the source of experimentally derived 3D-coordinates from several CYP 2C9 ligand-target complexes.
The results of this work consist of six ligand-based 3D-pharmacophore models, which can be used for the prediction of competitive inhibitors of CYP450 2C9. By using these pharmacophore models, scientists involved in preclinical research can reduce risk of molecule failures at later stages and therefore improve efficiency in terms of costs and time during early drug development, as potentially problematic compounds may

already be sorted out at a very early stage. However, the results of this thesis are not only of great interest for practical applications, but also provide valuable insights into the chemical features responsible for the molecular interaction between competitive inhibitors and the binding site of CYP450 2C9.

Zusammenfassung

Die Isoenzyme der CYP450 Superfamilie stellen die wichtigste Enzymklasse des Phase 1 Metabolismus´ dar. Neben der Biotransformation körpereigener Substanzen sind sie vor allem für den Metabolismus exogener Stoffe, der Xenobiotika, von großer Bedeutung, da diese dadurch hydrophiler und somit leichter ausscheidbar werden. Arzneistoffe stellen hierbei jedoch nicht nur Substrate von CYP450 Enzymen dar, viele von ihnen können diese auch hemmen, was zu stark erhöhten Plasmaspiegeln betroffener CYP450 Substrate und damit zu erhöhten Nebenwirkungen und Toxizität dieser führen kann. Das Ziel der vorliegenden Arbeit war es, qualitativ hochwertige und selektive 3D-Pharmakophor Modelle zu generieren, sowohl strukturbasiert als auch ligandenbasiert, um kompetitive Inhibitoren von CYP450 2C9 vorherzusagen.

Zunächst wurden experimentelle Daten bezüglich Inhibition von CYP450 2C9 aus der ChEMBL-Datenbank extrahiert. Anschließend wurden die aus verschiedenen biologischen Assays stammenden Daten gefiltert, kuriert, prozessiert und in Trainings- und Testdatensätze geteilt. Jene Datensätze sollten in weiterer Folge für die Generierung und Testung verschiedener prädiktiver 3D-Pharmakophormodelle verwendet werden. Das Programm LigandScout (4.4 Expert) wurde zur Verarbeitung und Filterung der diversen Datensätze, zur 3D-Pharmakophormodellierung, zum virtuellen Screening und zur Aktivitätsprofilerstellung verwendet. Darüber hinaus wurden die in der KNIME Analytics-Plattform (Version 3.6.1) bereitgestellten Inte:Ligand Expert Extensions (Version 1.5.1) zum Erstellen von Pipelining-Workflows verwendet, um die Verarbeitung der großen Datenmengen zu erleichtern. Für den strukturbasierten Ansatz stellte die Protein Data Bank (PDB) die Quelle für experimentell abgeleitete 3D-Koordinaten mehrerer CYP 2C9 Ligand-Target-Komplexe dar.

Das Ergebnis der vorliegenden Arbeit stellen sechs ligandenbasierte 3D-Pharmakophor-Modelle dar, welche zur Vorhersagung kompetitiver Inhibitoren von CYP450 2C9 verwendet werden können. Mithilfe jener Pharmakophormodelle können

Wissenschaftler in der präklinischen Forschung die Effizienz in Bezug auf Kosten und Zeit bei der frühen Arzneimittelentwicklung verbessern, da potenziell problematische Verbindungen möglicherweise bereits in einem sehr frühen Stadium aussortiert werden. Die Ergebnisse jener Arbeit sind jedoch nicht nur von großem Interesse für künftige praktische Anwendungen, sondern liefern auch wertvolle Einblicke bezüglich der für die molekulare Wechselwirkung zwischen kompetitiven Inhibitoren und der Bindetasche von CYP450 2C9 verantwortlichen chemischen Features.

Table of contents

1 Introduction

1.1 Motivation

Cytochrome P450 enzymes are of big relevance in the human body, since they are responsible for the production of endogenous substances like cholesterol, steroids, prostacyclins and thromboxane A2.[1]

Furthermore, they are of major importance for their contribution to metabolic processes mainly taking place in the liver. A minimum of 90% of all drugs are metabolized by cytochromes and therefore their influence on pharmacokinetics is immense. Unfortunately, drugs are not only metabolized by those enzymes, many drugs can also inhibit cytochromes, which leads to elevated plasma levels of additionally administered drugs, and therefore dangerous plasma concentrations and adverse effects can occur. In the case of pro-drugs, inhibition of cytochromes leads to less biotransformation into the active drug form and therefore decreased efficacy. In recent years, several drugs were withdrawn from the market because of adverse effects caused by cytochrome inhibition.[2] As patients are frequently prescribed multiple medications, drug-drug interactions by inhibition of cytochromes is a relevant issue and a big problem in medication management.

1.2 Aims

To reduce the risk of unforeseen drug-drug interactions and/or adverse outcomes due to CYP interactions, the pharmaceutical industry now routinely incorporates CYP enzyme assessments during preclinical stages. Similarly, the aim of this study was to create 3D-pharmacophore models, that can properly identify competitive inhibitors of the enzyme cytochrome P450 2C9. Predictive 3D-pharmacophore models offer different interesting and useful applications, that have a big contribution to the improvement of safe and meaningful medication. Probably the most relevant use case in pharma is the stage of early drug discovery, where pharmacophore models can be used to predict competitive inhibitors of CYP450 2C9 via activity profiling and/or virtual screening.[3] Thus, compounds retrieved by those pharmacophore models may

already be dropped from the pipeline before going to clinical trials, therefore saving time and valuable resources on compounds that will fail. In silico models assist pharmaceutical scientists with identifying problematic compounds at an early stage that might otherwise lead to adverse drug-drug interactions in later stages. Of course, this situation does not necessarily mean the end of a corresponding molecule. Moreover, the insights gained through pharmacophore modeling can be used for lead optimization of molecules that have been classified as problematic.[4] Just as potential inhibitors of CYP450 2C9 may be identified and sorted out by 3D-pharmacophore modeling, this technique could also be used to predict substrates of CYP2C9. Due to genetic polymorphisms of CYP enzymes, very strong activity fluctuations of CYP2C9 exist between population groups and single individuals. Because of that, there can be very large differences in plasma concentrations, especially with pure CYP2C9 substrates.[5]

Thus, already in early drug discovery it could be considered whether a drug candidate could possibly be a CYP2C9 substrate and could be sorted out accordingly. Although this approach is not part of this work, it would certainly be important to be considered when understanding adverse outcomes of drugs due to CYP metabolism, and this has already been done in other research projects.[6,7]

Another possible application of such 3D-pharmacophore models can be detecting new and selective competitive inhibitors of CYP450 2C9. Those inhibitors may be combined with drugs having a low bioavailability because of a high, CYP2C9 mediated, first-pass effect. In this way, competitive inhibition of the enzyme leads to reduced first-pass effect of additionally administered cytochrome substrates, and therefore the dose of such a drug can be decreased. But of course, for this use case the inhibitors of cytochrome 2C9 must have acceptable safety profiles and should not hit other human targets in a noteworthy way. Currently, there are no publications regarding those "positive" drug-drug interactions due to CYP2C9. However, this relatively new approach to pharmacokinetic improvement can already be found for important enzymes such as CYP3A4. (according to Kumar *et al.* 1999[8])

2 Background

2.1 Metabolism of xenobiotics

The human body is constantly confronted with foreign substances, the so-called xenobiotics. This not only includes drugs but also industrial chemicals, herbal remedies, nutritional products, pesticides, pollutants and chemicals in food. The reason why these substances enter our organism in the first place is a certain lipophilicity of the molecules, which allows absorption via the gastrointestinal tract or the skin. On the one hand, high lipophilicity of molecules has a positive effect on their absorption, but on the other hand, it makes their elimination from the body much more difficult, as compounds can be reabsorbed easily.[9] In the kidneys, which represent the most important excretory route of drugs, lipophilic molecules are subject to strong passive reabsorption, and because of that, elimination can be strongly decreased. Reduced excretion of a xenobiotic from our body means more time for the substance to influence the organism and, in the worst case, damage it. A shift in polarity from lipophilic to hydrophilic results in an increased renal and biliary elimination of xenobiotics, and therefore, an essential aspect of drug metabolism is the conversion of lipophilic compounds into much more hydrophilic ones.[10]

The liver is clearly the most important site of drug metabolism in the human body. Although metabolism also takes place in other parts of the body, the liver, with its multitude of different enzymes, is the dominant site of biotransformation. In addition to hepatic metabolism, organs such as the lungs, skin, kidneys, placenta and gastrointestinal tract are also involved in drug metabolism to a small extent. These processes taking place outside the liver are therefore referred to as extrahepatic metabolism.[11]

The actual course of drug biotransformation can be divided into two phases. Phase 1 metabolism serves to expose or generate polar functional groups in the molecule such as hydroxy, amino, or thiol groups. The chemical reactions of phase 1 metabolism are e.g., N- and O-dealkylation, aliphatic and aromatic hydroxylation, N- and S-oxidation,

deamination and epoxidation. By far the most important enzymes of this first phase are the enzymes of the cytochrome P450 superfamily. Following the phase 1 reactions, another enzyme-mediated chemical reaction, the so-called phase 2 reaction, takes place. Here the molecules are conjugated with hydrophilic groups such as glucuronic acid, sulphates, acetyl groups, methyl groups, glutathione or amino acids, which further increases the hydrophilicity of the molecules. The enzymes of phase 2 reactions are mainly transferases such as UDP-glucuronosyltransferases (UGTs), sulfotransferases (SULTs), N-acetyltransferases (NATs), glutathione S-transferases (GSTs) and various methyltransferases. In addition to the combination of phase 1 and subsequent phase 2, there is the possibility that molecules undergo only one of these two chemical reactions. This can be the case, for example, if the original molecule already has hydrophilic groups, and thus only phase 2 metabolism is necessary to increase hydrophilicity.[12,13]

The increase in hydrophilicity is probably the most important function of metabolism of xenobiotics, as it enables the body to effectively remove these foreign and potentially dangerous compounds from the organism. Another important point is that biotransformation in most cases already leads to a pharmacodynamically inactivation of drugs. Nevertheless, there are exceptions, and in some cases, metabolism leads to an increase in activity or activates the compounds in the first place. An example of this is the biotransformation of parathion into the highly toxic paraoxon, which represents the conversion of a pharmacologically inactive compound into a potent cholinesterase inhibitor.[14] Since the target of the following research is a cytochrome, namely CYP450 2C9, and due to the fact that the enzymes of the cytochrome P450 superfamily represent the dominant phase 1 enzymes, it is now necessary to describe those enzymes in detail.

2.2 Cytochrome P450 enzymes

To date, 57 functional CYP genes are known in the human genome and the resulting enzymes are of crucial importance for various physiological processes. The isoenzymes of the cytochrome P450 superfamily are by far the most important enzymes during phase 1 metabolism. Seven of the 57 isoenzymes are of particular importance and are responsible for the oxidative metabolism of over 90% of all drugs administered: CYP1A2 (4%), CYP2A6 (2%), CYP2C9 (10%), CYP2C19 (2%), CYP2E1 (2%), CYP2D6 (30%) and CYP3A4 (50%). Drugs can either be metabolized by a specific CYP450 isoenzyme or are the simultaneous substrate of a large number of CYP450s.[15] In addition to metabolism of drugs and various other xenobiotics, the enzymes of the cytochrome P450 superfamily are also relevant in other areas. The synthesis of steroid hormones, metabolism of fat-soluble vitamins and the conversion of polyunsaturated fatty acids are just a few examples of other tasks. Of course, CYP450 enzymes are not only found in mammalian cells, also insects (control of development via hormone biosynthesis or provision of insecticide resistance) and plants (hormone biosynthesis and herbicide degradation) benefit from this large and multifaceted group of enzymes.[16]

2.2.1 **Systematics and nomenclature**

The term P450 stands for "Pigment-450", which means that the pigment (in the case of cytochromes it is a heme) absorbs light at a maximum of 450 nm if it is present in a reduced state together with carbon monoxide. What is unusual here is that all other known heme-containing proteins (e.g., hemoglobin) absorb light at 420 nm. This phenomenon is due to the identity of the fifth ligand with the heme formed by a cysteine residue in P450s, as opposed to a fifth histidine ligand with proteins such as hemoglobin.[17]

The currently most used nomenclature for cytochrome classification was recommended by Nebert *et al.* in 1987 [18]. The prefix CYP applies to all species (except mice, here it is called Cyp) and defines that it is an enzyme of the cytochrome P450 superfamily. Cytochromes can be divided into families and subfamilies based on the agreement of their amino acid sequences. If cytochrome enzymes have a sequence identity of >40%, they belong to the same family. In the current nomenclature, the

family to which an isoenzyme belongs is identified by an Arabic number immediately after the abbreviation CYP (e.g., CYP2C9 belongs to family 2). If there is a sequence identity of >55% between cytochrome isoenzymes, the enzymes do not only belong to the same family but are also in the same subfamily. The membership to a certain subfamily is represented by a Latin letter immediately after the number for the family. (e.g., CYP2C9 belongs to subfamily C within family 2). An individual enzyme encoded by a gene is finally defined by a further Arabic number after the subfamily specification. For two CYP enzymes to be considered as different isoenzymes, they must differ in their amino acid sequence identity by at least 3%. This nomenclatural code, e.g., CYP2C9, is therefore unique for each cytochrome isoenzyme and allows quick assignment to the family and subfamily to which the enzyme belongs.[19]

2.2.2 **Localization and structure**

Basically, CYP450 enzymes are present in all tissues of the human body: Intestine, lung, kidney, brain, adrenal gland, gonads, heart, nasal and tracheal mucosa and skin. By far the largest abundance of CYP450 enzymes is found in the liver, where they make up about 2% of all microsomal proteins, which is much more than in other tissues. The extrahepatic part of CYP enzymes in metabolism is of local importance, whereas activity in the liver is essential for systematic metabolic clearance of drugs. In addition to the liver, CYP450 enzymes also play an important role in the epithelial cells of the gastrointestinal tract, particularly in the small intestine, where they strongly influence the bioavailability of xenobiotics.[20] Table 1 shows the localization of the most relevant CYP450 enzymes and some typical substrates.

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.

Table 1. Human CYP450 enzymes, localization and typical substrates taken from Anzenbacher *et. al*, 2001).[21]

Eukaryotic P450 enzymes are found in the cells of plants, fungi and animals, where they are bound inside to the membrane of the endoplasmic reticulum. In animal cells, CYP450 enzymes are also located on the inner membrane of mitochondria. The attachment of the enzyme to the cell membrane is achieved by an N-terminal helix, which acts as an anchor.[22] Since numerous X-ray structures of various isoenzymes are available, it is evident that all CYP450 enzymes have a certain basic structure in common. They all consist of approximately 500 amino acids, which form a single polypeptide chain.[6] There is a large alpha-helical domain consisting of the twelve helices A-L and a smaller beta-sheet domain containing the beta-sheets 1-4, which is demonstrated by Figure 1.[23] The three parallel helices D, L and I together with the antiparallel helix E form a four-helix bundle, which represents the conserved P450 structural core. The prosthetic group, the heme, is located between helix I and L and is connected non-covalently to the sulfur atom of an adjacent cysteine residue via its iron

atom. This amino acid residue represents the fifth ligand of the heme iron and is responsible for the characteristic and name-giving absorption maximum of cytochrome P450 enzymes at 450 nm. Since the heme is essential for the catalytic function of CYP450 enzymes, it is part of the binding site.[24]

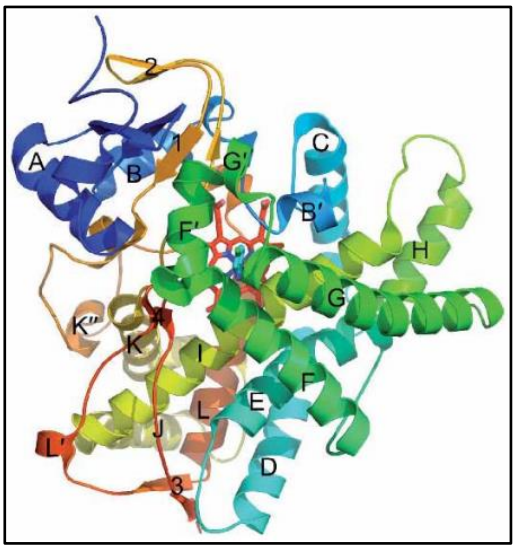


Figure 1. The typical tertiary structure of P450 enzymes based on the example of the CYP2B6-4-CPI complex (PDB ID: 3IBD). The protein consists of twelve alfa-helices (A-L) and four beta-sheets (1-4), with each secondary-structural element highlighted with a different color.

Figure taken from Gay, S.C. *et al.*, 2010.[23]

2.2.3 **Catalytic cycle**

The general function of cytochrome P450 isoenzymes is the oxidative metabolism of various endogenous and exogenous substances during phase 1 metabolism. They catalyze the regiospecific and stereospecific oxidation of substrates' non-activated hydrocarbons at body temperature, a process that would require very high temperatures if not catalyzed. The catalytic center of all cytochrome P450 enzymes is iron protoporphyrin IX (heme) and in most cases these enzymes serve as monooxygenases and therefore catalyze the hydroxylation of various substrates.[25] Simplified it can be said that this oxidation process consists of four steps, which are shown in Figure 2. First, a substrate binds to the active site of the enzyme, which is associated with the displacement of the sixth ligand of the heme, a distal water molecule. In the second step, an electron is picked up and thus the ferric (Fe3+) heme

is reduced to a ferrous (Fe2+) state. In the third step of the catalytic circle, molecular oxygen binds to the reduced iron and in the fourth step, again a 1-electron reduction takes place. Now some less well-defined steps follow, namely a protonation that leads to a hydroperoxo-ferric complex and the following scission of the O-O bond under water release. This scission leads to the formation of an active perferryl FeO species, which then reacts with the substrate. Once the reaction is complete, the product is released, closing the catalytic cycle.[26,27]

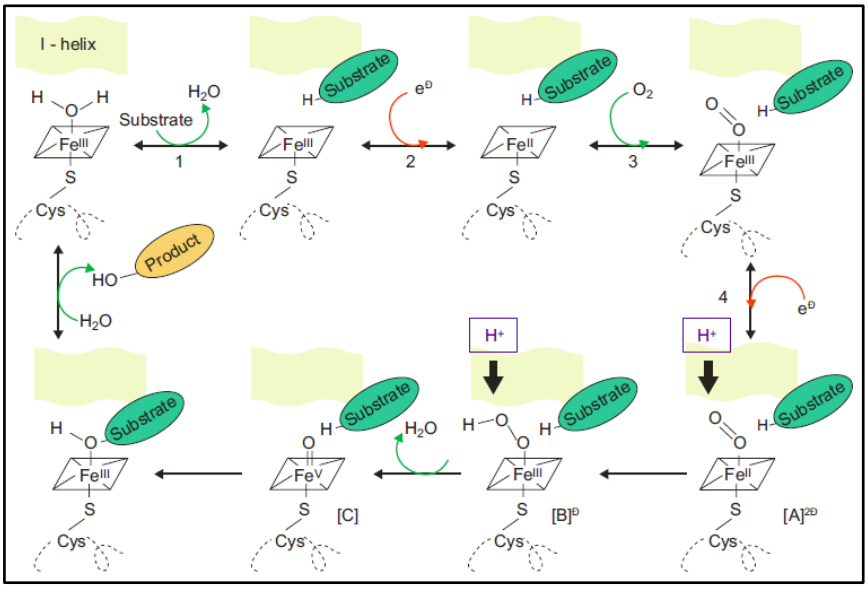


Figure 2. Catalytic mechanism of cytochrome P450 enzymes. Figure taken from Werck-Reichhart, D. *et al.*, 2000.[26]

Each time this corresponding process is to take place, two electrons must be provided somehow. Normally NADPH (nicotinamide adenine dinucleotide phosphate) is responsible for this, although in some cases the second electron can also be provided by cytochrome b5. But in most cases NADPH provides those two electrons, which are transferred to CPR in an initial step. CPR (NADPH-cytochrome P450 reductase) is a flavoprotein which, like the cytochrome P450 enzymes, is bound to the membrane of the endoplasmic reticulum by hydrophobic N-terminal sequences and it consists of two cofactors, FAD (flavin adenine dinucleotide) and FMN (flavin mononucleotide). After the binding of NADPH to CPR, the two electrons are first transferred to the cofactor FAD, which leads to a conformational change of the enzyme. This structural change of

CPR brings FMN and FAD closer together, allowing the electrons to be transferred from FAD to FMN. After electron transfer onto FMN, these two electrons are delivered one at a time at two distinct steps onto a cytochrome P450 enzyme at the respective points of the reaction cycle.[28–30]

2.2.4 **Inhibition of CYP450 enzymes**

Inhibition of a particular CYP450 enzyme means that a substrate that is normally metabolized by that enzyme can no longer be biotransformed for the duration of enzyme inhibition. With respect to drugs, this means that reduced metabolism of a CYP substrate leads to reduced elimination of the affected drug and an increase in plasma concentration. Depending on the intensity and mechanism of enzyme inhibition and also on the therapeutic range of the drug affected by inhibition, increased side effects and toxic reactions may occur. In general, inhibition of CYP450 enzymes can be divided into reversible and irreversible (mechanism-based inhibition) inhibition, whereby correct distinction is often difficult, which leads to another category, namely quasi-irreversible inhibition. By far the most common cause of drug-drug interactions is reversible inhibition of enzymes and this category can be further divided into competitive, non-competitive, uncompetitive and mixed forms of inhibition.[31]

Although the distinction and definition of these different forms is quite clear in theory, in practice it is usually difficult to distinguish between them. Reversible inhibition of CYP450 enzymes, unlike irreversible and quasi-irreversible inhibition, is only dose-dependent and normal metabolic enzyme function is resumed after elimination of the reversible inhibitor. In competitive enzyme inhibition, the binding of the inhibitor to the active site of the enzyme leads to a substrate not being able to bind to it and being metabolized. Often an inhibitor has a similar structural scaffold or interaction features to those of a substrate, which explains its binding to the active site. Of course, competitive CYP inhibitors can also be the substrate of the same enzyme, but this does not necessarily have to be the case. Depending on the complexity of competitive inhibitors and the specificity of the enzyme binding site, it is possible that only one specific isoenzyme is inhibited or that several different isoenzymes are affected.[32] In

non-competitive enzyme inhibition, the inhibitor does not bind to the active site of the enzyme, but to another site. This has no effect on the binding of the substrate but leads to the reduction of activity of later formed enzyme-substrate-inhibitor complex that is not able to function correctly. In contrast to non-competitive inhibition, an uncompetitive inhibitor does not bind to the free enzyme, but to the enzyme-substrate complex, which in turn leads to an unproductive enzyme performance. Mixed forms of enzyme inhibition are poorly defined and difficult to classify. However, it can be said that mixed type inhibitors simultaneously inhibit the enzyme in several ways, particularly combining competitive and non-competitive mechanisms.[32,33]

Irreversible inhibition of CYP450 enzymes leads to the formation of reactive metabolites, which either form stable complexes with the binding site of the enzyme or bind covalently to the heme or residues of the binding site. The result of this mechanism-based inhibition is a long-lasting irreversible blockade of the enzyme, which can only be ended by re-synthesis (suicidal substrate). Characteristic of this type of enzyme inhibition is that, unlike competitive inhibition, mechanism-based inhibition is not only dose-dependent, but also time-dependent.[34]

2.2.5 **Induction of CYP450 enzymes**

Just as CYP450 enzymes can be inhibited, there is also the possibility of induction of this enzyme group. In most cases, the induction of CYP450 enzymes leads to increased metabolism of xenobiotics, which is a protective mechanism. Even though this mechanism is a protective strategy for the body, CYP induction leads to significantly lower plasma concentrations of administered drugs and may therefore lead to a lack of pharmacological effects of drug therapy. However, it must be considered that CYPs, as already mentioned, can also lead to the activation of drugs and they play an important role in the production and oxidative metabolism of endogenous agents. Thus, the induction of cytochromes not only leads to increased degradation of drugs but also to certain disorders in the endogenous system. Furthermore, the formation of carcinogenic, mutagenic or cytotoxic xenobiotic metabolites can be increased in this way. Ligands that act as inducers bind to transcription factors within the cell and thus activate them. Important transcription factors such as pregnane X receptor (PXR),

constitutive androstane receptor (CAR) or aryl hydrocarbon receptor (AhR) lead to increased transcription of CYP450 genes and synthesis of mRNA. This results in an increase of CYP450 enzyme production via the genetic pathway.[35,36]

Figure 3 illustrates the process of CYP450 induction (in this case CYP3A4) via the transcription factor PXR. The lipophilic xenobiotic enters the cell and binds to PXR, whereupon this transcription factor is activated. The complex enters the cell nucleus and binds together with RXR (retinoid X receptor) to a corresponding recognition site of the DNA, which increases the transcription of the CYP3A4 gene and subsequently the translation and production of the enzyme.[33]

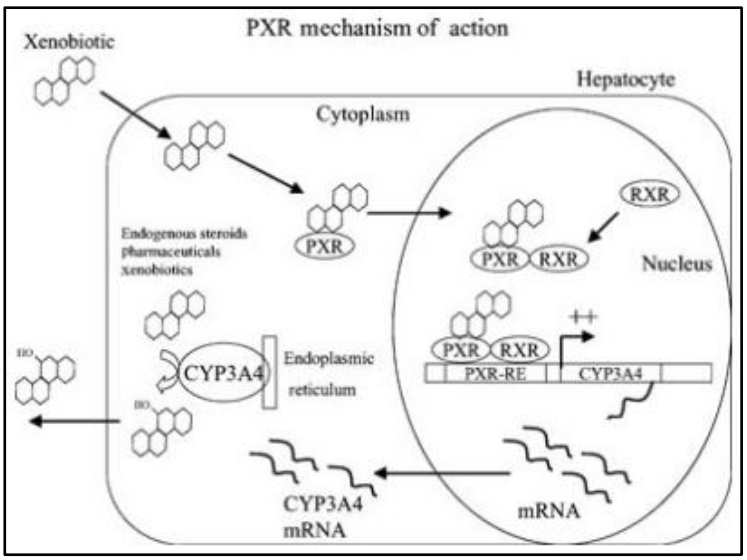


Figure 3. Scheme illustrating the induction of a CYP450 enzyme (CYP3A4) via the activation of PXR. Figure taken from Tompkins, L. M. *et al.*, 2007.[35]

2.3 Cytochrome P450 2C9

The genes for the enzymes of the CYP2C subfamily are located on chromosome 10 and code for the four isoenzymes CYP2C8, CYP2C9, CYP2C18 and CYP2C19, with CYP2C9 being by far the most relevant of these. After CYP3A4, CYP2C9 is the most common CYP enzyme in the liver and is responsible for phase 1 metabolism of 10-15% of all clinically relevant drugs, making this isoenzyme among the most important CYP450 enzymes in the human body.[4]

2.3.1 **Structure and binding site**

The two-domain protein CYP2C9 is composed of the classical basic structure of CYP450 enzymes, already explained in section 2.2.2.

An important structural part of CYP2C9 is the B-C loop, which represents an area of the active site and makes a major contribution to substrate specificity. A good example for the relevance of this area is the amino acid Phe114: The mutation Phe114Ile leads to the complete loss of diclofenac hydroxylation or affinity of the strong competitive inhibitor sulfaphenazole. The heme of CYP2C9 is located between helices I and L and the iron atom is penta-coordinated with Cys435.[6] Stabilization of the heme is provided by hydrogen bonds (H-bonds) between the propionyl residues of the heme and the side chains of some amino acids (Arg97, Trp120, Arg124, His368, Arg433). Contrary to earlier assumptions, the Arg97 residue does not play a role in the interaction with ligands but is of great importance for the stabilization of the heme, which has been shown by several mutagenesis studies (Ridderström *et al.,* 2000[37]).[38]

Within the binding site, the amino acid residue Arg108 has been regarded as the most relevant interaction partner between substrate molecules and CYP2C9. Electrostatic interactions or H-bonds between an enzyme and small lipophilic anions always seemed to be the crucial point of substrate selectivity. Curiously, in the first X-ray structure (PDB ID: 1OG5) of human CYP2C9 (Williams *et al.,* 2003[38]), with (S)-warfarin bound at a resolution of 2.55 Å, no interaction between the ligand and Arg108 was observed. Actually, Arg108 side-chain was oriented away from the active site and was therefore not available for interaction with substrates. Furthermore, in this first X-ray structure the substrate (S)-warfarin was positioned relatively far away (10Å) from the heme. A later X-ray structure of human CYP2C9 (Wester *et al*., 2004[39]) together with flurbiprofen (PDB ID: 1R9O) at a resolution of 2.0 Å clearly showed an interaction with Arg108. The conflicting results from 2003 (complex with (S)-warfarin) were viewed critically from this point onward. It is quite possible that the orientation of Arg108 away from the binding site in the publication from 2003 was caused by significantly more mutations in the enzyme. At this point it must be noted that no electron density map is available on the Protein Data Bank (PDB) for the crystal structure of the 1OG5

complex. From this point on it was finally possible to explain the origin of substrate specificity for negatively charged molecules, as this second model from 2004 clearly showed the interaction with the basic amino acid Arg108. Using X-ray structures of CYP2C9 (from rabbits) it became clear that there were often several different binding modes for substrates of CYP enzymes, which in turn would explain the two different binding modes of (S)-warfarin and flurbiprofen.[4] With a volume of 987 Å³ to 1271 Å³ (depending on the ligand in the binding site), CYP2C9 has a very large substrate binding pocket. However, this fact makes it difficult to understand why this enzyme has such a high selectivity for small substrate molecules, but not for larger ones.[23] In the X-ray structure of Williams *et al*., the substrate molecule (S)-warfarin is positioned relatively far away from the heme. This unusual placement of the ligand in the binding site suggests that this distal location might be a so-called loading site, and the substrate molecules would move to the actual active site after conformational changes of the enzyme. If this were indeed the case, there would be an explanation for this paradoxical selectivity for small substrate molecules within such a large binding site.[23]

2.3.2 **Substrates of CYP450 2C9**

Basically, CYP2C9 prefers oxidative metabolism of rather small and lipophilic compounds with acidic groups having pKa values ranging from 3.8-8.1, being negatively charged at physiological pH. Typical drug classes that represent substrates of CYP2C9 include non-steroidal anti-inflammatory drugs (NSAIDs), oral sulfonylurea hypoglycemics, oral anticoagulants, diuretics, uricosurics, angiotensin II blockers, antiarrhythmics and several others, as shown in Table 2. Figure 4 additionally shows the site of oxidation of typical substrates of CYP2C9, illustrated by arrows. The drugs (S)-flurbiprofen (4'-hydroxylation)[40], (S)-warfarin (7-hydroxylation)[40], tolbutamide (methylhydroxylation), phenytoin (4'-hydroxylation)[41], losartan (oxidation)[42] and diclofenac (4'-hydroxylation)[40] are probably the best known and most characteristic substrates of CYP2C9 and are therefore often used as probe substrates for the measurement of enzyme activities.[4,6]

Class	Antiinflam-matories	Oral hypoglycemics	Oral anticoagulants	Diuretics and uricosurics	Angiotensin II blockers
Substrates	Flurbiprofen Diclofenac Naproxen Piroxicam Suprofen Ibuprofen Mefenamic acid Celecoxib	Tolbutamide Glyburide Glipizide Glimepiride	(S)-Warfarin (S)-Acenocoumarol (Phenprocoumon)**	Torsemide Ticrynafen** Sulfinpyrazone sulfide	Losartan Irbesartan Candesartan
Class (con't.)	**Antiasthmatics**	**Anticonvulsants**	**Anticancer agents**	**Endogenous compounds**	**Miscellaneous**
Substrates (con't.)	Zafirlukast (Zileuton)	Phenytoin (Phenobarbital) (Trimethadione)	Cyclophosphamide (Tamoxifen)	Arachidonic acid 5-Hydro-xytryptamine Linoleic acid	Mestranol Fluvastatin Δ9-Tetrahydro-cannabinol (Benzopyrene) (Pyrene) (Fluoxetine) (Sildenafil) (Rosiglitazone)

*This listing is not intended to be exhaustive.

**Parentheses indicates that (where known) other P450s or metabolic pathways play a major role in clearance.

Table 2. Common substrates for CYP2C9 by therapeutic class. Table taken from Rettie, A. E.*et al.*, 2005.[4]

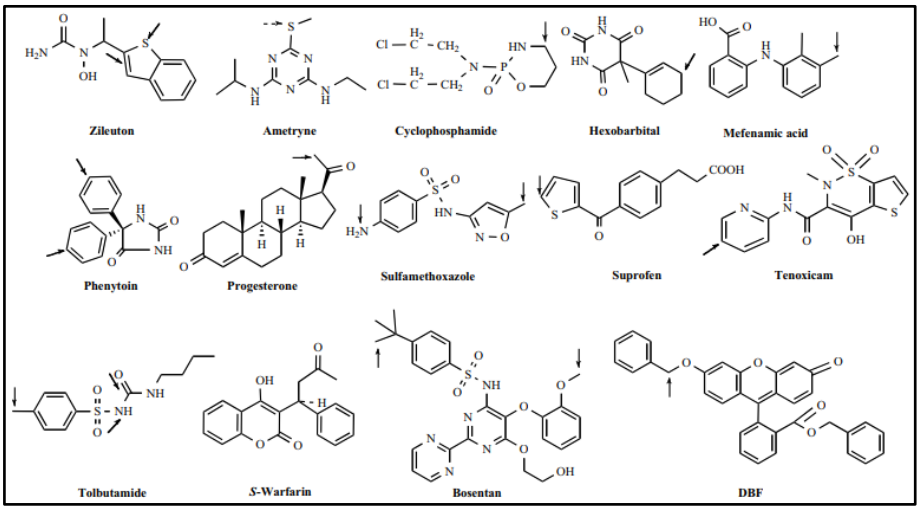


Figure 4. The site(s) of oxidation of common CYP2C9 substrates indicated by arrows. Figure taken from Zhou *et al.*, 2009.[6]

Despite the fact that the active site of CYP2C9 is already known in detail due to X-ray resolved structures, the entire 3D-structure of the enzyme including N-terminal

embedding in the membrane is still unknown. Considering that all CYP2C9 substrates have a certain lipophilicity in common, the question arises how these substrate molecules can enter the active site of the enzyme. It is possible that a special access channel exists in CYP2C9, which opens directly into the membrane of the endoplasmic reticulum and through which substrate molecules pass directly into the binding site. After biotransformation, the resulting hydrophilic metabolites are likely to be released directly into the hydrophilic environment of the cytosol via a special egress channel. Figure 5 illustrates the presumed principle behind the entry of substrates and the exit of metabolites from the enzyme.[43]

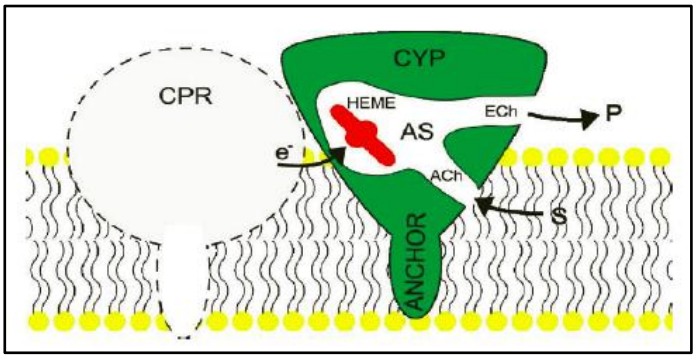


Figure 5. Hydrophobic substrates (S) enter the active site (AS) via an access channel (ACh) and the resulting products of biotransformation (P) leave the enzyme via an egress channel (ECh) to the cytosol. The oxidation reaction enabled by the heme (HEME) requires electrons provided by the cytochrome P450 reductase (CPR). Figure taken from Berka, K. *et al*., 2011.[43]

2.3.3 **Inhibitors of CYP450 2C9**

As explained in detail in section 2.2.4, there are different modes of enzyme inhibition: Reversible (competitive, non-competitive, uncompetitive, mixed-type), irreversible (mechanism-based) and quasi-irreversible inhibition. Inhibition of CYP450 isoenzymes does not necessarily affect only one specific enzyme; much more frequently, several different CYP isoenzymes are inhibited simultaneously. As with other CYP450 isoenzymes, reversible enzyme inhibition is the most common type of CYP2C9 inhibition and there are a lot of drugs to be mentioned. Since the actual assignment to the possible forms of reversible enzyme inhibition (competitive, non-competitive, uncompetitive, mixed forms) is difficult and in many cases not yet known, no further

differentiation is made between the different forms when enumerating the different drugs now.

The antiarrhythmic drug amiodarone is a weak reversible inhibitor of CYP2C9, whereas its metabolite desethylamiodarone is much more active.[44] The angiotensin-II receptor antagonists losartan > irbesartan > valsartan > candesartan inhibit the enzyme in different intensities.[45] Other enzyme inhibiting drug classes include the anti-HIV agent delavirdine and the uricosuric agent benzbromarone, which is considered one of the strongest inhibitors of CYP2C9.[46,47] Furthermore, fibrates such as gemfibrozil[48], NSAIDs such as phenylbutazone[49] or oral hypoglycemics such as glyburide inhibit CYP2C9, with glyburide being a simultaneous substrate of this enzyme.[50] Proton-pump inhibitors (PPIs) represent frequently used and therefore an important drug class, whose members inhibit CYP2C9 with different potencies (pantoprazole > omeprazole > lansoprazole > rabeprazole).[51] Particularly strong reversible inhibition of CYP2C9 is caused by the imidazole antifungal drugs fluconazole and clotrimazole.[52] The statin fluvastatin, which is also a substrate of CYP2C9, is a strong inhibitor of the enzyme, whereas other statins, such as pravastatin, lovastatin acid and simvastatin acid are only moderate to weak inhibitors of CYP2C9.[53] While the antibacterial drug sulfaphenazole is a strong reversible inhibitor of CYP2C9, its structural relatives sulfadiazine, sulfisoxazole, sulfabenzamide and sulfamethoxazole show no inhibitory activity on CYP2C9.[54,55] The antihistaminic drug cimetidine is also a very weak inhibitor of CYP2C9.[49] Other drugs with high inhibitory activity are nicardipine, nifedipine, dapsone, fluvoxamine, tamoxifen andtolbutamide.[56] Of course, not only approved drugs can inhibit CYP2C9. Plant ingredients also have the potential to do this: Curcumin and two analogues, demethoxycurcumin and bisdemethoxycurcumin show considerable inhibition of CYP2C9.[57] Flavonoids such as luteolin, baicalein, quercetin or galangin show high activity with respect to their enzyme inhibition[58–60] and also ingredients from garlic, gingko and ginseng can cause interactions due to CYP2C9 inhibition.[61,62]

Regarding irreversible inhibition (mechanism-based) there are a few example including drugs like silybin (CYP2C9, CYP3A4), (+/-)-suprofen (CYP2C9), tienilic acid (CYP2C9),

parathion (CYP2C9, CYP3A4, CYP1A2), cimetidine (CYP2C9, CYP2C19, CYP2D6), methimazole (CYP2C9, CYP2C19, CYP3A4, CYP2B6), hydrastine (CYP2C9, CYP3A4, CYP2D6) and also investigational compounds such as methylenedioxyphenyl benzothiazines (CYP2C9, CYP3A4, CYP1A1, CYP2D6), 2-phenyl-2-(1-piperidinyl)propane (CYP2C9, CYP2B1, CYP2B6) or3,5-Diethoxycarbonyl-1,4-dihydro-2,6-dimethyl-4-ethylpyridine (CYP2C9, CYP1A1, CYP1A2, CYP3A4).[6,63]

2.3.4 **Inducers of CYP450 2C9**

As in most cases for the induction of CYP450 isoenzymes, the induction of CYP2C9 also focuses on the activation of the pregnane X receptor (PXR). Typical inducers are the drugs rifampicin, phenobarbital and the plant substance hyperforin.[64] The drug bosentan, which is an endothelin receptor blocker and therefore used in the treatment of pulmonary hypertension, leads to the induction of the enzymes CYP2C9 and CYP3A4, which clinically leads to lower plasma levels of simultaneously applied simvastatin because of increased metabolism.[65] Avasimibe is a substance which reduces the deposition of cholesterol on arterial walls and thus lowers the risk of plaque formation and arteriosclerosis by inhibiting acyl-CoA:cholesterol acyltransferase (ACAT). Avasimibe leads to the induction of mRNA and protein levels of CYP2C9, CYP2C8 and CYP2B6, but is also a substrate of CYP2C9. In clinical studies, the longer-term administration of this compound resulted in a reduced AUC (area under the curve) of avasimibe due to the auto-induction of its own metabolism. In addition, with simultaneous administration of (S)-warfarin, avasimibe strongly reduced plasma levels of the anticoagulant agent and therefore reduced prothrombin times were observed.[66,67] The two anti-HIV drugs ritonavir and nelfinavir also represent potent inducers of several CYP450 isoenzymes, namely CYP2C9, CYP2C8, CYP2B6 and CYP3A4/3A5.[68] The cytostatic drug cyclophosphamide leads to an increased expression of CYP2C9 and CYP3A4 and thus also belongs to the group of clinically relevant CYP inducers.[69]

2.4 3D-pharmacophore modeling

Although Paul Ehrlich often appears in literature as the origin of pharmacophores, it was the American chemist Lemont B. Kier who actually introduced the concept and term of pharmacophores in a series of papers between 1967-1971.[70,71] Based on this preliminary work, the official definition of the term pharmacophore by IUPAC from 1998, which can be attributed to Camille G. Wermuth and Peter Günd, 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*."[72]

At this point it is important to underscore that a pharmacophore is by no means a simplified representation of certain functional groups of a molecule. Rather, it is an abstract concept that compresses bioactive substances to their available interaction characteristics (interaction features), which are responsible for ligand-target interactions. The observed ligand-target interactions are categorized into principal types of chemical features, such as hydrogen-bond donors (HBD), hydrogen-bond acceptors (HBA), positive (PI) and negative ionizable features (NI), aromatic ring interactions (AR), hydrophobic interactions (H) and different metal binding locations, the latter only available in LigandScout. Taking into account the exact chemical feature types and their localization in three-dimensional space, 3D-chemical feature-based pharmacophore models can be derived. For more than a decade 3D-pharmacophores have represented efficient in-silico tools in drug discovery for deciphering key interactions between proteins and ligands[73] i.e., the chemical features associated with

target interactions, and also for mining databases to identify bioactive molecules among millions of compounds.[74] In addition, they are widely used by chemists in drug discovery[75] and due to their abstract nature, 3D-pharmacophore models have been reported to be ideal filters for rapid and accurate screening of large virtual compound databases giving high enrichment factors and ability to scaffold hop.[76] Furthermore 3D-pharmacophores are useful also for identifying biological targets for ligands (target profiling)[77], repurposing existing drugs[78], exploring protein-protein interfaces[79], and profiling drug targets for side-effects.[80] In terms of 3D-pharmacophore model generation two approaches termed structure-based modeling and ligand-based modeling are widely used.[81,82]

2.4.1 **Structure-based pharmacophore modeling**

Due to the enormous increase in available 3D-structures of macromolecules (especially proteins) in recent years, the structure-based approach is becoming more and more important in the field of pharmacophore modeling. The Protein Data Bank (PDB), as the largest freely accessible database of macromolecular 3D-structures (often with co-crystallized ligands), is the starting point for structure-based modeling in most cases.[83] However, in industry, X-ray crystallography is used routinely to assess macromolecular-small molecule interactions and many compound analogues will be assessed with the protein target. In order to use structure-based modeling, such ligand-protein complexes must be available, preferably resolved in high resolution and/or derived from high quality structural modeling. However, when experimental data is not available computer-aided design approaches such as homology modeling, cavity prediction, apo-site pharmacophore modeling and molecular docking enables structure-based pharmacophores to be generated using information from the macromolecular structure. This structure-based approach is based on the identification of the presumably relevant interaction sites, which can be tracked down using either energy-based or geometry-based methods. Subsequently, the obtained information is put into the creation of a structure-based pharmacophore model.[84]

Even if this approach, of course depending on the quality of the underlying 3D structure of the ligand-target complex, comes closer to reality than ligand-based

pharmacophore modeling, there could be, in theory, limitations. On the one hand, if all possible binding site interactions are considered without the ligand and if all the chemical features of the ligand are considered without the binding pocket then there could be a problem that the resulting models would have a multitude of chemical features to be considered and the resulting pharmacophore models might not be suitable for hit finding via database screening. However, this is dealt with because the software algorithm analyzes all the possible chemical features of the ligand and similarly analyzes the binding cavity for complimentary features that must be in the correct geometries (distances and angles). A feature is only added to the pharmacophore model if there is a complimentary binding partner from the macromolecule. The fundamental concept of structure-based modeling is that the resulting chemical features in the model are based on information from the macromolecular structure. One limitation of structure-based modeling is that the resulting pharmacophore model is based on a single macromolecular-ligand complex, which represents one binding mode. Therefore, analysis of several structure-based models with interaction features derived from different ligands are needed to give a better picture of the structure-activity relationships (SAR) and X-ray derived data may not be available for all molecules that have been assessed at the macromolecular target. Apart from this limitation, structure-based 3D-pharmacophore modeling represents a promising and interesting method for directly deriving interactions between a compound and a macromolecule and datamining via pharmacophore-based virtual screening.[85]

2.4.2 **Ligand-based pharmacophore modeling**

Ligand-based pharmacophore modeling, in contrast to structure-based approaches, does not require the 3D-structure of the macromolecular target. It involves a set of selected ligands which have been proven to show activity at the desired target and are believed to interact with this target by a common mechanism of action. The choice of ligands is at the discretion of the modeler and depends on the outcome that is to be predicted. The aim of this approach is to identify the common chemical features shared by the set of ligands in 3D-space and represent them in a generated 3D-pharmacophore model. The guiding principle is that molecules with activity at the

same target must have some common chemical features in 3D-space. The goal of ligand-based modeling is thus to create models based on ligand information that predict biological outcomes that they have in common and simultaneously selectively do not fit the attributes that need to be avoid of inactive substances or those with undesired outcomes.[86,87]

Contrary to the structure-based approach, which adds pharmacophore features based on information from the macromolecular structure, ligand-based pharmacophore approaches identify common features of the ligands in 3D-space, without information from the macromolecular structure. However, there may be uncertainty about which features are responsible for the biological effect if the bioactive conformations of the ligands are not known. The procedure for ligand-based pharmacophore modeling roughly consists of three steps. First, a thorough and accurate sampling of the conformational space of each molecule must be conducted in order to ensure optimal alignment for the purpose of generating a high-quality model. An algorithm to do this is needed and there are many available. In the case of LigandScout, iCon is used to sample conformational space, generate 3D-conformations, prioritizes and stores them in a file.[88]

This step is followed by annotation of each conformation for each molecule its 3D-chemical features and the third step involves multiple alignment experiments of the 3D-chemical features of each conformation of each ligand and then the ligand-based 3D-pharmacophore model is constructed. Software programs for generating ligand-based pharmacophore models are LigandScout (available from Inte:Ligand GmbH; http://www.inteligand.com/), Catalyst (available from Dessault Systèmes), Phase (available from Schrodinger), and MOE (available from CCG). There are several notable differenc~~es~~ between these programs. LigandScout is unique in that it has improved and different algorithms for the pharmacophore feature definitions, generation of ligand conformations and alignment of molecules based on pattern matching and no limitation on the number of features that can be added to a model.[82,85]

2.4.3 **3D-Pharmacophore-based virtual screening**

Today, classical preclinical drug development research is divided into several steps, which are as follows: Target selection, hit identification, hit expansion, hit-to-lead transition, lead optimization and preclinical profiling.[89] Hit identification is crucial for the success or failure of the whole project.

To generate such hit molecules, in current drug discovery, usually large chemical libraries are screened against the desired intracellular or extracellular target. In primary assays, high-throughput screening (HTS) can be used to determine the activity of numerous substances on a given target in order to obtain hits for further steps in drug discovery. With the help of various technologies and advances in assay miniaturization, lab automation and the use of robots, it is possible to test daily 10 000-100 000 compounds for their activity on a specific biological target. Since the beginnings of HTS in the early to mid-1990s, substance libraries have grown steadily, which means that time is a big factor in the performance of high-throughput screening. In addition, the cost of purchasing the many compounds plays a major role and a limiting factor, which is why many academic labs, unlike industry, may not be able to afford HTS robotics. Virtual screening of 3D-pharmacophore models against virtual compound libraries thus makes it possible to filter out substances at an early stage and thus reduces the number of compounds that are subsequently used for HTS.[90]

While HTS experimentally measures the activity of molecules in a primary assay, virtual screening involves the use of a model, in this case a 3D-pharmacophore, to query a database of compounds. Virtual compounds in the database must match the defined features of the pharmacophore model as well as their 3D-geometries in order to be retrieved as a virtual hit. The hits retrieved are ranked with a geometric pharmacophore fit score. In order for a database to be queried with a model it must be prepared in advance. In the case of LigandScout an algorithm called idbgen is used to process compound libraries, e.g., from commercial providers, from public sources like ChEMBL or a library idea for new compounds to be prioritized for synthesis. idbgen computes conformations for each compound in the library and annotates each conformation with pharmacophore features so that a pharmacophore model can be

used to query the database. To conduct the virtual screening experiment another algorithm is needed. In the case of LigandScout, iScreen is a pattern matching algorithm that aligns the pharmacophore features of the compounds in the library to the features of the pharmacophore model and gives it a pharmacophore fit score. Compounds with features that do not match the model are not scored and are not added to the virtual hit list.

Libraries of compounds with known biological activities at the target of interest can be computed and used to determine how well a model can distinguish true active compounds from inactive compounds at a target of interest. If the model is highly predictive, it means that it identifies active and inactive molecules correctly. If the model performs poorly the features must be modified until it is predictive. A model that has been refined and tested is referred to as a validated model. After the model has been refined to a satisfactory level it can be used to query databases in a hit finding experiment. Further validation of the model involves the biological assessment of the virtual hits retrieved by the model (Figure 6).[80] Thus, both validated ligand-based and structure-based pharmacophore models can be used for hit identification as well as lead optimization in the process of early drug discovery.

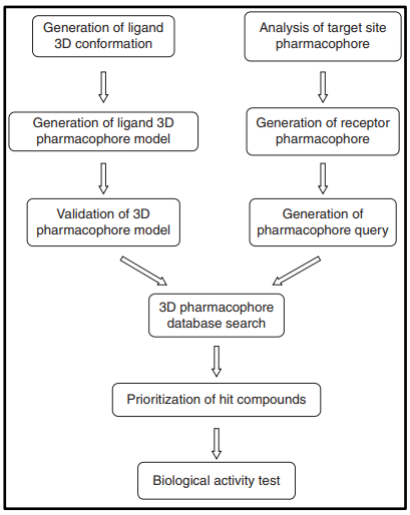


Figure 6. Basic procedure of structure-based and ligand-based pharmacophore virtual screening. Figure taken from Kim *et al.*, 2010.[91]

In order to validate the performance of a pharmacophore model, there are several parameters where sensitivity and specificity play an important role. Sensitivity indicates the true positive rate, i.e. how many of the true active compounds are retrieved by a model. This value is calculated by dividing the number of true positives by the sum of true positives and false negatives. Thus, a sensitivity of 1 means that all active compounds are recognized as active. A sensitivity of 0, on the other hand, indicates poor performance, since none of the active compounds were recognized as active. In contrast, specificity indicates the true negative rate. It indicates how many of the actually inactive compounds are classified as inactive, i.e. true negative. To do this, the number of true negatives is divided by the sum of true negatives and false positives, which in turn yields a value between 0 and 1. A specificity of 1 stands for ideal performance, since all inactive molecules are recognized as inactive, whereas a value of 0 represents the worst case.[92]

The enrichment factor (EF) is another parameter to assess the predictivity of a model. It indicates how many times more actives are detected by a pharmacophore model than if a random assignment would have taken place. Thus, with the help of the EF, it can be very well illustrated how much better a model performs compared to a random assignment. However, there are two limitations: First, the amount of this value depends very much on the ratio between active and inactive compounds in the database, which allows a comparison between the performance of models only if they have the same active/inactive ratio. Secondly, the EF does not provide any information about the ranking of the alignment scores in the hit list. In order to counteract this second problem, the program LigandScout provides the corresponding EFs at 1, 5 and 10 percent in addition to the overall enrichment factor.[93]

Probably the most illustrative way to present screening results is the receiver operating characteristic (ROC) curve analysis. After designating a library of active compounds and library of inactive compounds and screening both with a pharmacophore model, and assessing the hit list molecules, the sensitivity and specificity values are determined and then plotted as a function of sensitivity and 1-specificity. Good models lead to a large area under the curve, whereby it is important

that the curve rises very early. This means that true positives are retrieved by the model. Poor models, on the other hand, show a much less steep rise of the curve and/or the area under the curve is significantly lower. The dotted line corresponds to a random allocation of the compounds, which indicates a very poor performance of the model. If the AUC is even lower, it can be said that a pharmacophore model performs worse than random assignment because it retrieves inactive compounds which is undesireable.[94] All these different validation parameters are important not only to compare the performance of pharmacophore models, but also to assess the impact of refinement of the models during the training/refinement phase.

3 Materials

3.1 ChEMBL

The ChEMBL database (https://www.ebi.ac.uk/chembl/), also known as ChEMBLdb, is a freely accessible database containing compounds annotated with biological assay data. It is operated by the European Bioinformatics Institute (EBI), which is part of the European Molecular Biology Laboratory (EMBL), and it is located at the Wellcome Trust Genome Campus in Hinxton, UK. With well over 1.9 million compounds and 1.1 million assays, the ChEMBL database offers the most comprehensive range of freely available molecular data resources worldwide.[95,96]

With regard to other scientific research areas such as sequence analysis, structure elucidation of proteins or gene expression data, the journals in which the authors publish oblige them to submit their results to database. However, this is usually not the case in the area of bioactivity data of molecules, which forces one to manually extract data from various publications. Unfortunately, the contents of published articles from diverse journals are usually only available in very unstructured and non-transparent formats, making it difficult to find and extract this sort of experimental data. The ChEMBL database has therefore set itself the goal of solving this problem in addition to other databases such as PubChem (https://pubchem.ncbi.nlm.nih.gov) or DrugBank (https://www.drugbank.ca), with each of the various databases focusing on different areas and contents.[97]

For the ChEMBL, the topics of interest are experimental data from assays dealing with binding, function, absorption, distribution, metabolism and excretion of drug-like molecules. The corresponding targets of these data range from single proteins and protein complexes to measurements on tissues or whole organisms. The ChEMBL offers free access to a wide range of high quality data and thus gives everyone the opportunity to perform data-driven analyses of polypharmacology, bioisosteric replacements, chemogenomics, drug repurposing and predictive modelling.[98] The contents of the ChEMBL arise from various scientific publications in the field of medical

chemistry (e.g., Journal of Medicinal Chemistry, Bioorganic Medicinal Chemistry Letters, Journal of Natural Products and others) and are manually extracted from the corresponding original publications, precisely examined by experts for their correctness and standardized to common units where possible. For example, IC50 (half maximal inhibitory concentration) values are given in nM instead of µM/mM/M and half-lives are given in hours rather than in minutes/days/weeks. In this way, the data of different assays can be better compared by the user and it is easier to handle it. Besides information manually extracted from publications, the ChEMBL also contains complete screening results of other public databases (e.g., PubChem Bioassay[99]).[97,100]

3.2 Protein Data Bank (PDB)

The Protein Data Bank (PDB; http://www.rcsb.org/pdb/) was founded in 1971 at Brookhaven National Laboratories in New York as an archive for biological macromolecular crystal structures. Since 1998, the company has been managed by the Research Collaboratory for Structural Bioinformatics (RCSB). The PDB represents a large freely accessible collection of 3D structural data and related information of macromolecules and thus provides a pool of information from which researchers in the fields of biology, chemistry, computational science and students from various scientific fields benefit since years. Although the PDB archive initially contained only seven macromolecular structures, today users can access thousands of different structural biological data (152500 biological macromolecular structures).[101]

This immense structure collection contains proteins, nucleic acids and large macromolecular complexes of various organisms such as Homo sapiens, Escherichia coli, Mus musculus or Saccharomyces cerevisiae, whereby macromolecules of Homo sapiens make up the largest part of the library. In addition to the use of X-ray crystallography, NMR (nuclear magnetic resonance) and electron microscopy techniques are becoming increasingly important methods for generating the structural data provided by the PDB. Since the PDB is used all over the world today and the generated structural data originates not only from the United States but also from various countries in Europe and Asia, the worldwide PDB (wwPDB;

http://www.wwpdb.org/) was established in 2003 to ensure that all data files are uniform in content and format.[102]

Each entry in the PDB receives a so-called PDB ID, which represents a characteristic four-letter code with the general structure *nxyz*. Here *n* corresponds to an integer and *x, y* and *z* represent alphanumeric characters. The four-letter code is arbitrarily assigned to each entry and represents the unique and unchangeable identity of it. It is the definite reference between a structure from the PDB and the corresponding literature, which characterizes it. An essential part of the PDB is of course the efficient collection of high-quality data, which can be summarized as data processing. The steps of data processing are illustrated in Figure 7 and look as follows: After a structure has been deposited by the author, a corresponding PDB ID is immediately provided (step 1). Now the primary data are examined in the course of validation processes for errors and inhomogeneities. The subject of these validation processes is the assessment of the quality of the existing atomic model and the agreement between that model and the underlying experimental data. Then, the annotated entry and a corresponding comment including validation information is sent back to the author (step 2). After all points of criticism have been corrected, the author sends the data back again (step 3), whereby steps 2 and 3 can be repeated, depending on how satisfactory the corresponding corrections are. After final release by the author (step 4), the corrected data is available to users in the PDB archive.[103]

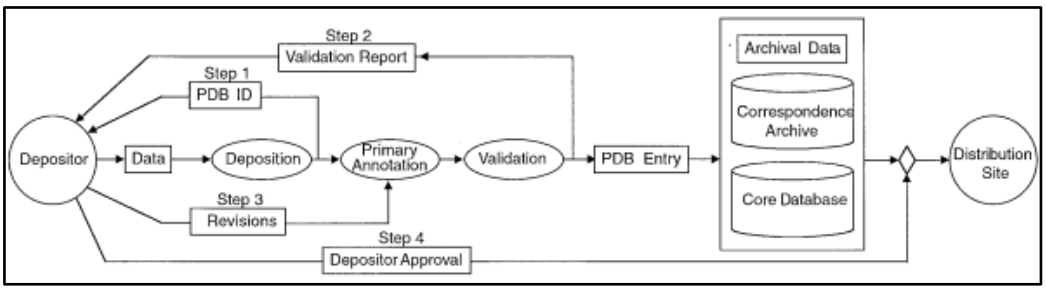


Figure 7. Steps of PDB data processing. Figure taken from Berman *et al.*, 2000.[101]

From the RCSB PDB website, the user has various options for searching, visualizing and analyzing structural data. A toolbar at the top of the website allows to search for structures using the PDB ID, names, sequences or ligand SMILES codes (Simplified Molecular Input Line Entry Specification). Furthermore, an extended search function on the left side of the website offers the possibility to refine the search (e.g., organism, UniProt molecule name, taxonomy, experimental method, X-ray resolution, release date, polymer type, enzyme classification). For each entry labelled with a specific PDB ID, all structural information about the macromolecule and any complexed small molecules is provided. Furthermore, information about the method of data generation, experimental data and validation details is provided. In addition to various options for visualizing the three-dimensional structures, the RCSB PDB naturally offers the free download of the various structural data, so that users can also continue working with these outside the RCSB PDB website.[104]

3.3 KNIME

With the rapidly increasing number of available compounds, bioassay records and 3D structures of macromolecules in databases such as the ChEMBL, PubChem or the PDB, the need for software to analyze and process those public data is clearly increasing. For scientists in the fields of bioinformatics, metabolomics, pharmacogenetics and cheminformatics, the program KNIME (Konstanz Information Miner) as a freely available analytics platform is an ideal tool for this purpose. KNIME is a Java-based modular data analytics and pipelining platform that was developed in 2004 at the University of Konstanz, Germany. The graphical user interface of KNIME allows data streams to be routed from a source input node to other nodes along the pipeline and in this way to be processed and analyzed.[105]

The processing of various data takes place within so-called nodes, which are the most general data processing units, and information can be transported between those nodes with the help of connections. KNIME provides a variety of different nodes that can be combined by simple drag & drop between input and export ports and therefore the program allows users to create and customize individual workflows to process their data. The nodes used can have different statuses: "Not executable" is indicated

by a red light and means that the node is not connected to its predecessor or that it has been configured incorrectly. If a yellow light illuminates on the node, this means "ready for execution" and nothing stands in the way of executing the node-associated function. If the data has been correctly processed in a node, the state "executed" is indicated by a green light.[106]

In principle, the nodes available in KNIME can be divided into four categories: First, there are nodes that are used to import files or databases and to export data. Then there are numerous nodes to manipulate imported data, which includes e.g., filtering, grouping, binning or sampling. Of course, KNIME also provides numerous nodes for visualizing imported contents, allowing interactive exploration of the data. Finally, data mining is also an essential aspect of KNIME: Various data mining algorithms like clustering, decision trees, neural networks and support vector machines are only some of the techniques that enable tasks like pattern recognition, dimensionality reduction and feature reconstruction. In addition to the proprietary nodes, it is also possible to implement self-programmed nodes and KNIME extensions, which can then be easily connected to proprietary ones. This allows KNIME to be ideally adapted to the needs of the user and provides a high degree of flexibility.[107]

A typical KNIME workflow starts with a node that reads in the data from some source, where the corresponding data are mostly text files. Of course, it is also possible to query databases, whereby there are separate nodes for this application. After data has been imported into a workflow, the information is stored in an internal table-based format. There are columns, which contain the data properties (e.g., integer, string, image, molecule) and an arbitrary number of rows corresponding to the column specifications. Data is then easily sent from node to node via manually set connections, so that the contents can be visualized, modified, transformed or modelled depending on the type of nodes used. The unique thing about this pipelining tool is that the nodes in KNIME keep the data and results permanently and therefore the user has the possibility to interrupt processes at any nodes, re-set connections and change workflows without losing previous results.[108] The following graphic (Figure 8) shows a typical workflow in KNIME using some of the Inte:Ligand Expert KNIME Extensions. For

the present scientific work, the version KNIME 3.6.1 was used, which was the latest edition at that time. In addition to some proprietary nodes, the various nodes of the Inte:Ligand/LigandScout Expert KNIME Extension were used to process the data.

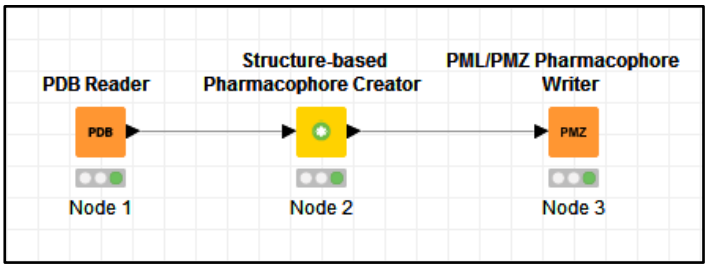


Figure 8. Inte:Ligand Expert KNIME Extensions for processing PDB file(s), automatically creating SB-Pharmacophore model(s) and writing the resulting model(s) to a file.

3.4 LigandScout

LigandScout is a commercially available molecular design software developed by Inte:Ligand GmbH (http://www.inteligand.com/), located in Vienna/Austria, which includes a user-friendly graphical user interface and command line tools for most of its algorithms. The program allows the generation of three-dimensional pharmacophore models, both structure-based and ligand-based approaches, and the execution of 3D virtual screening jobs. Since the original release of LigandScout in 2005, when it was the first tool developed to generate interactions from a protein ligand complex in a click of a button, it has evolved with multiple versions into an elaborate molecular design program with capabilities to draw, design and analyze molecules as well as predicting cavities on macromolecules. The latest version, LigandScout 4.4, has been used for the present work.

LigandScout is one of the most important programs for 3D-pharmacophore modeling and virtual screening, along with other frequently used software applications such as Catalyst (from Dessault Systèmes), MOE (from Chemical Computing Group) or Phase (from Schrödinger). In addition to its very clear and user-friendly interface, LigandScout offers two further advantages compared to other commercial programs. On the one hand, the software allows generation pharmacophore models without limiting the number of features, which guarantees high selectivity and quality of the

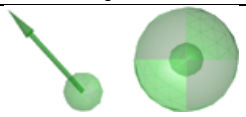
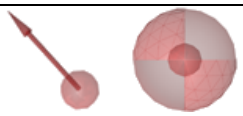
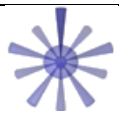
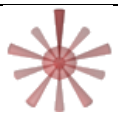
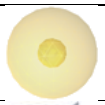
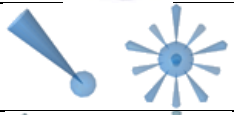
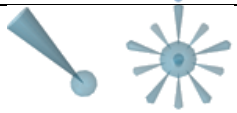
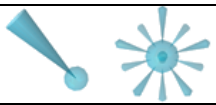
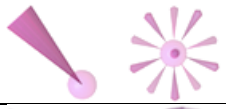
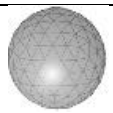
models. On the other hand, alignment by pattern recognition rather than graph matching algorithms is advantageous because it not lead to exponential increase in screening time with an increasing number of features.[73,109]

With regard to the user interface, LigandScout offers four different perspectives between which models and ligands can be easily copied. In the "Structure-Based" perspective, LigandScout offers the possibility to retrieve macromolecules from the PDB by entering the corresponding PDB 4-letter codes and then generate structure-based 3D pharmacophore models by pressing a single button. Another important feature within this view is docking of ligands into various binding sites, either using AutoDock or AutoDock Vina. Meanwhile, other functions are also possible, such as, cavity prediction and opening trajectories of molecular dynamics (MD) simulations and generating MD pharmacophores.

The entire process of ligand-based pharmacophore modeling takes place in the so-called "Ligand-Based" perspective. First, numerous conformations of molecules are generated, each conformation annotated with pharmacophore features and then these are aligned, and a corresponding ligand-based model is generated. Furthermore, molecules can be clustered by pharmacophore similarity in order to ensure rational pharmacophore models with high similarity of the features of the ligands used and further evaluate potential unique modes of binding to a macromolecule. Another useful application for molecular modeling is the so-called "Alignment" perspective, which, as the name suggests, can be used to align different pharmacophores and/or ligands. Once a pharmacophore or ligand has been set as reference, any number of pharmacophores and/or ligands can be aligned by features or reference points and subsequently a shared-feature or merged-feature pharmacophore model can be generated. The fourth perspective of the program is the "Screening" perspective, which is used to calculate libraries for virtual screening, load libraries for virtual screening and perform screening using one or more 3D pharmacophore models against various virtual databases. The used databases (LDB format) can be defined also as actives or decoys. Detailed screening results are then presented in form of a hit list and the performance of the pharmacophore models can also be assessed using ROC

curves automatically including different calculated parameters (sensitivity, specificity, ROC curve analysis, EF, AUC). The current version LigandScout 4.4 offers additional pharmacophore features shown in Table 3, with some features appearing not only as a sphere, but also as vector features.

Table 3. All pharmacophore features (spheres and vectors) available in LigandScout 4.4.

Feature Representation	Feature Type
	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

In addition to the software described above, Inte:Ligand also provides the Inte:Ligand Expert KNIME Extension package, which enables the implementation of various functionalities of LigandScout within KNIME. This extension was often used during the research work and is a practical way to provide the Inte:Ligand algorithms into the KNIME analytics platform to be easily used in pipelining and customized workflow creation.

4 Results and discussion

4.1 Creation of data sets

For the whole part of processing the ChEMBL libraries related to CYP2C9 and creating the data sets for modeling, the program KNIME (version 3.6.2) and the Inte:Ligand Expert KNIME extensions were used. Table 4 gives an overview of the most frequently applied IL Expert KNIME nodes used in this work.

Table 4. Most frequently applied IL Expert KNIME nodes in this work.

IL KNIME-Nodes	Description
SDF Reader*	Reads molecules from SDF (Molecule Structure-Data Files)
SDF Writer*	Saves molecules in SDF format (Structure-Data Files)
Library Filter*	Allows filtering of molecule libraries (two outputs: filtered molecules, filtered out molecules)
Row Filter	Allows filtering of molecule libraries (one output: filtered molecules)
Row Splitter	Allows filtering of molecule libraries (two outputs: filtered molecules, filtered out molecules)
String to Number	Converts values of a certain column from string to number
Table Merger*	Merges data from two molecule tables together
Duplicate Remover*	Calculates Extended-Connectivity-Fingerprints of the molecules in a table and keeps only the first detected molecule with same fingerprint
iCon*	Conformation generation algorithm: generates all possible conformations of molecules
Chiralator*	Calculates the chirality for undefined stereo-centers of a molecule
Activity profiling*	Parallel profiling of pharmacophore models against databases (LDB format) designed to predict targets and assess toxicity profiles. Illustrates whether and how strongly a molecule fits a model with heat map and hits lists and fit scores output.

***from LigandScout Expert KNIME Extensions**

The IL Expert KNIME ChEMBL extractor node was used to extract a library of compounds with target ID ChEMBL3397 representing all molecules in the database annotated with biological activities derived from assays involving organism Homo sapiens and CYP 2C9. In the case of competitive CYP2C9 inhibitors, the underlying data was extracted from ChEMBL and consisted of a total of 22218 unique compounds. The following properties from ChEMBL were extracted: Name, SMILES code, target

confidence, compound ID, reference, assay ID, compound synonyms, activity operator, assay type, assay description, known drug, activity value, activity units, name in reference, bioactivity type, target name, molecular formula, activity comment and target ID. Naturally, using information from different assays and sources leads to a certain inhomogeneity of data, since the respective experiments occurred at different times, locations and conditions. However, since pharmacophore models do not require input of an activity value as a descriptor, decision about the data for training sets were made by careful inspection and investigation of what is known drugs that strongly inhibit CYP2C9. The aim of this part was to create different data sets that would be suitable for model creation, training and testing. In order to be able to test the pharmacophore models for their applicability domain later on, different activity classes had to be defined. The original data coming from ChEMBL3397 was divided into very strong inhibitors, strong inhibitors, moderate inhibitors, weak inhibitors and inactive compounds. In this case, "inactive" means that, according to the underlying biological assay data, a corresponding compound has no inhibitory activity against the target CYP 2C9. Therefore, activity thresholds according to the compound´s Ki (inhibitory constant), IC50 (half maximal inhibitory concentration), potency and inhibition were defined, which is shown in Table 5-8.

Table 5. Activity threshold definitions of compounds evaluated in CYP 2C9 inhibition assays (Ki).

Activity Class at CYP 2C9	**Activity Threshold- Ki (nM)**
Very strong inhibitors	0.0-999.9 nM
Strong inhibitors	1000.0-4999.9 nM
Moderate inhibitors	5000.0-9999.9 nM
Weak inhibitors	10000.0-49999.9 nM
Inactives	>50000.0 nM

Table 6. Activity threshold definitions of compounds evaluated in CYP 2C9 inhibition assays (IC50).

Activity Class at CYP 2C9	**Activity Threshold- IC50 (nM)**
Very strong inhibitors	0.0-999.9 nM
Strong inhibitors	1000.0-4999.9 nM
Moderate inhibitors	5000.0-9999.9 nM
Weak inhibitors	10000.0-49999.9 nM
Inactives	>50000.0 nM

Table 7. Activity threshold definitions of compounds evaluated in CYP 2C9 inhibition assays (Potency).

Activity Class at CYP 2C9	Activity Threshold- Potency (nM)
Very strong inhibitors	0.0-999.9 nM
Strong inhibitors	1000.0-4999.9 nM
Moderate inhibitors	5000.0-12999.9 nM
Weak inhibitors	13000.0-29999.9 nM
Inactives	>30000.0 nM

Table 8. Activity threshold definitions of compounds evaluated in CYP 2C9 inhibition assays (Inhibition).

Activity Class at CYP 2C9	Activity Threshold- Inhibition (%)
Very strong inhibitors	90.0-100.0%
Strong inhibitors	75.0-89.9%
Moderate inhibitors	60.0-74.9%
Weak inhibitors	50.0-59.9%
Inactives	0.0-49.9%

The very first step of processing the data was to remove molecules with inconsistent information or doubtful origin. Furthermore, duplicates, that exist due to molecules being measured in multiple assays, had to be eliminated in order to avoid generation of artifacts during the evaluation the results. This primary reduction of compounds was of major importance to have certainty of reliable data with the highest possible quality. Afterwards, the original table of molecules was separated based on if the compounds were either known drugs or investigational compounds. The next step of processing the library of compounds was filtered by their bioactivity type, whereby Ki, IC50, potency and inhibition were the predominant types. Entries with unusual or missing bioactivity types were not considered and simply removed. In a next step those primary data sets underwent further division by the above depicted activity thresholds. This process led to five subsets (very strong inhibitors, strong inhibitors, moderate inhibitors, weak inhibitors, and inactive compounds) for each of the four bioactivity types. Finally, the subsets which belonged to different bioactivity types but fell into the same activity class, were merged again. Then the created data sets were either assembled into training or tests (see Figure 9).

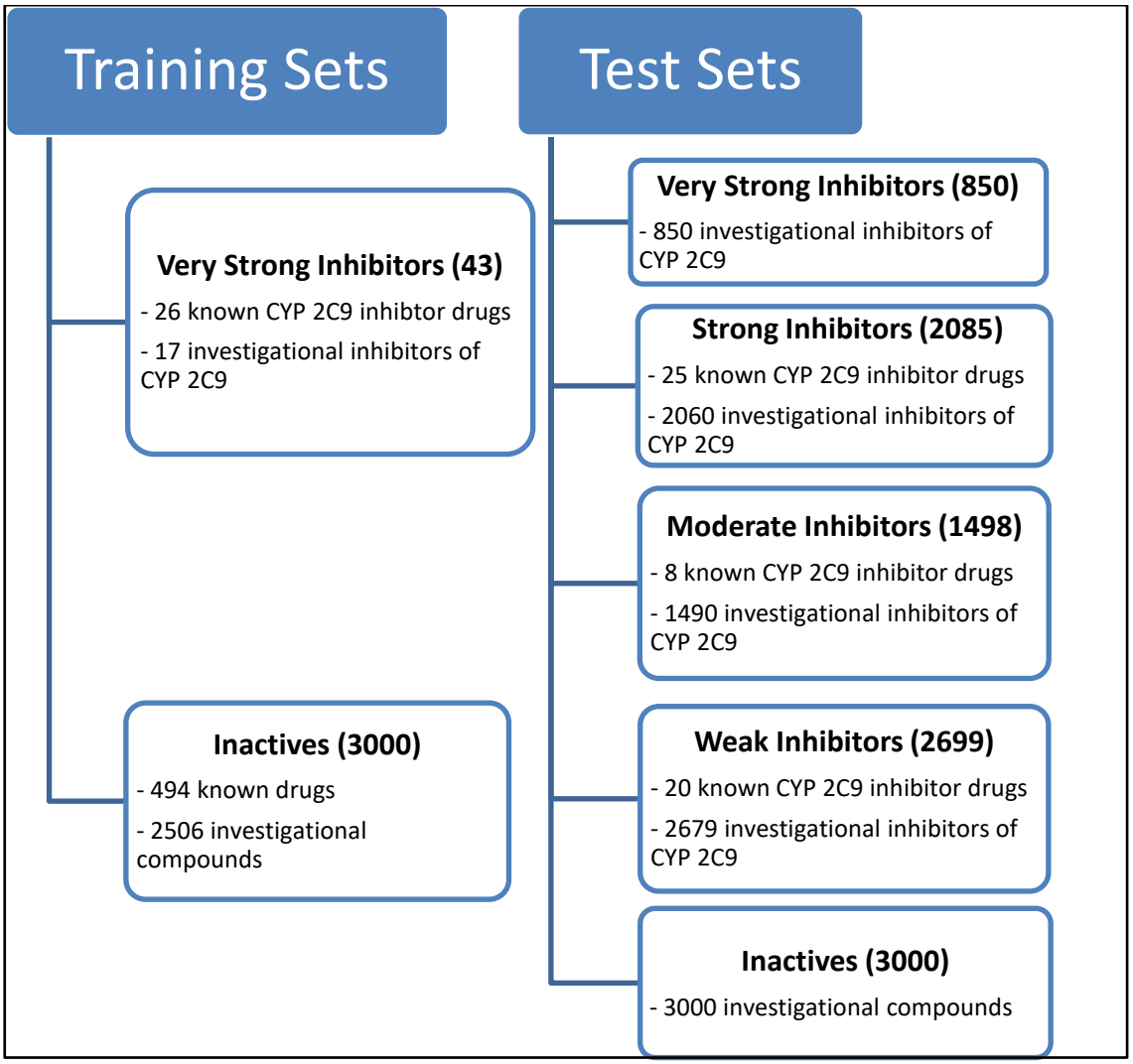


Figure 9. Assembled data sets derived from the ChEMBL database (Target ID:ChEMBL3397) for modeling and assessment of pharmacophore models designed to predict inhibitors of CYP2C9.

The creation of the very strong inhibitor training set now requires a somewhat more detailed explanation, since this set was of extraordinary importance for model creation and model testing. As the very strong inhibitor compounds of the training set would later serve for generating the actual ligand-based 3D pharmacophore models, these molecules had to be of particularly high quality and correctness. Based on these very strong competitive inhibitors of CYP 2C9, high-performance 3D-pharmacophore models were to be generated in order to retrieve very strong inhibitors and to predict potential inhibitors in the future. Accordingly, it was important that the models to be generated contained the essence of relevant features in three-dimensional space, so the very strong inhibitor training set had to be of high quality.

Some compounds in the ChEMBL3397 contained chiral centers that may not have been defined due to the fact that the enantiomers were not separated before the biological assessment. Therefore, the enantiomer responsible for the biological activity is not known. In fact, an estimated 50% of markets drugs are chiral and among those another 50% are mixtures, rather than a single enantiomer.[110] However, for 3D-pharmacophore modeling the 3D-orientation of the chemical features of the molecules is important. With regard to the very strong inhibitor training set compounds, all possible stereoisomers of the molecules had to be generated in a first step in order to be sure, that the active stereoisomer was present. The Chiralator node of the LigandScout Expert KNIME Extensions was used to generate all possible stereoisomers of the very strong inhibitor training set compounds.(see Figure 10)

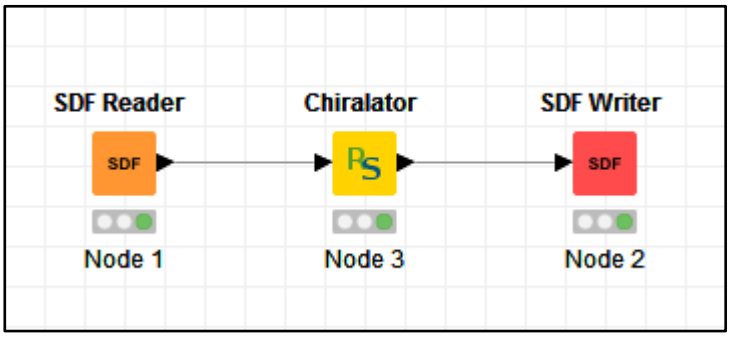


Figure 10. Creation of all stereoisomers of the very strong inhibitor training set compounds using nodes of the LigandScout Expert KNIME Extensions: Import of the very strong inhibitor training set compounds (SDF format) using the SDF Reader node, generation of all possible stereoisomers using the Chiralator node and subsequent export of the stereoisomers using the SDF Writer node.

After this was done, it was important to investigate if the eutomers (biologically active enantiomer/diastereomer of a molecule) of the compounds were known. Following this, if known, the inactive stereoisomers of the training set compounds were deleted. In case of uncertainty about the actually active stereoisomer of a compound, of course, all stereoisomers were included. Thus, e.g. in the case of fluvastatin, only the stereoisomer (3R, 5S) -fluvastatin was retained, since this is the eutomer. For example, in the case of miconazole, both enantiomers had to be included because it is not clear which of the two stereoisomers is actually active.

Finally, the very strong inhibitor compounds of the training set consisted of 43 molecules (26 drug molecules, 17 investigational molecules), which are listed in Table 9. According to the activity threshold, these 26 drug molecules were the only very strong inhibitor drugs of CYP450 2C9 in the data set and included many well-known inhibitors. In order to exclude any single-clusters for the later approach of ligand-based modeling, 17 very strong investigational inhibitor compounds were added manually.

Table 9. Strong inhibitors of CYP 2C9 used in the training set (sorted by activity value).

Name/ChEMBL ID	Known Drug	Bioactivity Type	Activity Value	Assay ID
CHEMBL1223034	No	IC50	2 nM	CHEMBL1227229
Rimonabant (CHEMBL111)	Yes	IC50	5 nM	CHEMBL1227229
Benzbromarone (CHEMBL388590)	Yes	Ki	19 nM	CHEMBL916313
CHEMBL1222969	No	IC50	>20 nM	CHEMBL1227229
CHEMBL458566	No	Ki	23 nM	CHEMBL972103
CHEMBL225481	No	Ki	40 nM	CHEMBL916313
CHEMBL457087	No	Ki	44 nM	CHEMBL972103
CHEMBL1836777	No	Inhibition	92 %	CHEMBL1837191
CHEMBL1359713	No	Potency	100 nM	CHEMBL1614027
CHEMBL250251	No	Ki	101 nM	CHEMBL945056
CHEMBL1596831	No	AC50	125 nM	CHEMBL1741325
*(3S,5R)-Noscapine, (3R,5S)-Noscapine, (3R,5R)-Noscapine, (3S,5S)-Noscapine (CHEMBL364713)	Yes	Potency	125 nM	CHEMBL1614027 (mixture)
Clotrimazole (CHEMBL104)	Yes	IC50	136 nM	CHEMBL3530854
Sulfaphenazole (CHEMBL1109)	Yes	IC50	180 nM	CHEMBL3367088
(S)-Dexmedetomidine (CHEMBL778)	Yes	Potency	199 nM	CHEMBL1614027
*(S)-Sulconazole, (R)-Sulconazole (CHEMBL1221)	Yes	IC50	200 nM	CHEMBL1909135 (mixture)
*(S)-Miconazole, (R)-Miconazole (CHEMBL91)	Yes	IC50	200 nM	CHEMBL1909135 (mixture)
*(S)-Econazole, (R)-Econazole (CHEMBL808)	Yes	IC50	200 nM	CHEMBL1909135 (mixture)
(1S,2R,13R,14S,17R,18S)-Danazol (CHEMBL1479)	Yes	IC50	300 nM	CHEMBL1909135

Dicoumarol (CHEMBL1466)	Yes	IC50	300 nM	CHEMBL1909135
*(S)-Nitrendipine, (R)-Nitrendipine (CHEMBL475534)	Yes	IC50	300 nM	CHEMBL1909135 (mixture)
CHEMBL1504465	No	AC50	354 nM	CHEMBL1741325
*(R)-Nicardipine, (S)-Nicardipine (CHEMBL1484)	Yes	IC50	378 nM	CHEMBL3530854 (mixture)
Disulfiram (CHEMBL964)	Yes	Potency	398 nM	CHEMBL1614027
(3R,5S)-Fluvastatin (CHEMBL2220442)	Yes	IC50	400 nM	CHEMBL1909135
CHEMBL475639	No	IC50	450 nM	CHEMBL1034809
CHEMBL1256759	No	Potency	501 nM	CHEMBL1614027
CHEMBL1480903	No	Potency	631 nM	CHEMBL1614027
CHEMBL1596756	No	Potency	631 nM	CHEMBL1614027
CHEMBL3209598	No	AC50	794 nM	CHEMBL1741325
CHEMBL1403311	No	AC50	891 nM	CHEMBL1741325
CHEMBL596515	No	IC50	<1000 nM	CHEMBL1073249
Fluconazole (CHEMBL106)**	Yes	IC50	9210 nM	CHEMBL3825882
Gemfibrozil (CHEMBL457)**	Yes	IC50	9600 nM	CHEMBL2185185
(2R,4S,5R)-Floxuridine (CHEMBL917)**	Yes	Inhibition	Unspecified	CHEMBL2071963

*** Mixture- component responsible for activity not reported. Stereochemistry elaborated for purpose of modeling.**

****recognized as very strong inhibitors in literature**

Concerning the inactive compounds of the training set and the different test sets, the different stereoisomers were not generated. Hence, for all the other molecules the one stereoisomer of the original ChEMBL data set was used for further actions. All the processed data sets were available in SDF format, but to be able to query them with a 3D-chemical feature-based pharmacophore they had to be converted into LigandScout database format (LDB) using the algorithm idbgen (Inte:Ligand).

Basically, all the further steps of the practical part were executed in the program LigandScout (version4.4.).

4.2 Training and refinement

4.2.1 **Structure-based modeling**

As a fundament for structure-based pharmacophore modeling, contents from the RCSB PDB (Protein Data Bank) were used. There are ten resolved X-ray structures of wildtype CYP2C9 and genetic variants available on the RCSB PDB. (see Table 10)

Table 10. Available X-ray structures of CYP450 2C9 from the PDB.

PDB 4-letter code	Ligand (name/ChEMBL ID)	Resolution	Reference
1OG2	Warfarin (substrate)	2.6 Å	Williams *et al*., 2003[38]
1OG5	Warfarin (substrate)	2.55 Å	Williams *et al*., 2003[38]
5A5I	ChEMBL3818307 (inhibitor)	2 Å	Skerratt *et al*., 2016[111]
5A5J	ChEMBL3818248 (inhibitor)	2.9 Å	Skerratt *et al*., 2016[111]
4NZ2	*Unspecified* (inhibitor)	2.45 Å	Brändén *et al*., 2014[112]
5W0C	*Unspecified* (inhibitor)	2.001 Å	Liu *et al.,* 2017[113]
5XXI	Losartan (substrate)	2.3 Å	Maekawa *et al*., 2017[114]
5X23	Losartan (substrate)	2 Å	Maekawa *et al*., 2017[114]
5X24	Losartan (substrate)	2.48 Å	Maekawa *et al*., 2017[114]
1R9O	Flurbiprofen (substrate)	2 Å	Wester *et al*., 2004[39]

The search for X-ray crystal structures of the human enzyme CYP450 2C9 with co-crystallized competitive inhibitors led to four X-ray structures. The PDB 4-letter codes of the protein-ligand complexes present were as follows: 5A5I (Skerratt *et al.,* 2016[111]), 5A5J (Skerratt *et al.,* 2016[111]), 4NZ2 (Brändén *et al.,* 2014[112]) and 5W0C (Liu *et al*., 2017[113]). Table 11 provides basic information about the four X-ray structures extracted from the PDB.

Table 11. Properties of the four X-ray structures of human CYP450 2C9 in complex with competitive inhibitors.

PDB 4-letter code	Resolution	R-free	R-work	Inhibitor (ChEMBL ID)	Inhibition value
5A5I	2 Å	0.245	0.212	ChEMBL3818307	Inhibition= 77% (3 µM)
5A5J	2.9 Å	0.288	0.213	ChEMBL3818248	Inhibition= 95% (3 µM)
4NZ2	2.45 Å	0.235	0.217	*Unspecified*	IC50= 31.8 nM
5W0C	2.001 Å	0.226	0.181	*Unspecified*	IC50= 100 nM

The resolution of the underlying data and the R-values (R-free, R-work) of the modelled 3D structures of the protein-ligand complexes were adequate in all four cases so that these X-ray crystal structures could be used for further work and for generating structure-based 3D-pharmacophore models of high quality. Even if 5A5J had the clearly weakest resolution (2.9 Å) among the four X-ray structures, it was still acceptable to use it for further purposes. Of particular interest, of course, were the inhibitory activity values of the co-crystallized competitive inhibitors. The two strongest competitive inhibitors in this case were the co-crystallized inhibitors from 5A5J, with an enzyme inhibition of 95.0% (at 3 µM) and the one from 4NZ2 with an IC50 of 31.8 nM. With a percentage enzyme inhibition of 77% (at 3µM) for the 5A5I competitive inhibitor and an IC50 of 100.0 nM for the 5W0C inhibitor, these two competitive inhibitors were less active than the other two. In the case of 5A5J and 5A5I, the ligands of the complexes were not completely surrounded by electron density. The ligands of 4NZ2 and 5W0C were completely surrounded by electron density. Unfortunately, despite intensive research, it was not possible to find out which reference compounds were involved in the four cases.

The corresponding structure-based 3D-pharmacophore models were generated fully automatically in the “Structure-Based” view of LigandScout.

Figures 11-14 represent the structure-based 3D pharmacophore models of 5A5I, 5A5J, 5W0C and 4NZ2. Furthermore, it is illustrated which amino acid side chains within the binding site were responsible for the enzyme-inhibitor interactions.

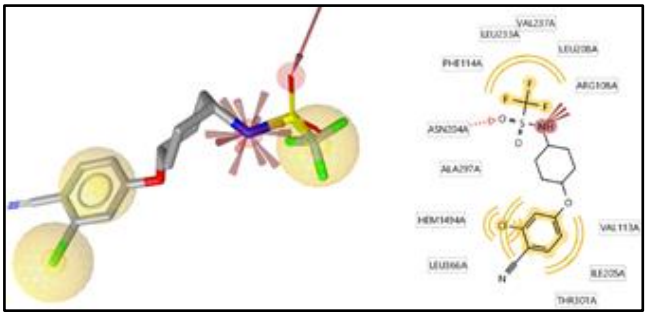


Figure 11. Structure-based pharmacophore model and key interactions between competitive inhibitor (ChEMBL ID: 3818307) and CYP 2C9 derived from 5A5I using the program LigandScout. The observed interactions are hydrophobic (yellow spheres), H-bond acceptor (red arrow) and a negative ionizable feature (red star).

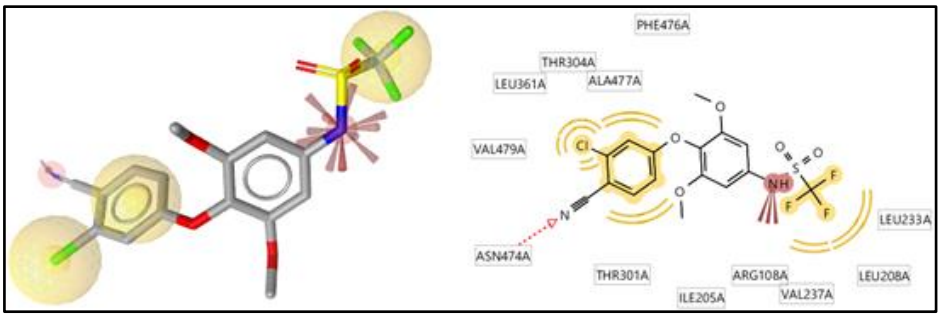


Figure 12. Structure-based pharmacophore model and key interactions between competitive inhibitor (ChEMBL ID: 3818248) and CYP 2C9 derived from 5A5J using the program LigandScout. The observed interactions are hydrophobic (yellow spheres), H-bond acceptor (red arrow) and a negative ionizable feature (red star).

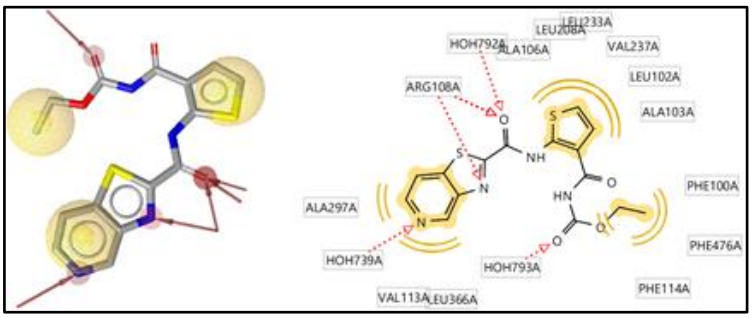


Figure 13. Structure-based pharmacophore model and key interactions between competitive inhibitor (ChEMBL ID: unspecified) and CYP 2C9 derived from 5W0C using the program LigandScout. The observed interactions are hydrophobic (yellow spheres) and H-bond acceptors (red arrows).

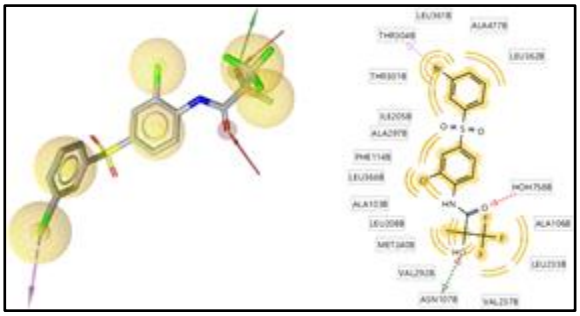


Figure 14. Structure-based pharmacophore model and key interactions between competitive inhibitor (ChEMBL ID: unspecified) and CYP 2C9 derived from 4NZ2 using the program LigandScout. The observed interactions are hydrophobic (yellow spheres), H-bond acceptors (red arrows), H-bond donor (green arrow) and an X-bond donor (purple arrow).

Since cytochrome P450 2C9 is of course a relatively unspecific enzyme, it was to be expected that also the interactions between competitive inhibitors and the amino acid side chains of the binding site would be different. Nevertheless, clear similarities could be observed regarding inhibitor-enzyme interactions between the four structure-based pharmacophore models. Figure 15 clearly shows that some amino acid side chains were of particular importance for enzyme inhibition, although it should be noted that these were almost exclusively lipophilic amino acids (Val237, Leu233, Leu208, Leu366, Thr301, Ala297, Ile205 and Phe114). The only exception was the basic amino acid Arg108, which seemed to be of major importance for interactions in most of the cases. These observations are supported by various publications dealing with the interaction between ligand and binding site of CYP 2C9 (e.g., Zhou *et al*., 2009[6]; Rettie *et al*., 2005[4]). Although publications mostly do not address the individual responsible unpolar amino acids, the relevance of this group of amino acids is always emphasized.

Although in literature the essential interaction between ligands and CYP 2C9 via Arg108 is usually described as an ionic interaction, there is a good chance that this interaction may also occur as an H-bond. The resulting structure-based 3D-pharmacophore model of 5W0C shows the interaction between inhibitor and binding site in the form of H-bonds with the amino acid Arg108. Another reaffirming point is the fact that almost none of the very strong inhibitor drugs contains negatively

ionizable features that would allow ionic interaction with Arg108 at physiological pH. Because of this fact, the interaction between ligand and CYP 2C9 via H-bonds seems quite plausible.

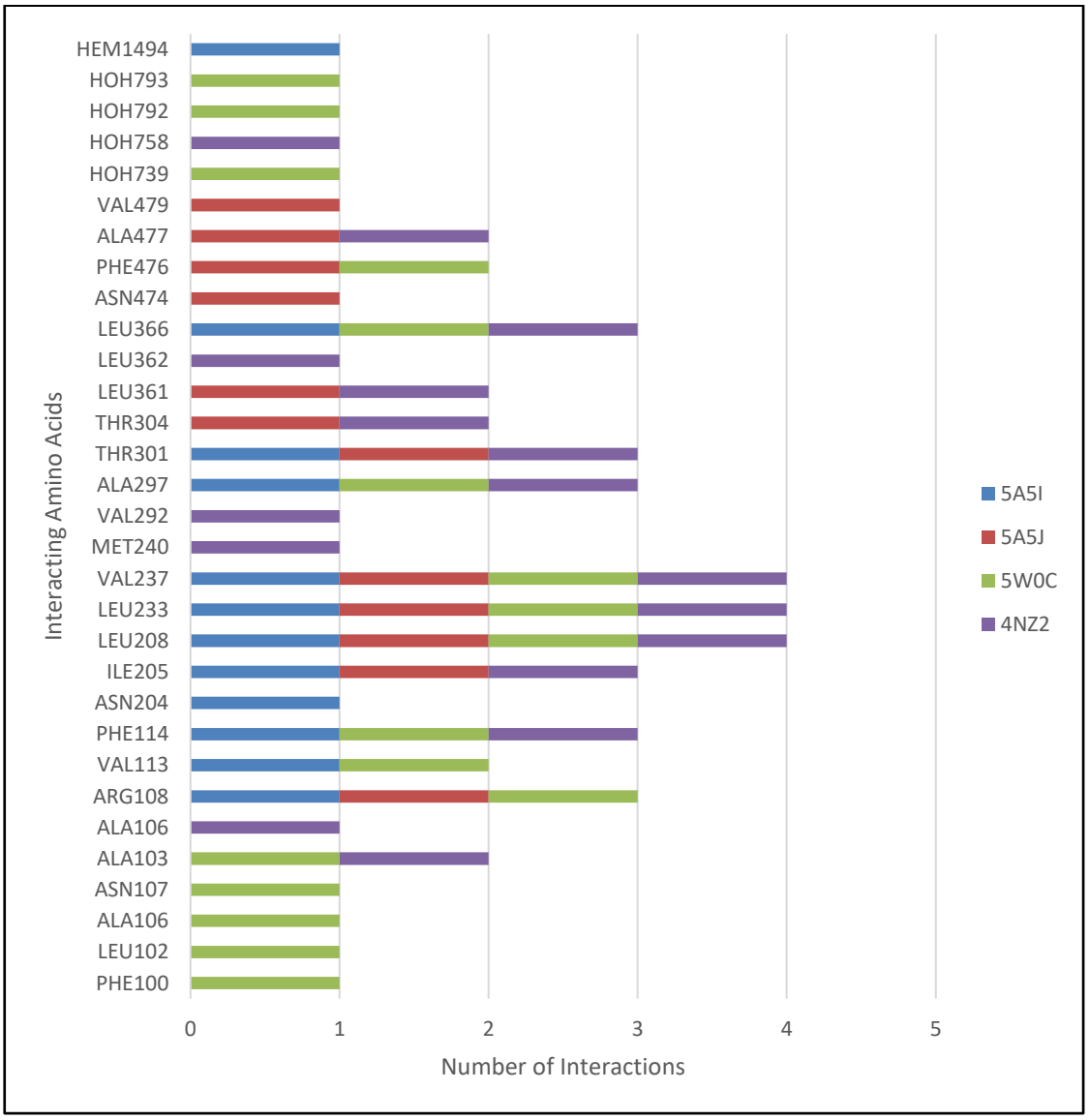


Figure 15. Observed interactions between competitive inhibitors and CYP2C9 derived from four X-ray derived experiments.

The four original structure-based models were screened against the 43 molecules of the very strong inhibitor training set in a first step. Figure 16 shows that the screening results were quite poor, because (except for 5A5I which at least recognized gemfibrozil) none of the models recognized even a single molecule as a hit. Gemfibrozil

is a lipid-lowering agent that inhibits CYP 2C9 with an IC50 of 9600 nM. The comparison of the two structures (gemfibrozil and the co-crystallized inhibitor of 5A5I) revealed clear structural similarities after alignment in LigandScout. (see Figure 17-18) With an inhibition of 77% (at 3µM), the co-crystallized inhibitor of 5A5I is quite comparable to gemfibrozil.

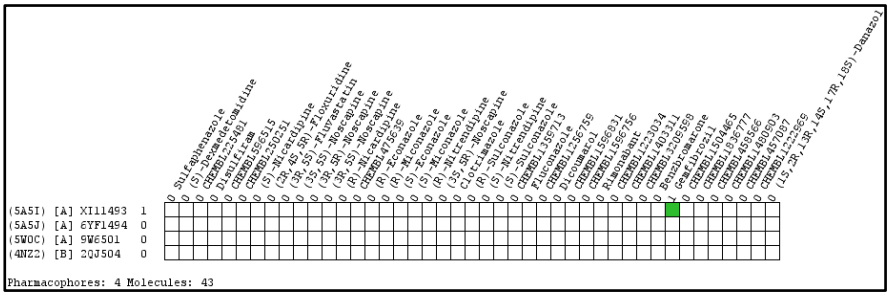


Figure 16. Hit retrieved from activity profiling using the four structure-based models against the very strong inhibitor training set compounds. The structure-based 3D-pharmacophore model from 5A5I was able to detect the competitive CYP 2C9 inhibitor gemfibrozil (IC50= 9600 nM).

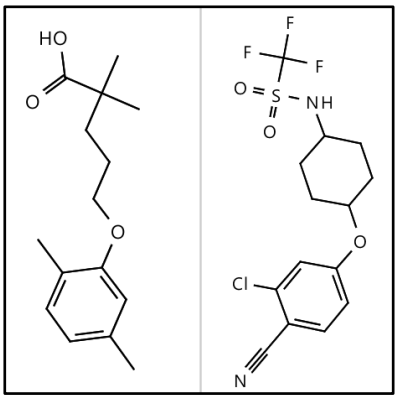


Figure 17. Comparison of the lipid-lowering agent gemfibrozil (left) with the co-crystallized CYP 2C9 inhibitor of 5A5I (right).

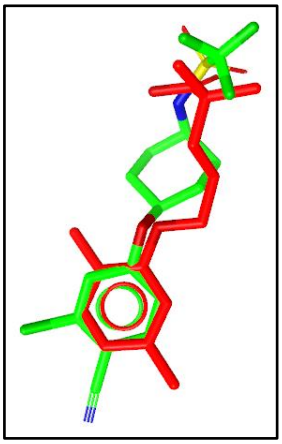


Figure 18. Alignment of the lipid-lowering agent gemfibrozil (red) and the co-crystallized CYP 2C9 inhibitor of 5A5I (green) in LigandScout.

Subsequently, the original structure-based 3D-pharmacophore models were screened against all other test sets (drugs and investigational compounds) to investigate their applicability for the detection of competitive CYP 2C9 inhibitors. Table 12-13 give detailed screening results of the four structure-based 3D-pharmacophore models with the diverse inhibitor drug and investigational inhibitor test sets.

Furthermore, it must be mentioned that the structure-based model generated from 5W0C could not even recognize its own ligand in a virtual screen. As a result, it was decided to do without 5W0C and its structure-based pharmacophore model and to continue working only with the other three models.

Table 12. Detailed screening results of the four structure-based 3D-pharmacophore models with the different CYP 2C9 drug test sets.

	5A5I	**5A5J**	**4NZ2**	**5W0C**
Very strong inhibitor drugs (43)	1 (2.3%)	0	0	0
Strong inhibitor drugs (25)	0	0	0	0
Moderate inhibitor drugs (8)	0	0	0	0
Weak inhibitor drugs (20)	0	0	0	0

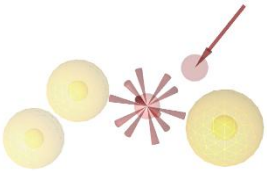
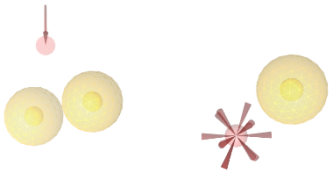
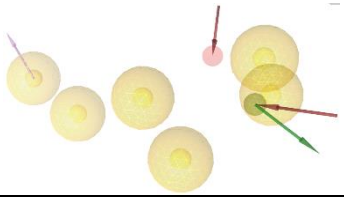
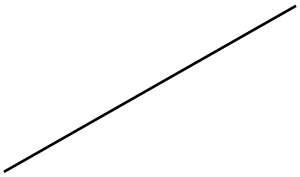
Table 13. Detailed screening results of the four structure-based 3D-pharmacophore models with the different CYP 2C9 investigational inhibitor test sets.

	5A5I	5A5J	4NZ2	5W0C
Very strong investigational inhibitors (850)	1 (0.12%)	0	0	0
Strong investigational inhibitors (2060)	9 (0.44%)	2 (0.09%)	0	0
Moderate investigational inhibitors (1490)	7 (0.47%)	0	0	0
Weak investigational inhibitors (2679)	9 (0.36%)	0	0	0

Subsequently, an attempt was made to generate a shared-feature pharmacophore model from the three leftover structure-based models (5A5I, 5A5J, 4NZ2). Perhaps a better result could be achieved with this approach and the essence of the interaction between inhibitors and enzyme could be presented with a shared model. Due to high inhibitory activity and good resolution, 4NZ2 was defined as the reference and subsequently the alignment was performed in LigandScout's "Alignment" view. The alignment of the 3D-pharmacophore models was performed taking into account reference points in the three-dimensional space to ensure that the alignment really made sense with respect to the binding site. Unfortunately, it was not possible to create a meaningful shared-feature pharmacophore model under these circumstances due to seemingly too big differences between the three structure-based models (the resulting shared-feature pharmacophore model consisted only of a single hydrophobic feature). Another attempt was to align the bioactive conformation of the three inhibitor ligands, but the subsequent generation of a shared-feature pharmacophore model led to an unacceptable performance.

The previously suggested assumption that the interaction between inhibitors and the binding site of CYP 2C9 could probably also be an H-bond enabled another possibility in the next step: Namely, that the negatively ionizable pharmacophore features of 5A5I and 5A5J were changed to H-bond acceptors, resulting in modified 3D pharmacophore models with new properties. Of course, the structure-based pharmacophore model of 4NZ2 was not changed since it did not contain any negative ionizable pharmacophore feature anyway. Table 14 gives information about the modification of the two pharmacophore models 5A5I and 5A5J.

Table 14. Refinement of structure-based pharmacophore models to improve prediction of very strong CYP 2C9 inhibitor drugs. Original and modified structure-based pharmacophore models.

	5A5I	5A5J	4NZ2
Orig.			
Mod.			

After the two structure-based models 5A5I and 5A5J had been changed as mentioned above, a new virtual screening job with the 43 very strong inhibitor training compounds was performed. Unfortunately, this modification of the pharmacophore features did not lead to any real improvement of the screening results. The modified structure-based model 5A5J again did not lead to a single hit and the modified model 5A5I only achieved two hits (gemfibrozil and ChEMBL1504465). Figure 19 shows the screening results of the two modified structure-based models based on an activity profile.

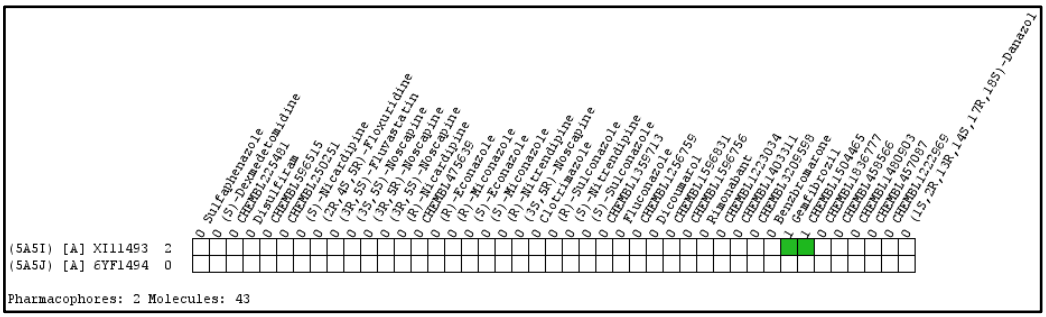


Figure 19. Activity profile of the modified structure-based models 5A5I and 5A5J.

As with the original structure-based pharmacophore models, the modified models of 5A5I and 5A5J and the original model 4NZ2 should now be used to generate a shared feature pharmacophore model, whereby again 4NZ2 was used as the reference and the pharmacophores should be aligned by reference points. This time it was possible to create such a shared-feature pharmacophore model, but unfortunately it consisted only of one pharmacophore feature, a hydrophobic interaction, making it not meaningful at all. Due to the fact that even the modified structure-based models did not bring any great success, this should be the end of structure-based approach in this thesis, and the focus would be on ligand-based modeling from that point on.

This first practical part of the work, structure-based 3D-pharmacophore modeling, provided interesting insights into the interaction of competitive antagonists and the binding site of the enzyme CYP 2C9. The structure-based 3D-pharmacophore models generated by the program LigandScout clearly confirmed what was theoretically described in section 2.3.1. In terms of structure-activity relationship (SAR), the great relevance of the interaction between competitive inhibitors and the basic amino acid Arg108 was clearly demonstrated, which apparently occurs either as an ionic interaction or as an H-bond. In addition to this essential interaction between molecule and enzyme, the numerous hydrophobic interactions within the binding site could be clarified. These reported interactions explain the selectivity of CYP 2C9 towards small, hydrophobic anions. Several attempts have been made to use the structure-based 3D pharmacophore models to identify competitive inhibitors of CYP 2C9 in virtual screening. Due to the complexity and high selectivity of the generated models, however, this was only partially achieved.

Since the aim of this diploma thesis was to generate 3D-pharmacophore models, which can detect as many inhibitors of CYP 2C9 as possible, especially very strong and strong inhibitors, the next step was the method of ligand-based 3D-pharmacophore modeling.

4.2.2 **Ligand-based modeling**

Figure 20 outlines the workflow used for the development of ligand-based pharmacophore models.

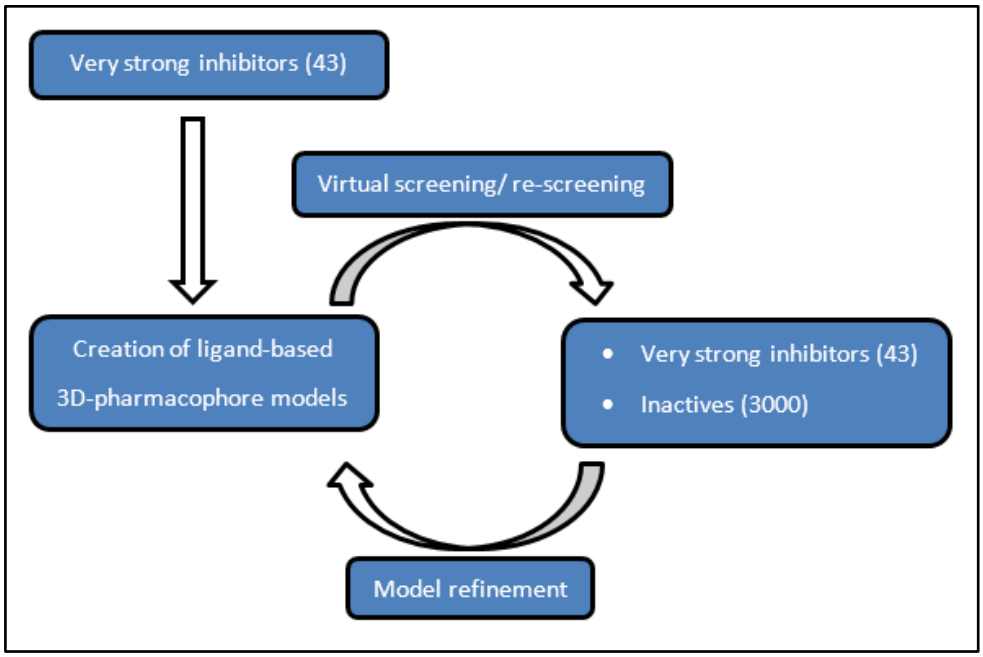


Figure 20. Basic workflow of ligand-based pharmacophore modeling to develop a ligand-based 3D-pharmacophore model for prediction of CYP 2C9 inhibitors.

In a first step the data set had to be loaded into the "Ligand-Based" view of LigandScout 4.4, and then all the possible conformations (maximum number of conformations per molecule was set 25) of the compounds had to be created. After conformation generation, the ligand set was clustered by pharmacophore similarity using LigandScout's Pharmacophore RDF-Code Similarity algorithm default settings. The step of ligand clustering based on pharmacophore features was the basis for the later ligand-based pharmacophore modeling, and therefore it was of particular importance. Figures 21-22 show all the detailed settings of conformation generation and ligand-clustering.

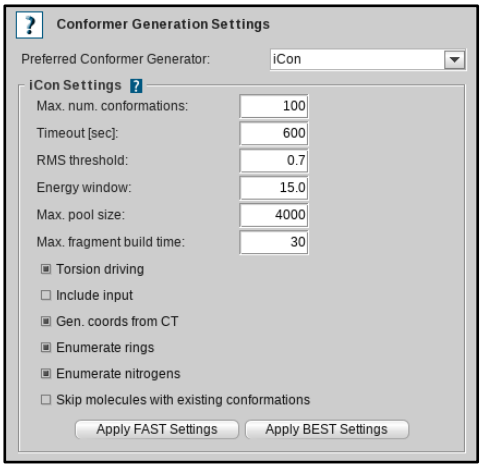


Figure 21. Conformer generation settings of LigandScout 4.4.

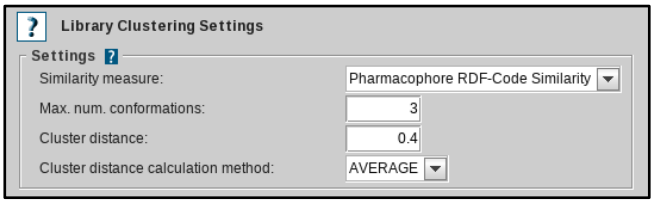


Figure 22. Library clustering settings of LigandScout 4.4.

Originally, the clustering process led to a total number of 16 clusters, but since some of the compound's stereoisomers were deleted during the research for the active enantiomers and diastereomers (as described in the section of data set creation), cluster 15 and cluster 1 were sorted out. As a result, the 43 compounds were now actually located within 14 clusters, even though the cluster names suggest a number of 16 clusters**.** Table 15 gives detailed information about the distribution of the very active training compounds into the different clusters.

Table 15. Very active training compounds- cluster IDs.

Compound Name	Cluster ID
CHEMBL596515	16
Disulfiram	16
CHEMBL225481	16
(S)-Dexmedetomidine	16
(3S,5R)-Noscapine	14
(3R,5S)-Noscapine	14
(3R,5R)-Noscapine	14
(3S,5S)-Noscapine	14

Compound Name	Cluster ID
CHEMBL1596756	8
CHEMBL1596831	8
Clotrimazole	8
CHEMBL1256759	7
Fluconazole	7
CHEMBL1359713	7
CHEMBL1222969	6
CHEMBL1223034	6

Sulfaphenazole	14
(R)-Nicardipine	13
(3R,5S)-Fluvastatin	13
(S)-Nicardipine	13
CHEMBL250251	13
(2R,4S,5R)-Floxuridine	12
(S)-Nitrendipine	11
(R)-Nitrendipine	11
CHEMBL475639	11
(S)-Miconazole	10
(S)-Econazole	10
(R)-Miconazole	10
(R)-Econazole	10
(S)-Sulconazole	9
(R)-Sulconazole	9

Rimonabant	6
CHEMBL1836777	5
CHEMBL1403311	5
Dicoumarol	5
(1S,2R,13R,14S,17R,18S)-Danazol	4
CHEMBL1480903	4
CHEMBL1504465	3
Benzbromarone	3
CHEMBL3209598	3
CHEMBL457087	2
CHEMBL458566	2
Gemfibrozil	2

The step of conformation generation and clustering was followed by the actual creation of ligand-based 3D-pharmacophore models. Basically, pharmacophore models were created for each of the clusters, but since many of the clusters led to inefficient pharmacophore models, only the most relevant ones are described in this section. For all of the models, the same pharmacophore-creation settings were used and were defined as follows:

- **Scoring function:** Pharmacophore fit and atom overlap
- **Pharmacophore type:** Merged feature pharmacophore
- **Number of omitted features for merged pharmacophore:** 4
- **Partially matching features optional, threshold (%):** 10.0
- **Create exclusion volumes:** Yes
- **Apply custom feature definitions:** Yes
- **Feature tolerance scale factor:** 1.0
- **Max. number of result pharmacophores:** 10

4.2.2.1 CYP2C9_Inhib._Model1

For the creation of ***CYP2C9_Inhib._Model1*** cluster 3, which contained the compounds CHEMBL1504465, benzbromarone and CHEMBL3209598, was used. This cluster was of great interest for the generation of ligand-based 3D-pharmacophore models, as the uricosuric drug benzbromarone is considered to be one of the most potent inhibitors of CYP 2C9.

The initial pharmacophore model consisted of the following pharmacophore features: Four hydrophobic interactions (one optional), two H-bond acceptors (one optional), one H-bond donor, two aromatic rings and two halogen-bond donors (both optional, both vector features). Table 16 gives a brief description of the initial model. In LigandScout's "Screening" view, a primary virtual screen of ***CYP2C9_Inhib._Model1*** against the training set (43 very active ligands, 3000 inactive ligands) led to five true positives and no false positives. Two of the five active hit list molecules were compounds that had been used to generate this pharmacophore model. Thus, this original pharmacophore model was able to find three additional molecules from the very strong inhibitor compounds of the training set (see Table 17).

With an EF of 70.8 (at 1%, 5%, 10%, 100%) the chance is consistently 70.8 times higher to receive a true positive (very strong inhibitor) with the initial pharmacophore model than a false positive (inactive compound). An AUC of 1.00 (at 1%, 5% and 10%) means that hits with the highest pharmacophore-fit score are only true positives. Especially the early EF and AUC (at 1%, 5% and 10%) are valuable parameters for assessing the performance of the pharmacophore model.

Table 16. Brief description of *CYP2C9_Inhib._Model1* (before model refinement), aligned with the very strong inhibitor drug benzbromarone.

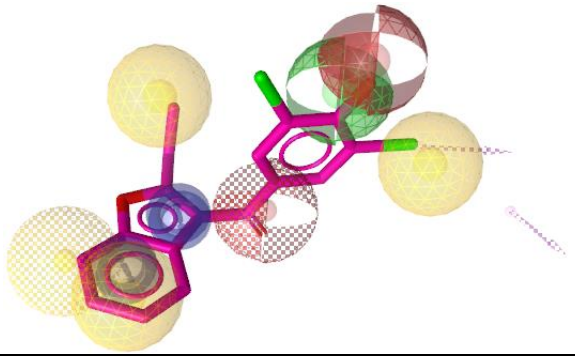
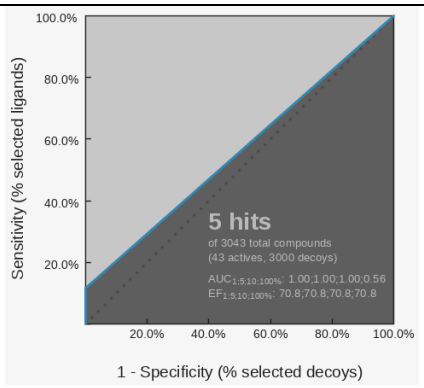
Model Name /Template Molecules	Figure of the Model /Model Features
Cluster 3	
CYP2C9_Inhib._Model1	
CHEMBL1504465 Benzbromarone CHEMBL3209598	4 H (1 optional) 2 HBA (1 optional) 1 HBD 2 AR 2 XBD (2 optional)

Table 17. Detailed screening results of *CYP2C9_Inhib._Model1* in training (before model refinement).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	5
False Positives	0
EF (1, 5, 10, 100%)	70.8, 70.8, 70.8, 70.8
AUC (1, 5, 10, 100%)	1.00, 1.00, 1.00, 0.56
Sensitivity (100%)	11.62
Specificity (100%)	100.0

Table 18-19 provide a summary of the refined model and the subsequent screening results of the training phase.

Table 18. Brief description of *CYP2C9_Inhib._Model1* (after model refinement), aligned with the retrieved very strong inhibitor drug (3R,5S)-fluvastatin.

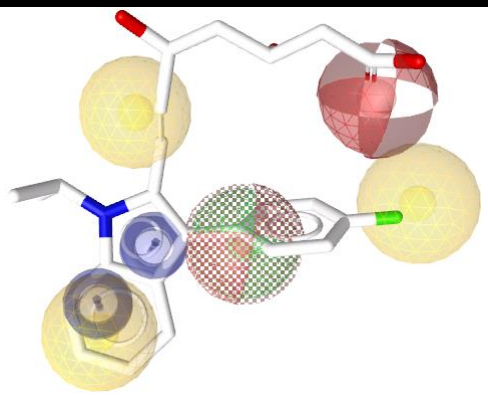
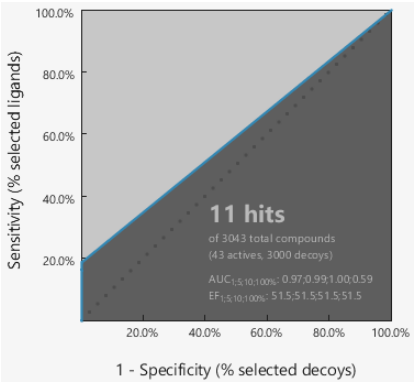
Model Name	Figure of the Model /Model Features
Cluster 3	
CYP2C9_Inhib._Model1	
3 H 2 HBA (1 optional) 1 HBD (1 optional) 2 AR	

Table 19. Detailed screening results of *CYP2C9_Inhib._Model1* in training (after model refinement).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	8
False Positives	3
EF (1, 5, 10, 100%)	51.5, 51.5, 51.5, 51.5
AUC (1, 5, 10, 100%)	0.97, 0.99, 1.00, 0.59
Sensitivity (100%)	18.60
Specificity (100%)	99.90

In the course of model refinement, the original pharmacophore model (11 pharmacophore features, including four optional ones) was heavily modified and reduced to eight pharmacophore features (two of them optional). The result of this

refinement was a pharmacophore model, which achieved eight true positives instead of five true positives. The number of false positives inevitably increased from zero to three, which explains a consistently slightly lower EF of 51.5 and an AUC of 0.97 (1%), 0.99 (5%), 100.0 (10%), and 0.59 (100%). Even though the refinement of this model led to slightly worse values of the mentioned parameters, it must be noted at this point that they are nevertheless still very good, and this compromise was gladly accepted for the gain in true positives. The increase, both true positive and false positive, is due in large part to the reduced number of pharmacophore features as the model thus lost selectivity. Optional features were used specifically to improve the ranking according to the pharmacophore-fit score.

4.2.2.2 *CYP*2C9_Inhib._Model2

For the next pharmacophore model, ***CYP2C9_Inhib._Model2***, cluster 6 served as a template. This cluster contained the following compounds: Rimonabant, CHEMBL1223034 and CHEMBL1222969. With an IC50 of 5 nM, the drug rimonabant, which has already been withdrawn from the market, represents the second most active molecule of the 43 training set compounds. Thus, the now discussed cluster 6 was of great importance for the generation of another ligand-based 3D-pharmacophore model.

The following pharmacophore features were part of the initial model: Five hydrophobic interactions, four H-bond acceptors (two optional), five aromatic rings (two optional) and three halogen-bond donors (all vector features). Table 20-21 give information about the initial pharmacophore model and the screening results before model refinement.

Table 20. Brief description of *CYP2C9_Inhib._Model2* (before model refinement), aligned with the very strong inhibitor drug rimonabant.

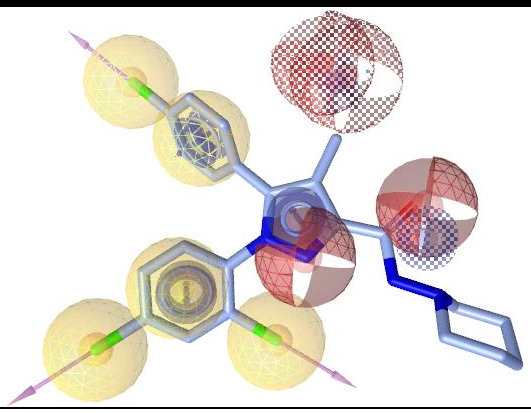
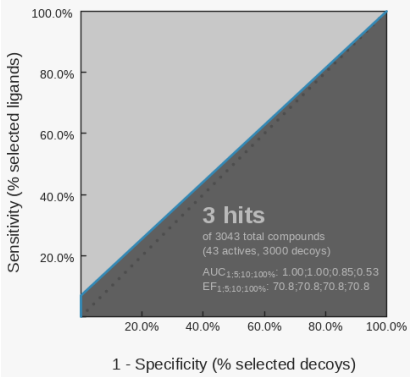
Model Name /Template Molecules	Figure of the Model /Model Features
Cluster 6	
CYP2C9_Inhib._Model2	
Rimonabant CHEMBL1223034 CHEMBL1222969	5 H 4 HBA (2 optional) 5 AR (2 optional) 3 XBD

Table 21. Detailed screening results of *CYP2C9_Inhib._Model2* in training (before model refinement).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	3
False Positives	0
EF (1, 5, 10, 100%)	70.8, 70.8, 70.8, 70.8
AUC (1, 5, 10, 100%)	1.00, 1.00, 0.85, 0.53
Sensitivity (100%)	6.98
Specificity (100%)	100.0

With 17 pharmacophore features (four of which were optional), this initial model was a highly specific and large model that resulted in only three hits. These were only true positives and not false positives. With a continuous EF of 70.8 and an AUC of 1.00 (1%),

1.00 (5%), 0.85 (10%) and 0.53 (100%), the two statistical parameters were accordingly high. However, the goal of detecting as many very strong inhibitors of CYP 2C9 as possible was not yet achieved with this highly selective model, which led to the extensive refinement of this ligand-based pharmacophore model.

Intensive refinement and a reduction in the number of pharmacophore features by more than half led to the desired result. Namely, a true positive increase from three to nine and an acceptable increase of false positive from zero to two. Thus, the pharmacophore model was now able to recognize three times as many very strong inhibitors as before refinement. Due to the practically unavoidable small increase in false positives, the EF dropped to a still high value of 57.9. Due to the very good ranking of the hit molecules according to their pharmacophore-fit score, the AUC even increased to 1.00 (1%), 1.00 (5%), 1.00 (10%) and 0.60 (100%). Thus, the original extremely selective pharmacophore model could be trimmed to a much more efficient one. Table 22-23 give information about the properties of ***CYP2C9_Inhib._Model2*** and the corresponding screening results in training after model refinement.

Table 22. Brief description of *CYP2C9_Inhib._Model2* (after model refinement), aligned with the retrieved very strong inhibitor drug (R)-sulconazole.

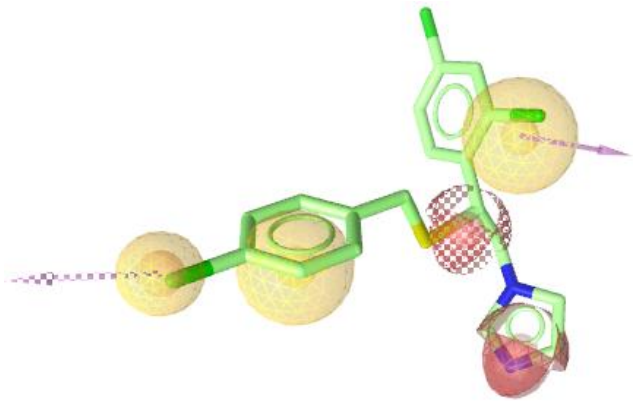
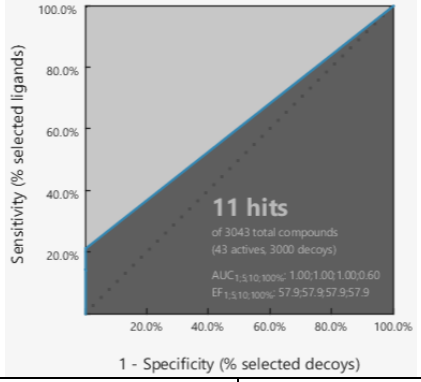
Model Name /Template Molecules	Figure of the Model /Model Features
	Cluster 6
2C9_Inhib._Model2	
Rimonabant CHEMBL1223034 CHEMBL1222969	3 H 2 HBA (1 optional) 2 XBD (1 optional)

Table 23. Detailed screening results of *CYP2C9_Inhib._Model2* in training (after model refinement).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	9
False Positives	2
EF (1, 5, 10, 100%)	57.9, 57.9, 57.9, 57.9
AUC (1, 5, 10, 100%)	1.00, 1.00, 1.00, 0.60
Sensitivity (100%)	20.93
Specificity (100%)	99.93

4.2.2.3 CYP2C9_Inhib._Model3

The next ligand-based pharmacophore model was generated from cluster 10. This cluster consisted of two drugs, miconazole and econazole, each drug with both enantiomers. Since azole antifungals are one of the most well-known inhibitors of CYP 2C9, cluster 10 was of great interest for further pharmacophore modeling. The resulting pharmacophore model was characterized by the following pharmacophore features: Six hydrophobic interactions (one optional), two H-bond acceptors, three aromatic rings and four halogen-bond donors (all vector features, one of them optional). Table 24-25 give information about the initial model and the first screening results in training of ***CYP2C9_Inhib._Model3*** before refinement.

Table 24. Brief description of *CYP2C9_Inhib._Model3* (before model refinement).

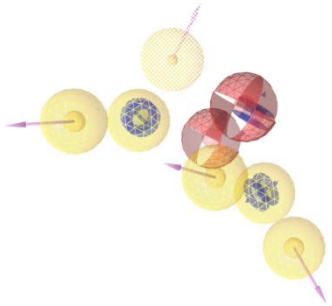
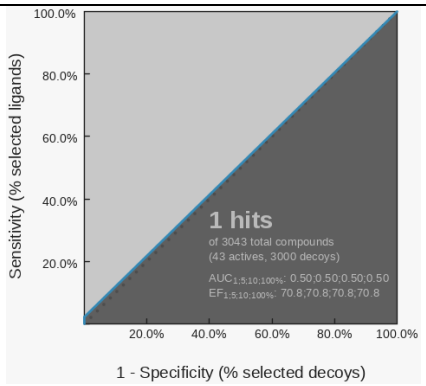
Model Name /Template Molecules	Figure of the Model /Model Features
Cluster 10	
CYP2C9_Inhib._Model3	
(R)-Miconazole (S)-Miconazole (R)-Econazole (S)-Econazole	6 H (1 optional) 2 HBA 3 AR 4 XBD (1 optional)

Table 25. Detailed screening results of *CYP2C9_Inhib._Model3* in training (before model refinement).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	1
False Positives	0
EF (1, 5, 10, 100%)	70.8, 70.8, 70.8, 70.8
AUC (1, 5, 10, 100%)	0.50, 0.50, 0.50, 0.50
Sensitivity (100%)	2.33
Specificity (100%)	100.0

With a number of 15 pharmacophore features, of which only two were optional, the initial model ***CYP2C9_Inhib._Model3*** was a highly selective pharmacophore model, which was not convincing in the first virtual screening. Only a single hit, albeit a true positive, could be achieved and a consistent AUC value of 0.5 testified to the rather poor performance of the model. With a very high continuous EF of 70.8, this

parameter was very good, but this is only due to the fact that no false positive was generated. In order to be able to apply this pharmacophore model in the future, it had to be greatly reduced in its complexity and selectivity in order to increase its ability to recognize very strong inhibitors of CYP 2C9.

Again, the details of the model and the training screen after refinement are shown in Table 26-27.

Table 26. Brief description of *CYP2C9_Inhib._Model3* (after model refinement), aligned with the retrieved very strong inhibitor drug (S)-econazole.

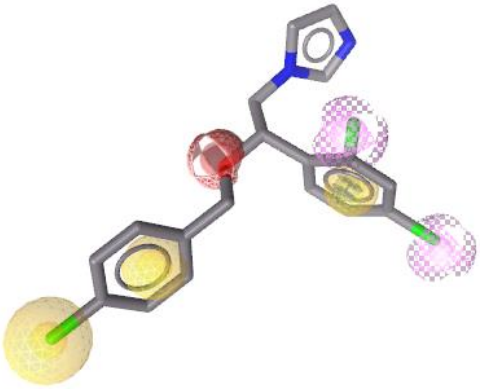
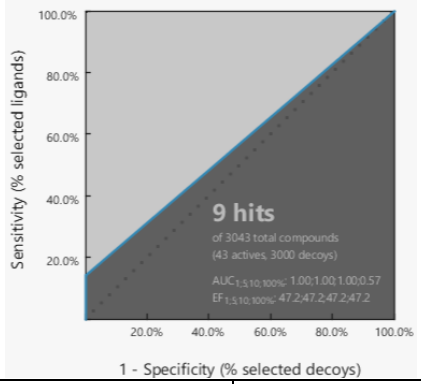
Model Name /Template Molecules	Figure of the Model /Model Features
Cluster 10	
CYP2C9_Inhib._Model3	
(R)-Miconazole (S)-Miconazole (R)-Econazole (S)-Econazole	3 H 1 HBA 1 AR 2 XBD (2 optional)

Table 27. Detailed screening results of *CYP2C9_Inhib._Model3* in training (after model refinement).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	6
False Positives	3
EF (1, 5, 10, 100%)	47.2, 47.2, 47.2, 47.2
AUC (1, 5, 10, 100%)	1.00, 1.00, 1.00, 0.57
Sensitivity (100%)	13.95
Specificity (100%)	99.90

Refinement of the original pharmacophore model led to the modification of the pharmacophore features and the reduction of 15 to 7 features. As expected, the number of hits increased from one to nine. However, three of the nine hits were false positives, which is the reason for the lower EF of 47.2 (1%, 5%, 10%, 100%) compared to the previous two models. Nevertheless, ***CYP2C9_Inhib._Model3*** was now able to detect significantly more very strong inhibitors than before, and also the AUC of 1.00 (1%), 1.00 (5%), 1.00 (10%) and 0.57 (100%) spoke in favor of the quality of this pharmacophore model.

4.2.2.4 CYP2C9_Inhib._Model4

For the fourth ligand-based 3D-pharmacophore model, no specific cluster was used this time. Rather, this time very strong inhibitors of CYP 2C9 were manually selected. The following three substances were used: Clotrimazole (IC50= 136 nM), (R)-sulconazole (IC50= 200 nM) and CHEMBL1359713 (potency= 100 nM).

The originally generated model contained the following pharmacophore features: Four hydrophobic interactions (one optional), two H-bond acceptors (one optional), two aromatic rings (one optional) and two halogen-bond donors (both vector features, one of them optional). The two following tables, Table 28-29, provide information about the initial ***CYP2C9_Inhib._Model4*** and its training results.

Although the number of features (ten pharmacophore features, four of which are optional) was rather small compared to the previous models in this initial pharmacophore model, the number of hits obtained was not very large. Four true positives and two false positives were the reason for the rather low EF of 47.2, compared to the previous models. An AUC value of 1.00 (1%), 1.00 (5%), 0.96 (10%) and 0.55 (100%), however, reflects the good ranking of the true positives in the hit list according to their pharmacophore-fit score.

Table 28. Brief description of *CYP2C9_Inhib._Model4* (before model refinement), aligned with the very strong inhibitor drug (S)-sulconazole.

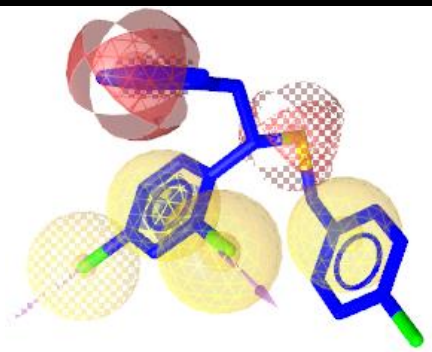
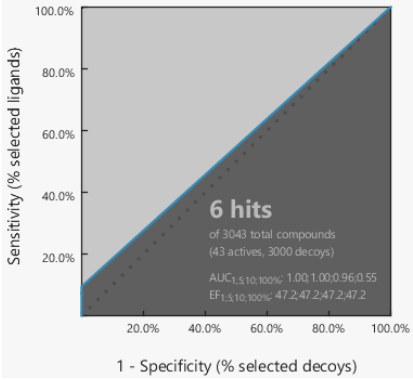
Model Name /Template Molecules	Figure of the Model /Model Features
CYP2C9_Inhib._Model4	
Clotrimazole (R)-Sulconazole CHEMBL1359713	4 H (1 optional) 2 HBA (1 optional) 2 AR (1 optional) 2 XBD (1 optional)

Table 29. Detailed screening results of *CYP2C9_Inhib._Model4* in training (before model refinement).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	4
False Positives	2
EF (1, 5, 10, 100%)	47.2, 47.2, 47.2, 47.2
AUC (1, 5, 10, 100%)	1.00, 1.00, 0.96, 0.55
Sensitivity (100%)	9.30
Specificity (100%)	99.93

This ligand-based pharmacophore model also underwent an elaborate refinement, which in turn improved its quality and performance. The aim of this refinement was, as expected, to change the properties of the pharmacophore model such that the number of true positive increased without further increasing the number of false positives.

A particularly intensive refinement and remodeling of the original pharmacophore model led to the ideal case that the number of true positives could be increased from four to six and the number of false positives could be reduced from two to one. The significantly improved performance of the pharmacophore model was confirmed by a strong gain in ER (60.7) and an increased AUC of 1.00 (1%), 1.00 (5%), 1.00 (10%), and 0.57 (100%).

Description and detailed screening results of the refined model ***CYP2C9_Inhib._Model4*** are shown in Table 30-31.

Table 30. Brief description of *CYP2C9_Inhib._Model4* (after model refinement), aligned with the retrieved very strong investigational inhibitor ChEMBL1223034.

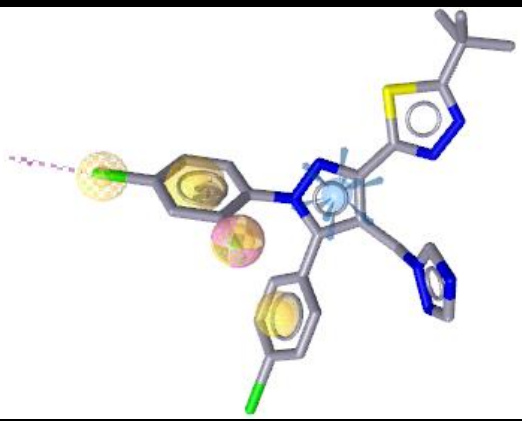
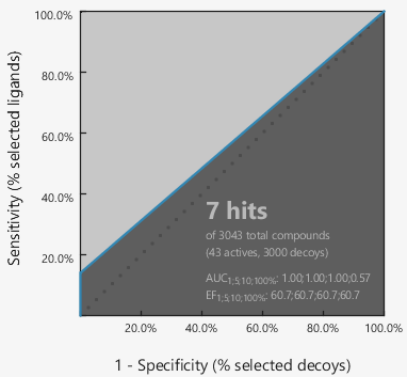
Model Name /Template Molecules	Figure of the Model /Model Features
CYP2C9_Inhib._Model4	
Clotrimazole (R)-Sulconazole CHEMBL1359713	4 H (1 optional) 1 AR 2 XBD (1 optional) 1 Iron-binding location (FEB)

Table 31. Detdetailed screening results of *CYP2C9_Inhib._Model4* in training (after model refinement).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	6
False Positives	1
EF (1, 5, 10, 100%)	60.7, 60.7, 60.7, 60.7
AUC (1, 5, 10, 100%)	1.00, 1.00, 1.00, 0.57
Sensitivity (100%)	13.95
Specificity (100%)	99.97

Figure 23 visualizes which of the active training compounds had already been covered by the four ligand-based models at this point. This so-called “activity profile” shows which compounds had already been recognized by a model and the intensity of the color showed how strongly a pharmacophore model fit with a respective molecule.

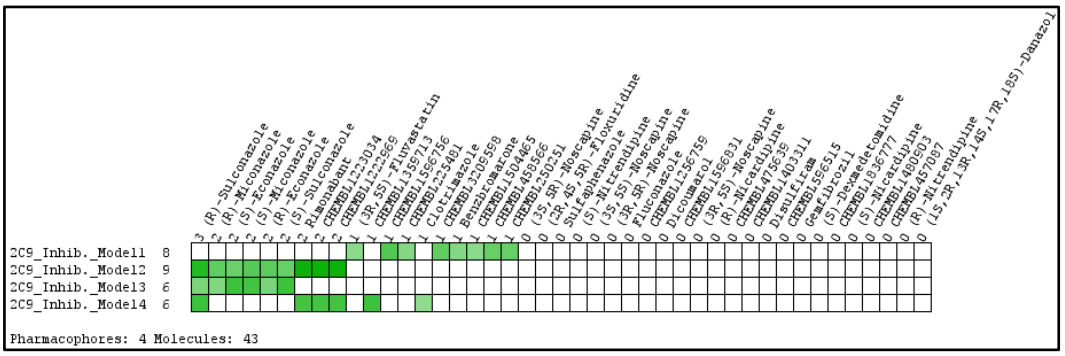


Figure 23. Activity profile of the four ligand-based pharmacophore models (CYP2C9_Inhib._Model1-4) after refinement in training.

Since the four ligand-based pharmacophore models mentioned together covered 19 (44%) of the 43 very strong inhibitors of the training set, two further pharmacophore models should be generated. The goal was to create as many ligand-based 3D-pharmacophore models to retrieve at least 50% of the very strong inhibitors from the training set. Because the more very strong inhibitors from the training set could be retrieved by the various pharmacophore models, the higher were the chances of achieving good results in the following test phase and in later practical application.

To make sure that these two additional models complemented the existing models and did not lead to hits that were already achieved by other models, a slightly different approach should be adopted. For this purpose, the 24 (56%) active training compounds, which had not yet been identified by any of the four ligand-based models, were clustered separately and then the two largest clusters were selected for the generation of the two additional models ***CYP2C9_Inhib._Model5*** and ***CYP2C9_Inhib._Model6***. In this way, the chances were very good to cover additional areas of the very strong inhibitor training set compounds and to record a broader range of possible arrangements of pharmacophore features in three-dimensional space.

4.2.2.5 CYP2C9_Inhib._Model5

After the already mentioned re-clustering of the 24 very active training compounds not yet identified by the four models (***CYP2C9_Inhib._Model1-4***), the largest cluster containing five molecules was used for this fifth ligand-based model. These five molecules were the four diastereomers of noscapine (potency= 125 nM) and the very strong inhibitor sulfaphenazole (IC50= 180 nM). The resulting ligand-based model ***CYP2C9_Inhib._Model5*** consisted of the following pharmacophore features: Two hydrophobic interactions (one optional), four H-bond acceptors (three optional) and one aromatic ring. The following two tables (Table 32-33) provide information on the composition and screening results of this initial pharmacophore model.

Table 32. Brief description of *CYP2C9_Inhib._Model5* (before model refinement), aligned with the very strong inhibitor drug (3R,5R)-noscapine.

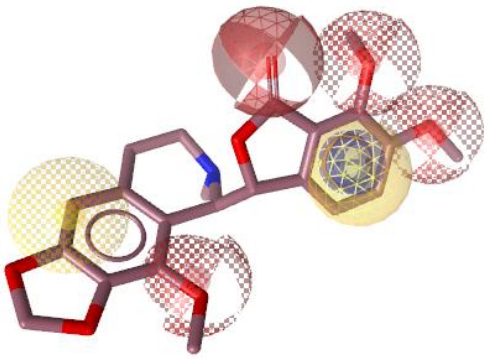
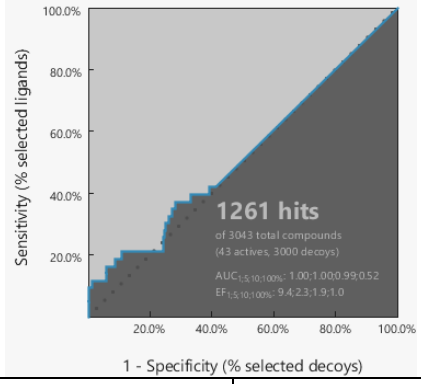
Model Name /Template Molecules	Figure of the Model /Model Features
CYP2C9_Inhib._Model5	
(3S,5R)-Noscapine (3R,5S)-Noscapine (3R,5R)-Noscapine (3S,5S)-Noscapine Sulfaphenazole	2 H (1 optional) 4 HBA (3 optional) 1 AR

Table 33. Detailed screening results of *CYP2C9_Inhib._Model5* in training (before model refinement).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	18
False Positives	1243
EF (1, 5, 10, 100%)	9.4, 2.3, 1.9, 1.0
AUC (1, 5, 10, 100%)	1.00, 1.00, 0.99, 0.52
Sensitivity (100%)	41.86
Specificity (100%)	58.57

The results of virtual screening with the initial pharmacophore model were unsatisfactory. Although the number of true positives (18) was very high, it faced a far higher number of false positives (1243), which gave the bad EF of 9.4 (1%), 2.3 (5%), 1.9 (10%) and 1.0 (100%) explained. On closer inspection, it quickly becomes clear what the reason for this extraordinarily poor performance was. Namely, the fact that although the initial pharmacophore model consisted of seven pharmacophore features, four of them were optional. The model was actually too unselective to properly distinguish between active and inactive. Thus, the refinement goal was to dramatically increase the selectivity of the pharmacophore model and reduce the number of false positives.

Tables 34-35 show ***CYP2C9_Inhib._Model5*** and the screening results in training after refinement of this pharmacophore model.

Table 34. Brief description of *CYP2C9_Inhib._Model5* (after model refinement), aligned with the retrieved very strong inhibitor drug fluconazole.

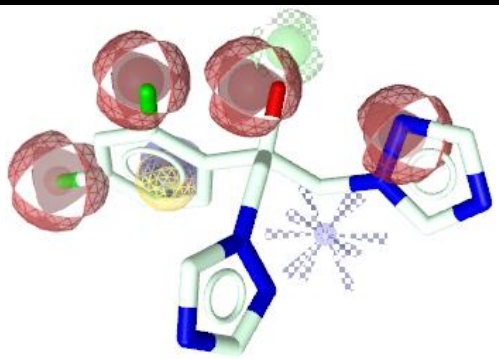
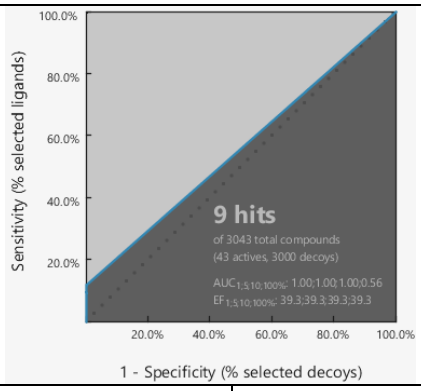
Model Name /Template Molecules	Figure of the Model /Model Features
CYP2C9_Inhib._Model5	
(3S,5R)-Noscapine (3R,5S)-Noscapine (3R,5R)-Noscapine (3S,5S)-Noscapine Sulfaphenazole	1 H 4 HBA 1 HBD (optional) 1 AR 1 PI (optional)

Table 35. Detailed screening results of *CYP2C9_Inhib._Model5* in training (after model refinement).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	5
False Positives	4
EF (1, 5, 10, 100%)	39.3, 39.3, 39.3, 39.3
AUC (1, 5, 10, 100%)	1.00, 1.00, 1.00, 0.56
Sensitivity (100%)	11.63
Specificity (100%)	99.87

After refinement, ***CYP2C9_Inhib._Model5*** consisted of eight pharmacophore features, two of which were optional. The significant increase in features and the associated gain in selectivity reduced the number of true positives to five and the number of false

positives to four. Performance improved in this way is evidenced by an EF of 39.3 and an AUC of 1.00 (1%), 1.00 (5%), 1.00 (10%) and 0.56 (100%). Thus, from an extremely nonspecific pharmacophore model, which could hardly distinguish between active and inactive compounds, a satisfactory model could be generated during the refinement.

4.2.2.6 CYP2C9_Inhib._Model6

For the sixth and thus last ligand-based pharmacophore model of the present work, the next largest cluster after the aforementioned re-clustering was used. It contained the following molecules: (R)-nicardipine, (S)-nicardipine, (R)-nitrendipine and (S)-nitrendipine. Based on these four molecules, the initial ligand-based model ***CYP2C9_Inhib._Model6*** was created and consisted of the following pharmacophore features: Three hydrophobic interactions, four H-bond acceptors, one H-bond donor (vector feature) and one aromatic ring. Since the initial version of ***CYP2C9_Inhib._Model6*** already performed very selective and specific, leading to no false positives, no further refinement of this ligand-based model was necessary. The reason for the high selectivity of this model was, of course, the fact that it consisted of nine pharmacophore features, none of which was optional. A continuous EF of 70.8 and an AUC of 1.00 (1%), 1.00 (5%), 0.97 (10%) and 0.55 (100%) testified to the very good performance of ***CYP2C9_Inhib._Model6***.

Table 36-37 give information about the properties of the model and the screening results during test phase.

Table 36. Brief description of *CYP2C9_Inhib._Model6*, aligned with the very strong inhibitor drug (R)-nitrendipine.

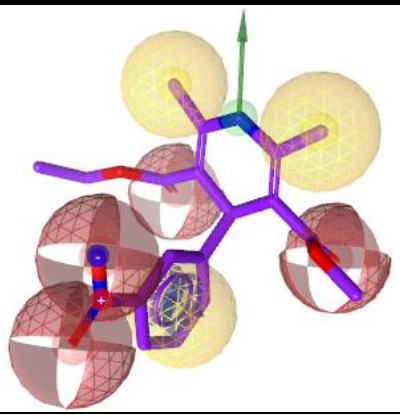
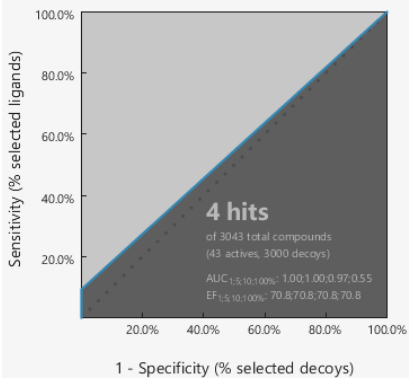
Model Name /Template Molecules	Figure of the Model /Model Features
CYP2C9_Inhib._Model6	
(R)-Nicardipine (S)-Nicardipine (R)-Nitrendipine (S)-Nitrendipine	3 H 4 HBA 1 HBD 1 AR

Table 37. Detailed screening results of *CYP2C9_Inhib._Model6* in training.



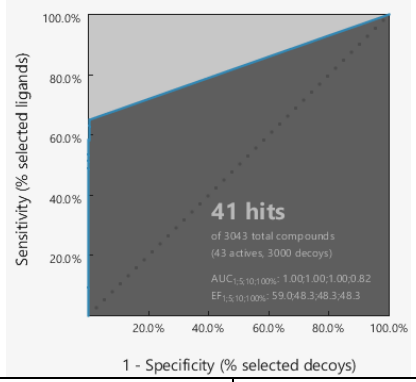
Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	4
False Positives	0
EF (1, 5, 10, 100%)	70.8, 70.8, 70.8, 70.8
AUC (1, 5, 10, 100%)	1.00, 1.00, 0.97, 0.55
Sensitivity (100%)	9.30
Specificity (100%)	100.0

The refinement of these six ligand-based 3D-pharmacophore models consisted essentially of marking features optional or obligate, changing feature tolerances or feature weights. Sometimes the position of features in 3D space was slightly changed,

or features were deleted, sometimes new ones were created. Since the refinement of the original models to the final pharmacophore models required uncountable steps of slight changes and re-screening jobs, only the initial and the final models are shown for each of the six models in the pages above. The exact description of the individual refinement steps of each of the six models would have taken too long and probably went beyond the scope of this work.

Interestingly, the previously presented ligand-based 3D-pharmacophore models covered different areas of the actives of the training set and completed each other extraordinarily well. Table 38 shows the ROC curve and detailed screening results of the six ligand-based pharmacophore models (***CYP2C9_Inhib._Model1-6***) in a parallel screen against the 43 actives and the 3000 inactives of the training set. Figure 24 shows which areas of active training compounds were covered by the six pharmacophore models and which were not. The plan to apply the six ligand-based 3D-pharmacophore models in the form of parallel screening offers a great advantage over the classical application of a single pharmacophore model. Since the different ligand-based models ***CYP2C9_Inhib._Model1-6*** differ significantly in terms of their pharmacophore features and their orientation in three-dimensional space, they can together cover the space of CYP 2C9 inhibitors much better and more comprehensive than individual models would be able to do so. This also increases the chance of predicting future potential inhibitors of CYP 2C9 during early drug discovery.

Table 38. Screening results of the final six ligand-based models *(CYP2C9_Inhib._Model1-6)* in training (parallel screen).



Total Database Compounds	3043
Active Ligands in Database	43
Inactive Ligands in Database	3000
True Positives	28
False Positives	13
EF (1, 5, 10, 100%)	59.0, 48.3, 48.3, 48.3
AUC (1, 5, 10, 100%)	1.00, 1.00, 1.00, 0.82
Sensitivity (100%)	65.12
Specificity (100%)	99.57

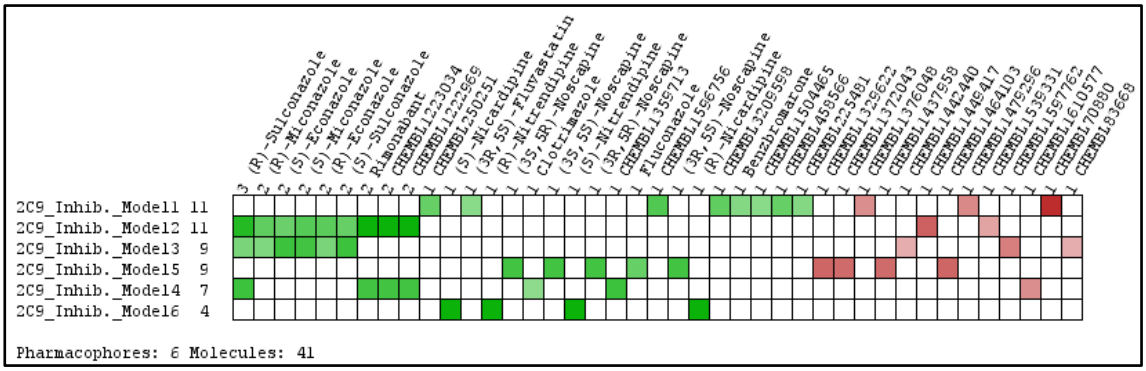


Figure 24. Activity profile of the final six ligand-based models (CYP2C9_Inhib._Model1-6) in training in a parallel screen (true positives are shown in green, false positives in red).

Since with these ligand-based pharmacophore models two thirds of the active training compounds had already been identified, it was not necessary to generate further pharmacophore models at this point. Rather, it was time to test the various ligand-based pharmacophore models and investigate their applicability.

4.3 Model testing

After the training phase was finished, the final part of the work took place, namely the testing of the different 3D-pharmacophore models, which should define the field of application of the models. As already mentioned, however, the structure-based models were not used here anymore, as they were far too poorly working at the beginning. As a result, only the six ligand-based pharmacophore models are discussed in terms of model testing. Since the six refined ligand-based models (***CYP2C9_Inhib._Model1-6***) already complemented each other very well in the training phase in the form of a parallel screen, the models should also be used in combination in the test phase. Figure 25 again shows the six finalized ligand-based 3D-pharmacophore models.

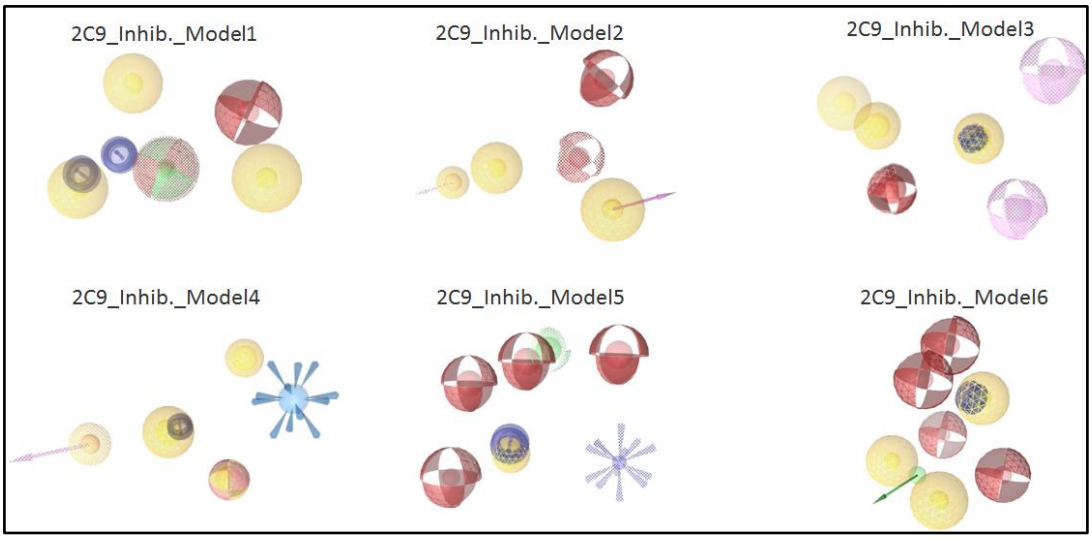


Figure 25. The six finalized ligand-based models *CYP2C9_Inhib._Model1-6.*

As already mentioned, several test sets were available: A very strong (850 investigational compounds), a strong (25 known drugs, 2060 investigational compounds), a moderate (8 known drugs, 1490 investigational compounds) and a weak inhibitor test set (20 known drugs, 2679 investigational compounds). As in the training phase, there was also a set of 3000 inactive compounds for model testing, against which the models were tested to determine the false positive rate. Table 39-41 show which known drugs were present in the individual test sets.

Table 39. Known drugs classified as strong CYP 2C9 inhibitors.

Dyclonine	Nimesulide
Escitalopram	Nimodipine
Felodipine	Nordihydroguaiaretic acid
Finasteride	Oxiconazole
Glyburide	Podophyllotoxin
Indinavir	Raloxifene
Isradipine	Sulphadiazine
Lacidipine	Telmisartan
Lifitegrast	Tribromsalan
Losartan	Troglitazone
Niclosamide	Voriconazole
Nicorandil	Zafirlukast
Nifedipine	

Clustering those 25 strong inhibitor drugs in LigandScout predominantly resulted in single clusters, which in turn shows the strong differences in the existing interaction features of the drugs (out of 17 resulting clusters, 14 were single clusters). Once again, this fact shows how unspecific the binding site of the enzyme CYP 2C9 is. In the following testing of the six 3D-pharmacophore models, the fact that the drug´s pharmacophore features in three-dimensional space are very different presents a considerable challenge in retrieving those strong inhibitor drugs.

Table 40. Known drugs of the moderate CYP 2C9 inhibitor test set.

Amiodarone	Levoketoconazole
Clonidine	Phenytoin
Fenofibrate	Ritonavir
Fluvoxamine	Thyroxine

The clustering process of the eight moderate inhibitor test set drugs in LigandScout resulted in eight single clusters. This means that these drugs have little to no similarity with respect to their pharmacophore features in three-dimensional space.

Table 41. Known drugs of the weak inhibitor test set.

Abiraterone	Omeprazole
Bepridil	Paroxetine
Cilostazol	Phenylbutazone
Dexrabeprazole	Propranolol
Duloxetine	Quinidine
Ivacaftor	Sertraline
Lansoprazole	Simvastatin
Linezolid	Tamoxifen
Loratadine	Ticlopidine
Megestrol acetate	Verapamil

The attempt to cluster the twenty weak inhibitor test sets resulted in the formation of 17 clusters, of which 16 were single clusters. Again, this shows how strongly the individual drugs differ with respect to their three-dimensional pharmacophore features.

4.3.1 **Very strong inhibitor test results**

The parallel screen of the six ligand-based 3D pharmacophore models (***CYP2C9_Inhib._Model1-6***) against the 850 very strong inhibitors (all investigational compounds) and 3000 inactive molecules led to the following result: Overall, the models together were able to detect 32 of the 850 very strong inhibitor test set compounds, which represents 3.76%. Remarkably, there was almost no overlap between the six models (only CHEMBL1446716, CHEMBL1375127 and CHEMBL1232474 were recognized by more than one pharmacophore model), confirming the idea of the parallel screen. Regarding the inactive compounds classified as active (false positives), the models together led to only 16 hits from 3000 compounds, which corresponds to 0.53%. Table 42 exactly breaks down the hit rate of the individual models and in Figure 26 the activity profile shows how strongly a pharmacophore model matches the test set compounds based on the intensity of the color, whereby green color stands for an active hit (true positive) and red color for an inactive hit (false positive). Molecules which are not recorded as a hit are not visible in this figure.

Table 42. Detailed results of the very strong inhibitor test set.

	Very Strong Inhibitor hits (true positives)	Inactive hits (false positives)
CYP2C9_Inhib._Model1	8 (0.94%)	2 (0.07%)
CYP2C9_Inhib._Model2	10 (1.18%)	1 (0.03%)
CYP2C9_Inhib._Model3	8 (0.94%)	4 (0.13%)
CYP2C9_Inhib._Model4	4 (0.47%)	2 (0.07%)
CYP2C9_Inhib._Model5	2 (0.24%)	7 (0.23%)
CYP2C9_Inhib._Model6	4 (0.47%)	0 (0.00%)
Parallel results	**32 (3.76%)**	**16 (0.53%)**

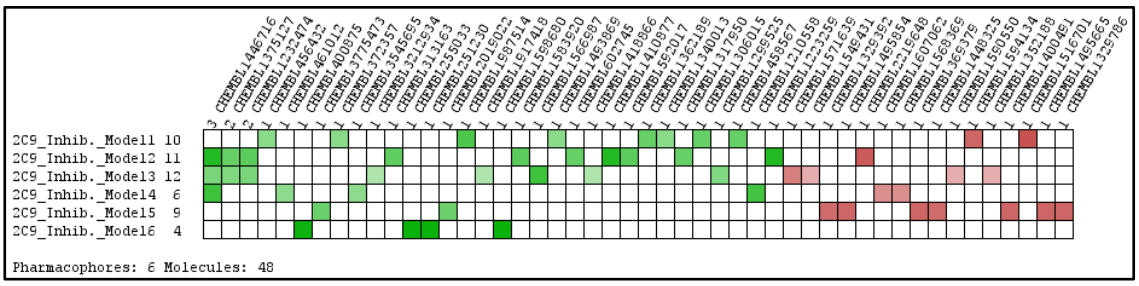


Figure 26. Activity profile of the six ligand-based models against the very strong inhibitor test set.

The hit list of this parallel test screen is available in the appendix and provides the detailed screening results (name, matching features, pharmacophore-fit score, active/decoy, bioactivity type, activity values, activity units, compound ID und assay ID).

4.3.2 **Strong Inhibitor test results**

As already mentioned, the next test set consisted of 2085 active compounds (of which 25 were known drugs and 2060 investigational compounds) and 3000 inactive molecules. This time the six ligand-based pharmacophore models were able to identify 52 active compounds from this test set, which corresponds to a hit rate of 2.49%. Regarding the false positives, there were of course 16 hits again, since the same 3000 inactive compounds were always used for the different test sets. Except for the active compound CHEMBL285932, which was recognized by both ***CYP2C9_Inhib._Model2*** and ***CYP2C9_Inhib._Model4***, once again there was no overlap of the pharmacophore models with respect to hits; once again the meaningfulness of the parallel screen was

confirmed. Table 43 shows the exact contribution of the six pharmacophore models to the respective hits and the activity profile of Figure 27 illustrates how well the pharmacophore models fit the respective hits.

Table 43: Detailed results of the strong inhibitor test set.

	Strong Inhibitor hits (true positives)	Inactive hits (false positives)
CYP2C9_Inhib._Model1	13 (0.62%)	2 (0.07%)
CYP2C9_Inhib._Model2	7 (0.34%)	1 (0.03%)
CYP2C9_Inhib._Model3	12 (0.58%)	4 (0.13%)
CYP2C9_Inhib._Model4	7 (0.34%)	2 (0.07%)
CYP2C9_Inhib._Model5	10 (0.48%)	7 (0.23%)
CYP2C9_Inhib._Model6	4 (0.19%)	0 (0.00%)
Parallel results	**52 (2.49%)**	**16 (0.53%)**

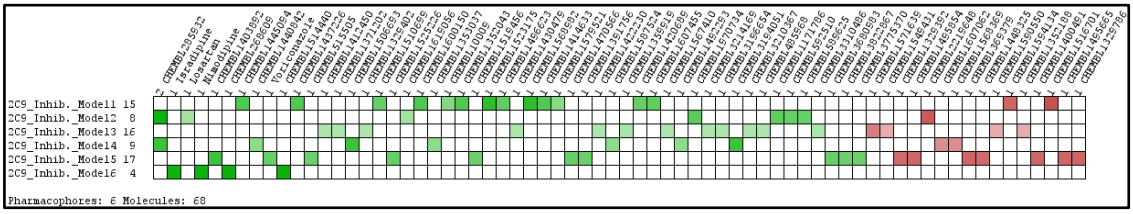


Figure 27. Activity profile of the six ligand-based models against the strong inhibitor test set.

In the appendix of this thesis the hit list with detailed information about the strong inhibitor test screen is available.

4.3.3 **Moderate Inhibitor test results**

For the next test run, the six ligand-based pharmacophore models were screened in parallel against the moderate inhibitor test set, which contained 8 known drugs and 1490 investigational compounds, and the inactive set (3000 inactive compounds). This test screen led to the following result: 23 compounds of 1498 active molecules were detected by the pharmacophore models, corresponding to a percentage of 1.54%. Of 3000 inactive compounds, 16 were of course again classified as active, corresponding to a rate of 0.53%. Except for CHEMBL476833, CHEMBL368348 and CHEMBL173038,

which were recognized by several pharmacophore models, each hit was recognized by a single model, as shown in the activity profile in Figure 28. Table 44 again gives information about the involvement of the individual models ***CYP2C9_Inhib._Model1-6*** in the overall result and the detailed hitlist can be seen in the appendix.

Table 44. Detailed results of the moderate inhibitor test set.

	Moderate Inhibitor hits (true positives)	**Inactive hits (false positives)**
CYP2C9_Inhib._Model1	5 (0.33%)	2 (0.07%)
CYP2C9_Inhib._Model2	8 (0.53%)	1 (0.03%)
CYP2C9_Inhib._Model3	6 (0.40%)	4 (0.13%)
CYP2C9_Inhib._Model4	2 (0.13%)	2 (0.07%)
CYP2C9_Inhib._Model5	5 (0.33%)	7 (0.23%)
CYP2C9_Inhib._Model6	0 (0.00%)	0 (0.00%)
Parallel results	**23 (1.54%)**	**16 (0.53%)**

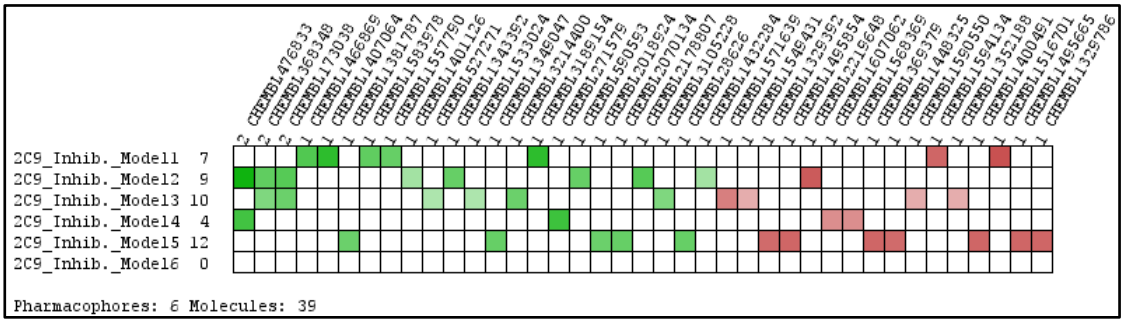


Figure 28. Activity profile of the six ligand-based models against the moderate inhibitor test set.

4.3.4 **Weak inhibitor test results**

The fourth and final test set, against which the six ligand-based pharmacophore models were screened in parallel, consisted of 2699 active compounds (20 known drugs and 2679 investigational compounds) and 3000 inactive molecules. This last test screen resulted in 46 true positives, corresponding to 1.70% of the active compounds, and 16 false positives (0.53%). With the exception of CHEMBL270902 and CHEMBL1096182, each of the hit molecules was discovered by a single pharmacophore model. This therefore also justified the use of the six ligand-based pharmacophore models in a parallel screen in this test set, which contained the compounds with the lowest inhibitory activities. Table 45 shows how much the individual models

contributed to the screening result and Figure 29 illustrates, in the form of an activity profile, how strongly the pharmacophore models matched the corresponding hits.

Table 45: Detailed results of the weak inhibitor test set.

	Active hits (true positives)	**Inactive hits (false positives)**
CYP2C9_Inhib._Model1	6 (0.22%)	2 (0.07%)
CYP2C9_Inhib._Model2	14 (0.52%)	1 (0.03%)
CYP2C9_Inhib._Model3	17 (0.63%)	4 (0.13%)
CYP2C9_Inhib._Model4	6 (0.22%)	2 (0.07%)
CYP2C9_Inhib._Model5	5 (0.19%)	7 (0.23%)
CYP2C9_Inhib._Model6	0 (0.00%)	0 (0.00%)
Parallel results	**46 (1.70%)**	**16 (0.53%)**

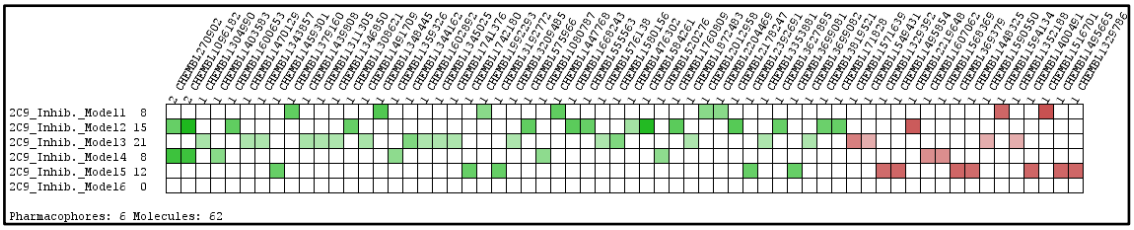


Figure 29. Activity profile of the six ligand-based models against the weak inhibitor test set.

Again, the exact screening results of this test run are attached in the form of a hit list in the appendix.

4.4 Discussion

The testing of the six ligand-based 3D-pharmacophore models was of course the groundbreaking step, as this last part determined the model´s advantages, disadvantages and their applicability domain. No doubt all models are extremely selective in their performance, which is demonstrated by a hit rate in the smallest percentage. The best result was achieved with the very strong inhibitor test set, whereby the six models together classified 3.76% of all active compounds as very strong inhibitors. The three models ***CYP2C9_Inhib._Model1-3*** were of particular importance, as they retrieved a large number of the very strong inhibitors in the test set, but also in the other test sets (Figure 30). Upon detection of strong, moderate and weak inhibitors of CYP 2C9 the three models resulted in significantly fewer hits. The two models ***CYP2C9_Inhib._Model4*** and ***CYP2C9_Inhib._Model6*** generally had a slightly weaker performance than the other three models mentioned. Again, they scored the most hits with the very strong inhibitor test set and showed weaker performance in detecting strong, moderate and weak inhibitors.

The plan not to consider the six ligand-based models as individual and independent units, but to see them as a complementary combination should prove to be a good decision, as the models complemented each other extraordinarily well in almost all test screens and there was hardly any overlap of hits. The only one of the six ligand-based pharmacophore models that was significantly worse than the others in the final test phase was ***CYP2C9_Inhib._Model5***, which in some cases achieved more false positives than true positives. Therefore, the question arises whether the use of this model is really justified with regard to sensitivity and specificity, or whether ***CYP2C9_Inhib._Model5*** should be avoided. Since this one model makes up almost half of all false positives (7 out of 16 false positives), but provides the smallest contribution to the true positives of the very strong inhibitor test set, it is only advisable to avoid ***CYP2C9_Inhib._Model5***. This decision probably increases the quality of the results when the pharmacophore models are applied in practice.

In summary, the five ligand-based 3D-pharmacophore models ***CYP2C9_Inhib._Model1,2,3,4,6*** represent very selective but at the same time high-quality and efficient models capable of detecting very strong competitive inhibitors of CYP450 2C9. In the future, these models can be used to predict competitive inhibitors of CYP2C9 during early drug discovery.

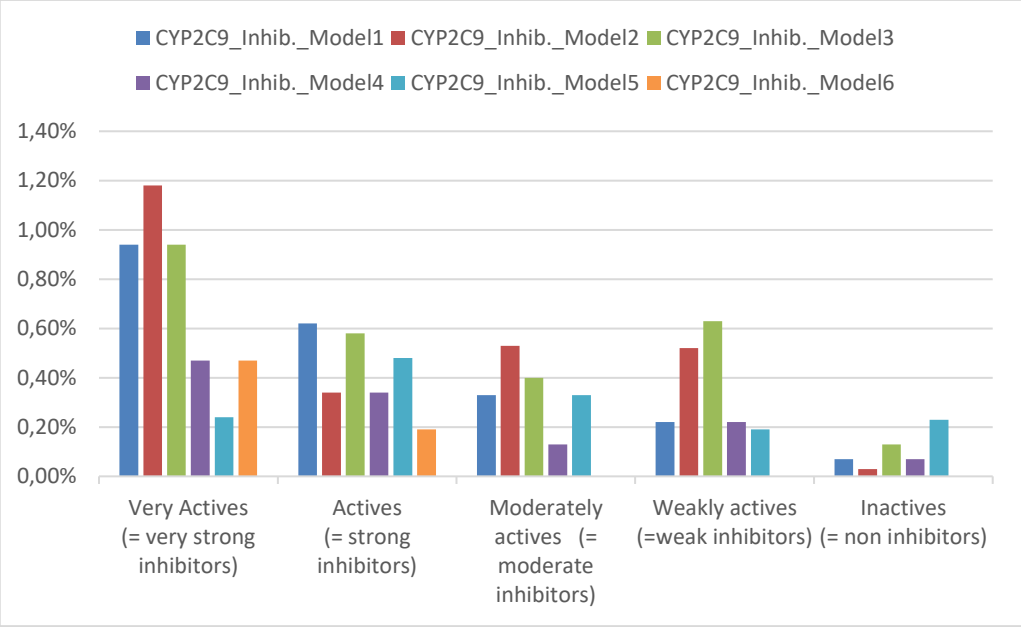


Figure 30. Performance of the six ligand-based models in the different test screens.

Comparing the three best performing ligand-based 3D-pharmacophore models ***CYP2C9_Inhib._Model1-3*** with the previously discussed structure-based 3D-pharmacophore models of 5A5I, 5A5J and 4NZ2, some differences become apparent. On the one hand, the three structure-based models are mainly dominated by hydrophobic features. Quite characteristic of course are the negative ionizable features of the structure-based models 5A5I and 5A5J. In addition, there are a few H-bond acceptors and even fewer H-bond donors. In contrast, the three ligand-based 3D-pharmacophore models are clearly "more colorful". In addition to hydrophobic pharmacophore features and H-bond acceptors, the considerable number of aromatic features and the presence of some halogen-bond donors are remarkable. Negative ionizable features are not found in any of the ligand-based models. This prominent

pharmacophore feature of the structure-based 3D-pharmacophore models was evidently replaced by H-bond acceptors. Overall, the ligand-based 3D-pharmacophore models have, on average, significantly more pharmacophore features than the structure-based pharmacophore models. They seem much more complex and inhomogeneous. It is all the more interesting that the structure-based models in virtual screening achieved almost no hits and were even more selective than the ligand-based models.

Finally, it can of course be argued that detecting 3.76% of all very active compounds is a very low yield. However, there are a few important points to keep in mind: Firstly, the low rate of true positives is followed by a much lower number of false positives, i.e. inactive compounds that were wrongly classified as active. With 0.53% the percentage of false positives is extremely small and this fact shows the high quality of the models. If one takes a closer look at the hit lists in the appendix of this thesis, one can see that the true positives are ranked very high in terms of pharmacophore-fit score, compared to the false positives. Furthermore, one must be aware that cytochrome P450 enzymes such as CYP3A4, CYP2D6 and CYP2C9 are highly unspecific enzymes, which was already discussed at the beginning of this thesis. This non-specificity of enzymes is of major importance for our body and enables the metabolism of countless substances. But it is also exactly this fact that makes the application of pharmacophore modeling with such a target so difficult and time-consuming. If one decides to design pharmacophore models rather unselective and fuzzy, of course one will achieve a high rate of true positives, unfortunately also a high rate of false positives. But if one wants to design high quality models with great specificity,low sensitivity must be considered, which is the case here. However, the aim of this work was to generate such high quality and highly selective pharmacophore models from the beginning. The goal of identifying competitive inhibitors of CYP450 2C9 and distinguishing them from non-inhibitors can thus be considered achieved and in the future those ligand-based 3D-pharmacophore models can be used to predict competitive inhibitors of CYP450 2C9 at an early stage and guide medicinal chemists to modify compounds to avoid undesired outcomes.

5 Abbreviations

ACAT	Acyl-CoA:cholesterol acyltransferase
AhR	Aryl hydrocarbon receptor
AR	Aromatic ring
AUC	Area under the curve
CAR	Constitutive androstane receptor
CPR	NADPH-cytochrome P450 reductase
CPU	Central processing unit
CYP450	Cytochrome P450
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
EF	Enrichment factor
EMBL	European Molecular Biology Laboratory
FAD	Flavin adenine dinucleotide
FEB	Iron binding location
FMN	Flavin mononucleotide
GST	Glutathione S-transferase
H	Hydrophobic interaction
HBA	Hydrogen-bond acceptor
HBD	Hydrogen-bond donor
H-bond	Hydrogen-bond
HIV	Human immunodeficiency virus
HTS	High throughput screening
IC50	Half maximal inhibitory concentration
ID	Identifier
IUPAC	International Union of Pure and Applied Chemistry
Ki	Inhibitory constant
KNIME	Konstanz Information Miner
LDB	LigandScout Data Base
MD simulation	Molecular dynamics simulation
MGB	Magnesium binding location
MNB	Manganese binding location
MOE	Molecular Operating Environment
mRNA	Messenger ribonucleic acid
NADPH	Nicotinamide adenine dinucleotide phosphate
NAT	N-acetyltransferase
NI	Negative ionizable area
NMR	Nuclear magnetic resonance
NSAID	Non-steroidal anti-inflammatory drugs
PDB	Protein Data Bank
PI	Positive ionizable area
PPI	Proton-pump inhibitor
PXR	Pregnane X receptor

QSAR	Quantitative structure-activity relationship
RCSB	Research Collaboratory f. Structural Bioinformatics
RDF	Resource Description Framework
ROC	Receiver operating characteristic
RXR	Retinoid X receptor
SDF	Structure-Data File
SMILES	Simplified Molecular Input Line Entry Specification
SULT	Sulfotransferase
UGT	UDP-glucuronosyltransferase
XBD	Halogen bond donor
ZNB	Zinc binding location

6 References

1. Lynch, T. & Price, A. The Effect of Cytochrome P450 Metabolism on Drug Response, Interactions, and Adverse Effects. *Am. Fam. Physician* **76**, 391–396 (2007).

2. Ogu, C. C. & Maxa, J. L. Drug Interactions Due to Cytochrome P450. *Baylor Univ. Med. Cent. Proc.* **13**, 421–423 (2000).

3. Ahlström, M. M., Ridderström, M. & Zamora, I. CYP2C9 structure-metabolism relationships: Substrates, inhibitors, and metabolites. *J. Med. Chem.* **50**, 5382–5391 (2007).

4. Rettie, A. E. & Jones, J. P. Clinical and Toxicological Relevance of CYP2C9: Drug-Drug Interactions and Pharmacogenetics. *Annu. Rev. Pharmacol. Toxicol.* **45**, 477–494 (2005).

5. Cavallari, L. H., Jeong, H. & Bress, A. Role of cytochrome P450 genotype in the steps toward personalized drug therapy. *Pharmgenomics. Pers. Med.* **4**, 123–136 (2011).

6. Zhou, S.-F., Zhou, Z.-W., Yang, L.-P. & Cai, J.-P. Substrates, Inducers, Inhibitors and Structure-Activity Relationships of Human Cytochrome P450 2C9 and Implications in Drug Development. *Curr. Med. Chem.* **16**, 3480–3675 (2009).

7. Mestres, J., Bryant, S. D., Zamora, I. & Gasteiger, J. Shaping the future of safer innovative drugs in Europe. *Nat. Biotechnol.* **29**, 789–790 (2011).

8. Kumar, G. N. *et al.* Potent Inhibition of the Cytochrome P-450 3A-Mediated Human Liver Microsomal Metabolism of a Novel HIV Protease Inhibitor by Ritonavir : A Positive Drug-Drug Interaction. *Drug Metab. Dispos.* **27**, 902–908 (1999).

9. Parkinson, A. Biotransformation of Xenobiotics. in *Casarett and Doull's Toxicology: The Basic Science of Poisons* (2009).

doi:10.1002/9780470744307.gat004

10. Meyer, U. A. Overview of enzymes of drug metabolism. *J. Pharmacokinet. Biopharm.* **24**, 449–459 (1996).

11. Rice, J. E. Drug Metabolism. in *Organic Chemistry Concepts and Applications for Medicinal Chemistry* 145–201 (Academic Press, 2014). doi:10.1016/B978-0-12-800739-6.00007-3

12. Kitteringham, B., Munir, N. R. & Park, K. The Role of Active Metabolites in Drug Toxicity. *Drug Saf.* **11**, 114–144 (1994).

13. Jancova, P., Anzenbacher, P. & Anzenbacherova, E. Phase II Drug Metabolizing Enzymes. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub.* **154**, 103–116 (2010).

14. Gram, T. E. & Gillette, J. R. Biotransformation of Drugs. in *Fundamentals of Biochemical Pharmacology* 571–609 (Pergamon Press, 1971). doi:10.1016/B978-0-08-017775-5.50028-3

15. Zhou, S., Liu, J. & Chowbay, B. Polymorphism of human cytochrome P450 enzymes and its clinical impact. *Drug Metab. Rev.* **41**, 89–295 (2009).

16. Denisov, I. G., Makris, T. M., Sligar, S. G. & Schlichting, I. Structure and Chemistry of Cytochrome P450. *Chem. Rev.* **105**, 2253–2277 (2005).

17. Furge, L. L. & Guengerich, F. P. Cytochrome P450 Enzymes in Drug Metabolism and Chemical Toxicology. *Biochem. Mol. Biol. Educ.* **34**, 66–74 (2006).

18. Nebert, D. W. *et al.* The P450 Gene Superfamily : Recommended Nomenclature. *DNA* **6**, 1–11 (1987).

19. R.L., S. & D.J., E. Recent advances: The cytochrome P450 enzymes. *Ann. Pharmacother.* **29**, 619–624 (1995).

20. Pelkonen, O. *et al.* Inhibition and induction of human cytochrome P450 enzymes : current status. *Arch. Toxicol.* **82**, 667–715 (2008).

21. Anzenbacher, P. & Anzenbacherová, E. Cytochromes P450 and metabolism of xenobiotics. *Cell. Mol. Life Sci.* **58**, 737–747 (2001).

22. Williams, P. A., Cosme, J., Sridhar, V., Johnson, E. F. & McRee, D. E. Mammalian microsomal cytochrome P450 monooxygenase: Structural adaptations for membrane binding and functional diversity. *Mol. Cell* **5**, 121–131 (2000).

23. Gay, S. C., Roberts, A. G. & Halpert, J. R. Structural features of cytochromes P450 and ligands that affect drug metabolism as revealed by X-ray crystallography and NMR. *Future Med. Chem.* **2**, 1451–1468 (2010).

24. Denisov, I. G., Makris, T. M., Sligar, S. G. & Schlichting, I. Structure and Chemistry of Cytochrome P450. *Chem. Rev.* **105**, 2253–2277 (2005).

25. Meunier, B., de Visser, S. P. & Shaik, S. Mechanism of oxidation reactions catalyzed by cytochrome P450 enzymes. *Chem. Rev.* **104**, 3947–3980 (2004).

26. Werck-Reichhart, D. & Feyereisen, R. Cytochromes P450: a success story. *Genome Biol.* **1**, 1–9 (2000).

27. Montellano, P. R. O. & De Voss, J. J. Substrate Oxidation by Cytochrome P450 Enzymes. in *Cytochrome P450: Structure, Mechanism, and Biochemistry* (ed. Montellano, P. R. O.) 183–245 (Springer, 2005).

28. Farooq, Y. & Roberts, G. C. K. Kinetics of electron transfer between NADPH-cytochrome P450 reductase and cytochrome P450 3A4. *Biochem. J.* **432**, 485–494 (2010).

29. Wang, M. *et al.* Three-dimensional structure of NADPH-cytochrome P450 reductase: Prototype for FMN- and FAD-containing enzymes. *Proc. Natl. Acad. Sci.* **94**, 8411–8416 (1997).

30. Pandey, A. V. & Flück, C. E. NADPH P450 oxidoreductase: Structure, function, and pathology of diseases. *Pharmacol. Ther.* **138**, 229–254 (2013).

31. Pelkonen, O. *et al.* Inhibition and induction of human cytochrome P450 enzymes: current status. *Arch. Toxicol.* **82**, 667–715 (2008).

32. Hollenberg, P. F. Characteristics and common properties of inhibitors, inducers, and activators of cyp enzymes. *Drug Metab. Rev.* **34**, 17–35 (2002).

33. Zhang, Z. Y. & Wong, Y. N. Enzyme kinetics for clinically relevant CYP inhibition. *Curr. Drug Metab.* **6**, 241–257 (2005).

34. Pelkonen, O., Mäenpääj, J., Taavitsainen, P., Rautio, A. & Raunio, H. Inhibition and induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica* **28**, 1203–1253 (1998).

35. Tompkins, L. M. & Wallace, A. D. Mechanisms of cytochrome P450 induction. *J. Biochem. Mol. Toxicol.* **21**, 176–181 (2007).

36. Denison, M. S. & Whitlock, J. P. Xenobiotic-inducible Transcription of Cytochrome P450 Genes. *J. Biol. Chem.* **270**, 18175–18178 (1995).

37. Ridderstro, M. *et al.* Arginines 97 and 108 in CYP2C9 Are Important Determinants of the Catalytic Function. *Biochem. Biophys. Res. Commun.* **270**, 983–987 (2000).

38. Williams, P. A. *et al.* Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature* **424**, 464 (2003).

39. Wester, M. R. *et al.* The structure of human cytochrome P450 2C9 complexed with flurbiprofen at 2.0 Å resolution. *J. Biol. Chem.* **279**, 35630–35637 (2004).

40. Yamazaki, H. *et al.* Comparative Studies on the Catalytic Roles of Cytochrome P450 2C9 and Its Cys-and Leu-Variants in the Oxidation of Warfarin, Flurbiprofen, and Diclofenac by Human Liver Microsomes. *Biochem. Pharmacol.* **56**, 243–251 (1998).

41. Giancarlo, G. M., Venkatakrishnan, K., Granda, B. W., von Moltke, L. L. & Greenblatt, D. J. Relative contributions of CYP2C9 and 2C19 to phenytoin 4-hydroxylation in vitro: inhibition by sulfaphenazole, omeprazole, and ticlopidine. *Eur. J. Clin. Pharmacol.* **57**, 31–36 (2001).

42. Lee, C. B. *et al.* Tolbutamide, flurbiprofen, and losartan as probes of CYP2C9

activity in humans. *J. Clin. Pharmacol.* **43**, 84–91 (2003).

43. Berka, K., Hendrychová, T., Anzenbacher, P. & Otyepka, M. Membrane position of ibuprofen agrees with suggested access path entrance to cytochrome P450 2C9 active site. *J. Phys. Chem. A* **115**, 11248–11255 (2011).

44. Ohyama, K. *et al.* Inhibitory effects of amiodarone and its N-deethylated metabolite on human cytochrome P450 activities: Prediction of in vivo drug interactions. *Br. J. Clin. Pharmacol.* **49**, 244–253 (2000).

45. Taavitsainen, P., Kiukaanniemi, K. & Pelkonen, O. In vitro inhibition screening of human hepatic P450 enzymes by five angiotensin-II receptor antagonists. *Eur J Clin Pharmacol* **56**, 135–140 (2000).

46. Voorman, R. L., Payne, N. A., Wienkers, L. C., Hauer, M. J. & Sanders, P. E. Interaction of delavirdine with human liver microsomal cytochrome P450: inhibition of CYP2C9, CYP2C19, and CYP2D6. *Drug Metab. Dispos.* **29**, 41–47 (2001).

47. Takahashi, H. *et al.* Potentiation of anticoagulant effect of warfarin caused by enantioselective metabolic inhibition by the uricosuric agent benzbromarone. *Clin. Pharmacol. Ther.* **66**, 569–581 (1999).

48. Wen, X., Wang, J. S., Backman, J. T., Kivisto, K. T. & Neuvonen, P. J. Gemfibrozil is a potent inhibitor of human cytochrome P450 2C9. *Drug Metab. Dispos.* **29**, 1359–1361 (2001).

49. Miners, J. O. *et al.* Tolbutamide hydroxylation by human liver microsomes. Kinetic characterisation and relationship to other cytochrome P-450 dependent xenobiotic oxidations. *Biochem. Pharmacol.* **37**, 1137–1144 (1988).

50. Kim, K. A. & Park, J. Y. Inhibitory effect of glyburide on human cytochrome P450 isoforms in human liver microsomes. *Drug Metab. Dispos.* **31**, 1090–1092 (2003).

51. Li, X. Q., Andersson, T. B., Ahlstrom, M. & Weidolf, L. Comparison of inhibitory

effects of the proton pump-inhibiting drugs omeprazole, esomeprazole, lansoprazole, pantoprazole, and rabeprazole on human cytochrome P450 activities. *Drug Metab. Dispos.* **32**, 821–827 (2004).

52. Back, D. J., Tjia, J. F., Karbwang, J. & Colbert, J. In vitro inhibition studies of tolbutamide hydroxylase activity of human liver microsomes by azoles, sulphonamides and quinolines. *Br. J. Clin. Pharmacol.* **26**, 23–29 (1988).

53. Transon, C., Leemann, T. & Dayer, P. In Vitro Comparative Inhibition Profiles of Major Human Drug Metabolising Cytochrome P450 Isozymes (CYP2C9, CYP2D6 and CYP3A4) by HMG-CoA Reductase Inhibitors. *Eur. J. Clin. Pharmacol.* **50**, 209–215 (1996).

54. Mancy, A., Dijols, S., Poli, S., Guengerich, P. & Mansuy, D. Interaction of Sulfaphenazole Derivatives with Human Liver Cytochromes P450 2C: Molecular Origin of the Specific Inhibitory Effects of Sulfaphenazole on CYP 2C9 and Consequences for the Substrate Binding Site Topology of CYP 2C9. *Biochemistry* **35**, 16205–16212 (1996).

55. Komatsu, K. *et al.* Prediction of in Vivo Drug-Drug Interactions Between Tolbutamide and Various Sulfonamides in Humans Based on in Vitro Experiments. *Drug Metab. Dispos.* **28**, 475–481 (2000).

56. Kumar, V. *et al.* CYP2C9 Inhibition: Impact of Probe Selection and Pharmacogenetics on in Vitro Inhibition Profiles. *Drug Metab. Dispos.* **34**, 1966–1975 (2006).

57. Volak, L. P., Ghirmai, S., Cashman, J. R. & Court, M. H. Curcuminoids Inhibit Multiple Human Cytochromes P450, UDP-glucuronosyltransferase, and Sulfotransferase Enzymes, whereas Piperine is a Relatively Selective CYP3A4 Inhibitor. *Drug Metab. Dispos.* **36**, 1594–1605 (2008).

58. Kim, J. Y. *et al.* Effects of Flavonoids Isolated from Scutellariae Radix on Cytochrome P450 Activities in Human Liver Microsomes. *J. Toxicol. Environ. Heal. A.* **65**, 373–381 (2002).

59. Foti, R. S., Wahlstrom, J. L. & Wienkers, L. C. The In Vitro Drug Interaction Potential of Dietary Supplements Containing Multiple Herbal Components. *Drug Metab. Dispos.* **35**, 185–188 (2007).

60. von Moltke, L. L. *et al.* Inhibition of Human Cytochromes P450 by Components of Ginkgo Biloba. *J. Pharm. Pharmacol.* **56**, 1039–1044 (2004).

61. Foster, B. C. *et al.* An In Vitro Evaluation of Human Cytochrome P450 3A4 and P-glycoprotein Inhibition by Garlic. *J. Pharm. Pharm. Sci.* **4**, 176–184 (2001).

62. Gaudineau, C., Beckerman, R., Welbourn, S. & Auclair, K. Inhibition of Human P450 Enzymes by Multiple Constituents of the Ginkgo Biloba Extract. *Biochem. Biophys. Res. Commun.* **318**, 1072–1078 (2004).

63. Fontana, E., Dansette, P. M. & Poli, S. M. Cytochrome P450 Enzymes Mechanism Based Inhibitors: Common Sub-Structures and Reactivity. *Curr. Drug Metab.* **6**, 413–451 (2005).

64. Chen, Y.; Ferguson, S.S.; Negishi, M.; Goldstein, J. A. Induction of Human CYP2C9 by Rifampicin, Hyperforin, and Phenobarbital is Mediated by the Pregnane X Receptor. *J. Pharmacol. Exp. Ther.* **308**, 495–501 (2004).

65. Dingemanse, J., Schaarschmidt, D. & van Giersbergen, P. L. M. Investigation of the Mutual Pharmacokinetic Interactions Between Bosentan, a Dual Endothelin Receptor Antagonist, and Simvastatin. *Clin. Pharmacokinet.* **42**, 293–301 (2003).

66. Robertson, D. G., Breider, M. A. & Milad, M. A. Preclinical Safety Evaluation of Avasimibe in Beagle Dogs: An ACAT Inhibitor with Minimal Adrenal Effects. *Toxicol. Sci.* **59**, 324–334 (2001).

67. Sahi, J. *et al.* Effects Of Avasimibe on Cytochrome P450 2C9 Expression in Vitro and in Vivo. *Drug Metab. Dispos.* **32**, 1370–1376 (2004).

68. Dixit, V. *et al.* Cytochrome P450 Enzymes and Transporters Induced by Anti-Human Immunodeficiency Virus Protease Inhibitors in Human Hepatocytes: Implications for Predicting Clinical Drug Interactions. *Drug Metab. Dispos.* **35**,

1853–1859 (2007).

69. Chang, T.K.; Yu, L.; Maurel, P.; Waxman, D. J. Enhanced Cyclophosphamide and Ifosfamide Activation in Primary Human Hepatocyte Cultures: Response to Cytochrome P-450 Inducers and Autoinduction by Oxazaphosphorines. *Cancer Res.* **57**, 1946–1954 (1997).

70. Drie, J. H. Van. Monty Kier and the Origin of the Pharmacophore Concept. *Internet Electron. J. Mol. Des.* **6**, 271–279 (2007).

71. Güner, O. F. & Bowen, J. P. Setting the Record Straight: The Origin of the Pharmacophore Concept. *J. Chem. Inf. Model.* **54**, 1269–1283 (2014).

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

73. Wolber, G. & Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **45**, 160–169 (2005).

74. Langer, T. & Bryant, S. D. In Silico Screening: Hit Finding from Database Mining. in *Wermuth's The Practice of Medicinal Chemistry* 210–222 (Elsevier Ltd, 2008).

75. Langer, T. Pharmacophores in Drug Research. *Mol. Inform.* **29**, 470–475 (2010).

76. Karaboga, A. S. *et al.* Highly Specific and Sensitive Pharmacophore Model for Identifying CXCR4 Antagonists. Comparison with Docking and Shape-Matching Virtual Screening Performance. *J. Chem. Inf. Model.* **53**, 1043–1056 (2013).

77. Duwensee, K. *et al.* Leoligin , the major lignan from Edelweiss , activates cholesteryl ester transfer protein. *Atherosclerosis* **219**, 109–115 (2011).

78. Wei, Y. *et al.* New Use for an Old Drug : Inhibiting ABCG2 with Sorafenib. *Mol. Cancer Ther.* **11**, 1693–1702 (2012).

79. Rechfeld, F. *et al.* Thienoquinolines as Novel Disruptors of the PKC ε /RACK2 Protein − Protein Interaction. *J. Med. Chem.* **57**, 3235–3246 (2014).

80. Bryant, S. & Langer, T. Data Mining Using Ligand Profiling and Target Fishing. in *Data Mining in Drug Discovery* (eds. Hoffmann, R. D., Gohier, A. & Pospisil, P.) 257–269 (Wiley-VCH Verlag GmbH & Co. KGaA, 2014).

81. Seidel, T., Ibis, G., Bendix, F. & Wolber, G. Strategies for 3D pharmacophore-based virtual screening. *Drug Discov. Today Technol.* **7**, 221–228 (2010).

82. Langer, T. & Wolber, G. Pharmacophore definition and 3D searches. *Drug Discov. Today* **1**, 203–207 (2004).

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

84. Drwal, M. & Griffith, R. Combination of ligand- and structure-based methods in virtual screening. *Drug Discov. Today Technol.* **10**, 395–401 (2013).

85. Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov. Today* **15**, 444–450 (2010).

86. Caporuscio, F. & Tafi, A. Pharmacophore Modelling: A Forty Year Old Approach and its Modern Synergies. *Curr. Med. Chem.* **18**, 2543–2553 (2011).

87. Chang, C., Ekins, S., Bahadduri, P. & Swaan, P. W. Pharmacophore-based discovery of ligands for drug transporters. *Adv. Drug Deliv. Rev.* **58**, 1431–1450 (2006).

88. Poli, G., Seidel, T. & Langer, T. Conformational Sampling of Small Molecules With iCon : Performance Assessment in Comparison With OMEGA. *Front. Chem.* **6**, (2018).

89. Sun, H. Pharmacophore-based virtual screening. *Curr. Med. Chem.* **15**, 1018–1024 (2008).

90. Mayr, L. M. & Bojanic, D. Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* **9**, 580–588 (2009).

91. 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).

92. Kirchmair, J. *et al.* Fast and Efficient in Silico 3D Screening: Toward Maximum Computational Efficiency of Pharmacophore-Based and Shape-Based Approaches. *J. Chem. Inf. Model.* **47**, 2182–2196 (2007).

93. Markt, P., Schuster, D. & Langer, T. Pharmacophore Models for Virtual Screening. in *Virtual Screening* (ed. Sotriffer, C.) 115–152 (Wiley-VCH Verlag GmbH & Co. KGaA, 2011).

94. Triballeau, N., Acher, F., Brabet, I., Pin, J.-P. & Bertrand, H.-O. Virtual Screening Workflow Development Guided by the 'Receiver Operating Characteristic' Curve Approach . Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **48**, 2534–2547 (2005).

95. Bender, A. Compound bioactivities go public. *Nat. Chem. Biol.* **6**, 309 (2010).

96. Warr, W. A. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput. Aided. Mol. Des.* **23**, 195–198 (2009).

97. Gaulton, A. *et al.* ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, 1100–1107 (2012).

98. Papadatos, G., Gaulton, A., Hersey, A. & Overington, J. P. Activity, assay and target data curation and quality in the ChEMBL database. *J. Comput. Aided. Mol. Des.* **29**, 885–896 (2015).

99. Wang Y, Suzek T, Z. J. et al. PubChem BioAssay: 2014 update. *Nucleic Acids Res.* **42**, 1075–1082 (2013).

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

101. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

102. Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **35**, 301–303 (2007).

103. Berman, H. M. *et al.* The Protein Data Bank. *Acta Cryst.* **D58**, 899–907 (2002).

104. Rose, P. W. *et al.* The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45**, D271–D281 (2017).

105. 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 (2012).

106. Sieb, C., Meinl, T. & Berthold, M. R. Parallel and Distributed Data Pipelining With Knime. *Mediterr. J. Comput. Networks* **3**, 43–51 (2007).

107. Feltrin, L. KNIME an Open Source Solution for Predictive Analytics in the Geosciences. *IEEE Geoscience and Remote Sensing Magazine* **3**, 28–38 (2015).

108. Berthold, M. R. *et al.* KNIME - the Konstanz Information Miner. *ACM SIGKDD Explor. Newsl.* **11**, 26–31 (2009).

109. Spitzer, G. M. *et al.* One Concept, Three Implementations of 3D Pharmacophore-Based Virtual Screening: Distinct Coverage of Chemical Search Space. *J. Chem. Inf. Model.* **50**, 1241–1247 (2010).

110. Mcconathy, J., Ph, D., Owens, M. J. & Ph, D. Stereochemistry in Drug Action. *Prim. Care Companion J Clin Psychiatry* **5**, 70–73 (2003).

111. Skerratt, S. E., De Groot, M. J. & Phillips, C. Discovery of a Novel Binding Pocket for Cyp 2C9 Inhibitors: Crystallography, Pharmacophore Modelling and Inhibitor SAR. *Med.Chem.Comm.* **7**, 813 (2016).

112. Brändén, G., Sjögren, T., Schnecke, V. & Xue, Y. Structure-based ligand design to overcome CYP inhibition in drug discovery projects. *Drug Discov. Today* **19**, 905–911 (2014).

113. Liu, R. *et al.* Determinants of the Inhibition of DprE1 and CYP2C9 by Antitubercular Thiophenes. *Angew. Chemie - Int. Ed.* **56**, 13011–13015 (2017).

114. Maekawa, K. *et al.* Structural Basis of Single Nucleotide Polymorphisms in Cytochrome P450 2C9. *Biochemistry* **56**, 5476–5480 (2017).

7 Appendix

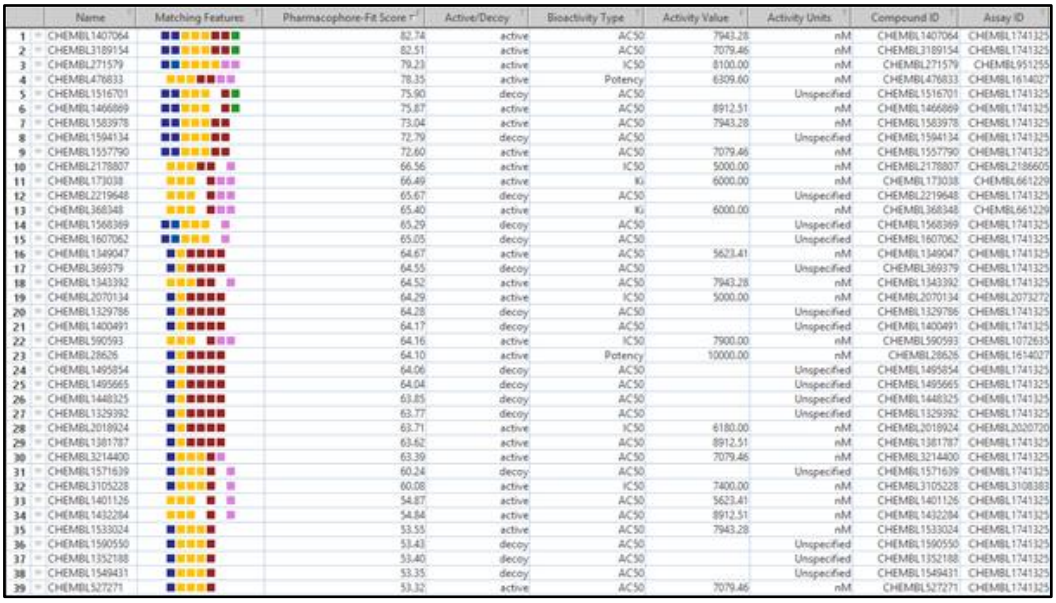
Hit list of the six ligand-based models: very strong CYP 2C9 inhibitor test set (sorted by pharmacophore-fit score).

	Name	Matching Features	Pharmacophore-Fit Score	Active/Decoy	Bioactivity Type	Activity Value	Activity Units	Compound ID	Assay ID
1	CHEMBL255033		98.84	active	Potency	631.00	nM	CHEMBL255033	CHEMBL1614027
2	CHEMBL251230		98.78	active	Potency	316.20	nM	CHEMBL251230	CHEMBL1614027
3	CHEMBL1598680		98.61	active	Potency	199.50	nM	CHEMBL1598680	CHEMBL1614027
4	CHEMBL400875		97.59	active	Potency	794.30	nM	CHEMBL400875	CHEMBL1614027
5	CHEMBL1210558		79.19	active	IC50	100.00	nM	CHEMBL1210558	CHEMBL1228414
6	CHEMBL1987514		76.88	active	AC50	891.25	nM	CHEMBL1987514	CHEMBL1741325
7	CHEMBL1516701		75.90	decoy	AC50		Unspecified	CHEMBL1516701	CHEMBL1741325
8	CHEMBL1446716		75.01	active	Potency	20.00	nM	CHEMBL1446716	CHEMBL1614027
9	CHEMBL1223259		74.93	active	IC50	20.00	nM	CHEMBL1223259	CHEMBL1227229
10	CHEMBL1410877		74.77	active	AC50	316.23	nM	CHEMBL1410877	CHEMBL1741325
11	CHEMBL1594134		72.79	decoy	AC50		Unspecified	CHEMBL1594134	CHEMBL1741325
12	CHEMBL458567		72.75	active	Ki	41.00	nM	CHEMBL458567	CHEMBL972103
13	CHEMBL1306015		72.72	active	AC50	446.68	nM	CHEMBL1306015	CHEMBL1741325
14	CHEMBL1362189		72.49	active	AC50	281.84	nM	CHEMBL1362189	CHEMBL1741325
15	CHEMBL1566987		70.65	active	Potency	398.10	nM	CHEMBL1566987	CHEMBL1614027
16	CHEMBL456432		66.23	active	Ki	4.20	nM	CHEMBL456432	CHEMBL972103
17	CHEMBL1493869		66.13	active	AC50	63.10	nM	CHEMBL1493869	CHEMBL1741325
18	CHEMBL592017		66.07	active	IC50	690.00	nM	CHEMBL592017	CHEMBL1072635
19	CHEMBL1340013		65.92	active	AC50	112.20	nM	CHEMBL1340013	CHEMBL1741325
20	CHEMBL1583920		65.85	active	AC50	446.68	nM	CHEMBL1583920	CHEMBL1741325
21	CHEMBL372357		65.71	active	IC50	840.00	nM	CHEMBL372357	CHEMBL1100451
22	CHEMBL2219648		65.67	decoy	AC50		Unspecified	CHEMBL2219648	CHEMBL1741325
23	CHEMBL1232474		65.63	active	AC50	158.49	nM	CHEMBL1232474	CHEMBL1741325
24	CHEMBL1568369		65.29	decoy	AC50		Unspecified	CHEMBL1568369	CHEMBL1741325
25	CHEMBL313163		65.26	active	AC50	630.96	nM	CHEMBL313163	CHEMBL1741325
26	CHEMBL461012		65.26	active	IC50	800.00	nM	CHEMBL461012	CHEMBL963888
27	CHEMBL3545695		65.24	active	IC50	440.00	nM	CHEMBL3545695	CHEMBL3385973
28	CHEMBL1607062		65.05	decoy	AC50		Unspecified	CHEMBL1607062	CHEMBL1741325
29	CHEMBL369379		64.55	decoy	AC50		Unspecified	CHEMBL369379	CHEMBL1741325
30	CHEMBL602745		64.50	active	IC50	870.00	nM	CHEMBL602745	CHEMBL1072635
31	CHEMBL1375127		64.36	active	Potency	20.00	nM	CHEMBL1375127	CHEMBL1614027
32	CHEMBL1329786		64.28	decoy	AC50		Unspecified	CHEMBL1329786	CHEMBL1741325
33	CHEMBL1400491		64.17	decoy	AC50		Unspecified	CHEMBL1400491	CHEMBL1741325
34	CHEMBL1495854		64.06	decoy	AC50		Unspecified	CHEMBL1495854	CHEMBL1741325
35	CHEMBL1495665		64.04	decoy	AC50		Unspecified	CHEMBL1495665	CHEMBL1741325
36	CHEMBL1317950		64.04	active	Potency	12.60	nM	CHEMBL1317950	CHEMBL1614027
37	CHEMBL3775473		63.96	active	IC50	397.00	nM	CHEMBL3775473	CHEMBL3776779
38	CHEMBL1448325		63.85	decoy	AC50		Unspecified	CHEMBL1448325	CHEMBL1741325
39	CHEMBL1329392		63.77	decoy	AC50		Unspecified	CHEMBL1329392	CHEMBL1741325
40	CHEMBL2019022		63.74	active	IC50	550.00	nM	CHEMBL2019022	CHEMBL2020720
41	CHEMBL1299525		60.27	active	AC50	14.13	nM	CHEMBL1299525	CHEMBL1741325
42	CHEMBL1571639		60.24	decoy	AC50		Unspecified	CHEMBL1571639	CHEMBL1741325
43	CHEMBL1590550		53.43	decoy	AC50		Unspecified	CHEMBL1590550	CHEMBL1741325
44	CHEMBL1352188		53.40	decoy	AC50		Unspecified	CHEMBL1352188	CHEMBL1741325
45	CHEMBL1549431		53.35	decoy	AC50		Unspecified	CHEMBL1549431	CHEMBL1741325
46	CHEMBL1418866		53.20	active	AC50	79.43	nM	CHEMBL1418866	CHEMBL1741325
47	CHEMBL1917418		53.19	active	IC50	800.00	nM	CHEMBL1917418	CHEMBL1919388
48	CHEMBL3212934		53.09	active	AC50	501.19	nM	CHEMBL3212934	CHEMBL1741325

Hit list of the six ligand-based models: strong CYP 2C9 inhibitor test set (sorted by pharmacophore-fit score).

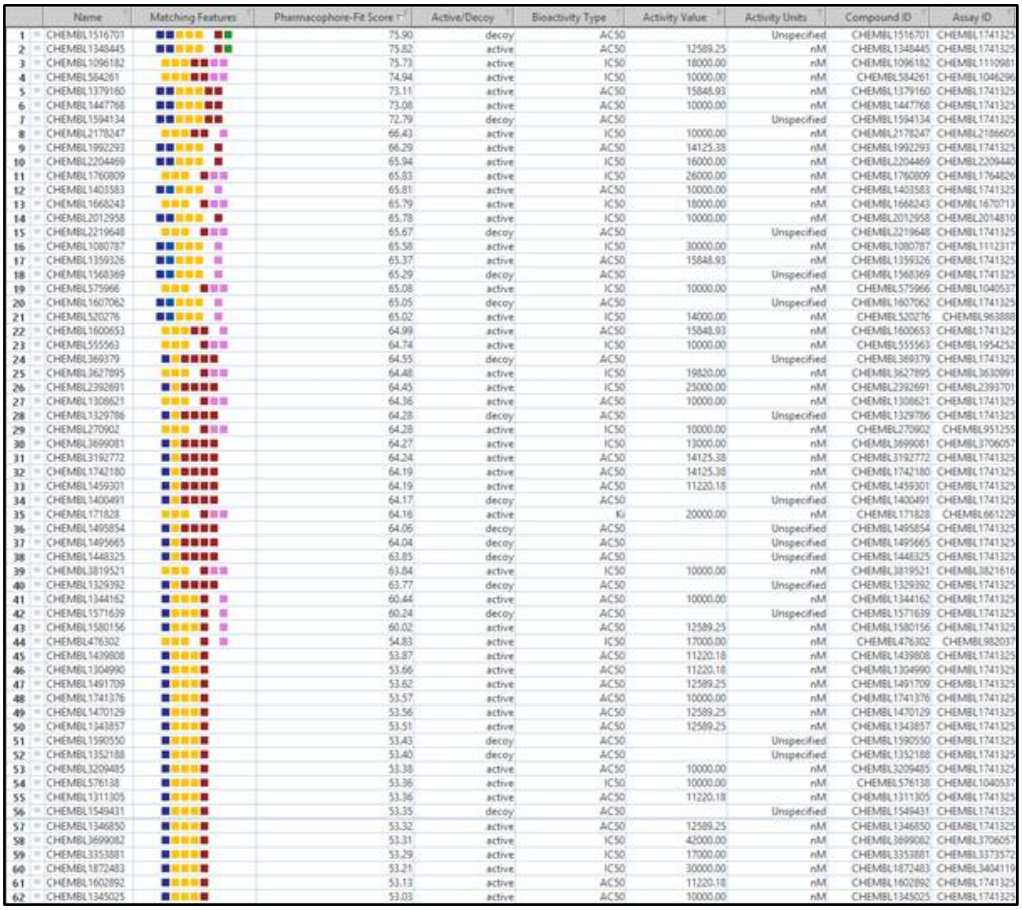
	Name	Matching Features	Pharmacophore-Fit Score	Active/Decoy	Bioactivity Type	Activity Value	Activity Units	Compound ID	Assay ID
1	Isradipine		97.59	active	IC50	3660.00	nM	CHEMBL1648	CHEMBL3530854
2	Voriconazole		64.53	active	IC50	3620.00	nM	CHEMBL638	CHEMBL1635163
3	Nimodipine		98.79	active	IC50	1690.00	nM	CHEMBL1428	CHEMBL3541096
4	Losartan		54.78	active	IC50	2882.00	nM	CHEMBL191	CHEMBL1909135
5	CHEMBL1403892		70.56	active	Potency	1584.90	nM	CHEMBL1403892	CHEMBL1614027
6	CHEMBL268609		98.79	active	Potency	1584.90	nM	CHEMBL268609	CHEMBL1614027
7	CHEMBL1445094		75.69	active	Potency	1258.90	nM	CHEMBL1445094	CHEMBL1614027
8	CHEMBL1440842		65.15	active	Potency	1584.90	nM	CHEMBL1440842	CHEMBL1614027
9	CHEMBL285932		78.40	active	Potency	1995.30	nM	CHEMBL285932	CHEMBL1614027
10	CHEMBL1514440		98.76	active	Potency	3162.30	nM	CHEMBL1514440	CHEMBL1614027
11	CHEMBL515505		63.96	active	Potency	3162.30	nM	CHEMBL515505	CHEMBL1614027
12	CHEMBL1437226		76.06	active	Potency	1000.00	nM	CHEMBL1437226	CHEMBL1614027
13	CHEMBL1412450		53.39	active	AC50	4466.84	nM	CHEMBL1412450	CHEMBL1741325
14	CHEMBL1371202		53.06	active	AC50	3162.28	nM	CHEMBL1371202	CHEMBL1741325
15	CHEMBL1506693		79.53	active	AC50	3162.28	nM	CHEMBL1506693	CHEMBL1741325
16	CHEMBL1510699		72.80	active	AC50	3548.13	nM	CHEMBL1510699	CHEMBL1741325
17	CHEMBL1329402		53.35	active	AC50	2818.38	nM	CHEMBL1329402	CHEMBL1741325
18	CHEMBL1525226		64.48	active	AC50	3162.28	nM	CHEMBL1525226	CHEMBL1741325
19	CHEMBL1619056		54.45	active	AC50	1995.26	nM	CHEMBL1619056	CHEMBL1741325
20	CHEMBL1600150		76.10	active	AC50	1412.54	nM	CHEMBL1600150	CHEMBL1741325
21	CHEMBL1353037		65.13	active	AC50	3981.07	nM	CHEMBL1353037	CHEMBL1741325
22	CHEMBL10009		66.70	active	AC50	1258.93	nM	CHEMBL10009	CHEMBL1741325
23	CHEMBL1352043		75.80	active	AC50	3981.07	nM	CHEMBL1352043	CHEMBL1741325
24	CHEMBL1519456		63.84	active	AC50	1000.00	nM	CHEMBL1519456	CHEMBL1741325
25	CHEMBL1523175		82.24	active	AC50	2818.38	nM	CHEMBL1523175	CHEMBL1741325
26	CHEMBL1496623		73.05	active	AC50	3548.13	nM	CHEMBL1496623	CHEMBL1741325
27	CHEMBL1430479		53.32	active	AC50	1258.93	nM	CHEMBL1430479	CHEMBL1741325
28	CHEMBL1568982		82.28	active	AC50	1258.93	nM	CHEMBL1568982	CHEMBL1741325
29	CHEMBL1414633		79.84	active	AC50	3548.13	nM	CHEMBL1414633	CHEMBL1741325
30	CHEMBL1579521		66.48	active	AC50	1122.02	nM	CHEMBL1579521	CHEMBL1741325
31	CHEMBL1470566		64.42	active	AC50	1000.00	nM	CHEMBL1470566	CHEMBL1741325
32	CHEMBL1391756		63.73	active	AC50	4466.84	nM	CHEMBL1391756	CHEMBL1741325
33	CHEMBL1422230		53.51	active	AC50	2238.72	nM	CHEMBL1422230	CHEMBL1741325
34	CHEMBL1587524		65.69	active	AC50	2238.72	nM	CHEMBL1587524	CHEMBL1741325
35	CHEMBL1359919		53.76	active	AC50	1778.28	nM	CHEMBL1359919	CHEMBL1741325
36	CHEMBL1420689		72.82	active	AC50	1995.26	nM	CHEMBL1420689	CHEMBL1741325
37	CHEMBL1607455		72.92	active	AC50	2511.89	nM	CHEMBL1607455	CHEMBL1741325
38	CHEMBL1567410		53.31	active	AC50	1995.26	nM	CHEMBL1567410	CHEMBL1741325
39	CHEMBL1492293		55.23	active	AC50	4466.84	nM	CHEMBL1492293	CHEMBL1741325
40	CHEMBL1970734		65.83	active	AC50	3981.07	nM	CHEMBL1970734	CHEMBL1741325
41	CHEMBL3214169		53.10	active	AC50	2238.72	nM	CHEMBL3214169	CHEMBL1741325
42	CHEMBL3196654		53.53	active	AC50	1584.89	nM	CHEMBL3196654	CHEMBL1741325
43	CHEMBL3194051		79.90	active	AC50	3548.13	nM	CHEMBL3194051	CHEMBL1741325
44	CHEMBL3210367		53.34	active	AC50	1122.02	nM	CHEMBL3210367	CHEMBL1741325
45	CHEMBL485968		53.43	active	IC50	3000.00	nM	CHEMBL485968	CHEMBL948449
46	CHEMBL1171786		65.12	active	IC50	4168.69	nM	CHEMBL1171786	CHEMBL1174522
47	CHEMBL592510		64.78	active	IC50	1100.00	nM	CHEMBL592510	CHEMBL1072633
48	CHEMBL589625		64.19	active	IC50	2900.00	nM	CHEMBL589625	CHEMBL1072633
49	CHEMBL3310486		53.18	active	IC50	2400.00	nM	CHEMBL3310486	CHEMBL3363904
50	CHEMBL3680983		64.34	active	IC50	4300.00	nM	CHEMBL3680983	CHEMBL3705202
51	CHEMBL3822867		63.74	active	IC50	4190.00	nM	CHEMBL3822867	CHEMBL3825882
52	CHEMBL3775370		63.94	active	IC50	1562.00	nM	CHEMBL3775370	CHEMBL3776779
53	CHEMBL1571639		60.24	decoy	AC50		Unspecified	CHEMBL1571639	CHEMBL1741325
54	CHEMBL1549431		53.35	decoy	AC50		Unspecified	CHEMBL1549431	CHEMBL1741325
55	CHEMBL1329392		63.77	decoy	AC50		Unspecified	CHEMBL1329392	CHEMBL1741325
56	CHEMBL1495854		64.06	decoy	AC50		Unspecified	CHEMBL1495854	CHEMBL1741325
57	CHEMBL2219648		65.67	decoy	AC50		Unspecified	CHEMBL2219648	CHEMBL1741325
58	CHEMBL1607062		65.05	decoy	AC50		Unspecified	CHEMBL1607062	CHEMBL1741325
59	CHEMBL1568369		65.29	decoy	AC50		Unspecified	CHEMBL1568369	CHEMBL1741325
60	CHEMBL369379		64.55	decoy	AC50		Unspecified	CHEMBL369379	CHEMBL1741325
61	CHEMBL1590550		53.43	decoy	AC50		Unspecified	CHEMBL1590550	CHEMBL1741325
62	CHEMBL1448325		63.85	decoy	AC50		Unspecified	CHEMBL1448325	CHEMBL1741325
63	CHEMBL1594134		72.79	decoy	AC50		Unspecified	CHEMBL1594134	CHEMBL1741325
64	CHEMBL1352188		53.40	decoy	AC50		Unspecified	CHEMBL1352188	CHEMBL1741325
65	CHEMBL1400491		64.17	decoy	AC50		Unspecified	CHEMBL1400491	CHEMBL1741325
66	CHEMBL1516701		75.90	decoy	AC50		Unspecified	CHEMBL1516701	CHEMBL1741325
67	CHEMBL1495665		64.04	decoy	AC50		Unspecified	CHEMBL1495665	CHEMBL1741325
68	CHEMBL1329786		64.28	decoy	AC50		Unspecified	CHEMBL1329786	CHEMBL1741325

Hit list of the six ligand-based models: moderate CYP 2C9 inhibitor test set (sorted by pharmacophore-fit score).



	Name	Matching Features	Pharmacophore-Fit Score	Active/Decoy	Bioactivity Type	Activity Value	Activity Units	Compound ID	Assay ID
1	CHEMBL1407064		82.74	active	AC50	7943.28	nM	CHEMBL1407064	CHEMBL1741325
2	CHEMBL3189154		82.51	active	AC50	7079.46	nM	CHEMBL3189154	CHEMBL1741325
3	CHEMBL271579		79.23	active	IC50	8100.00	nM	CHEMBL271579	CHEMBL951259
4	CHEMBL476833		78.35	active	Potency	6309.60	nM	CHEMBL476833	CHEMBL1614027
5	CHEMBL1516701		75.90	decoy	AC50		Unspecified	CHEMBL1516701	CHEMBL1741325
6	CHEMBL1466869		75.87	active	AC50	8912.51	nM	CHEMBL1466869	CHEMBL1741325
7	CHEMBL1583978		73.04	active	AC50	7943.28	nM	CHEMBL1583978	CHEMBL1741325
8	CHEMBL1594134		72.79	decoy	AC50		Unspecified	CHEMBL1594134	CHEMBL1741325
9	CHEMBL1557790		72.60	active	AC50	7079.46	nM	CHEMBL1557790	CHEMBL1741325
10	CHEMBL2178807		66.56	active	IC50	5000.00	nM	CHEMBL2178807	CHEMBL2186605
11	CHEMBL173038		66.49	active	Ki	6000.00	nM	CHEMBL173038	CHEMBL661229
12	CHEMBL2219648		65.67	decoy	AC50		Unspecified	CHEMBL2219648	CHEMBL1741325
13	CHEMBL368348		65.40	active	Ki	6000.00	nM	CHEMBL368348	CHEMBL661229
14	CHEMBL1568369		65.29	decoy	AC50		Unspecified	CHEMBL1568369	CHEMBL1741325
15	CHEMBL1607062		65.05	decoy	AC50		Unspecified	CHEMBL1607062	CHEMBL1741325
16	CHEMBL1349047		64.67	active	AC50	5623.41	nM	CHEMBL1349047	CHEMBL1741325
17	CHEMBL369379		64.55	decoy	AC50		Unspecified	CHEMBL369379	CHEMBL1741325
18	CHEMBL1343392		64.52	active	AC50	7943.28	nM	CHEMBL1343392	CHEMBL1741325
19	CHEMBL2070134		64.29	active	IC50	5000.00	nM	CHEMBL2070134	CHEMBL2073272
20	CHEMBL1329786		64.28	decoy	AC50		Unspecified	CHEMBL1329786	CHEMBL1741325
21	CHEMBL1400491		64.17	decoy	AC50		Unspecified	CHEMBL1400491	CHEMBL1741325
22	CHEMBL590593		64.16	active	IC50	7900.00	nM	CHEMBL590593	CHEMBL1072635
23	CHEMBL28626		64.10	active	Potency	10000.00	nM	CHEMBL28626	CHEMBL1614027
24	CHEMBL1495854		64.06	decoy	AC50		Unspecified	CHEMBL1495854	CHEMBL1741325
25	CHEMBL1495665		64.04	decoy	AC50		Unspecified	CHEMBL1495665	CHEMBL1741325
26	CHEMBL1448325		63.85	decoy	AC50		Unspecified	CHEMBL1448325	CHEMBL1741325
27	CHEMBL1329392		63.77	decoy	AC50		Unspecified	CHEMBL1329392	CHEMBL1741325
28	CHEMBL2018924		63.71	active	IC50	6180.00	nM	CHEMBL2018924	CHEMBL2020720
29	CHEMBL1381787		63.62	active	AC50	8912.51	nM	CHEMBL1381787	CHEMBL1741325
30	CHEMBL3214400		63.39	active	AC50	7079.46	nM	CHEMBL3214400	CHEMBL1741325
31	CHEMBL1571639		60.24	decoy	AC50		Unspecified	CHEMBL1571639	CHEMBL1741325
32	CHEMBL3105228		60.08	active	IC50	7400.00	nM	CHEMBL3105228	CHEMBL3108383
33	CHEMBL1401126		54.87	active	AC50	5623.41	nM	CHEMBL1401126	CHEMBL1741325
34	CHEMBL1432284		54.84	active	AC50	8912.51	nM	CHEMBL1432284	CHEMBL1741325
35	CHEMBL1533024		53.55	active	AC50	7943.28	nM	CHEMBL1533024	CHEMBL1741325
36	CHEMBL1590550		53.43	decoy	AC50		Unspecified	CHEMBL1590550	CHEMBL1741325
37	CHEMBL1352188		53.40	decoy	AC50		Unspecified	CHEMBL1352188	CHEMBL1741325
38	CHEMBL1549431		53.35	decoy	AC50		Unspecified	CHEMBL1549431	CHEMBL1741325
39	CHEMBL527271		53.32	active	AC50	7079.46	nM	CHEMBL527271	CHEMBL1741325

Hit list of the six ligand-based models: weak CYP 2C9 inhibitor test set (sorted by pharmacophore-fit score).



	Name	Matching Features	Pharmacophore-Fit Score	Active/Decoy	Bioactivity Type	Activity Value	Activity Units	Compound ID	Assay ID
1	CHEMBL1516701		75.90	decoy	AC50		Unspecified	CHEMBL1516701	CHEMBL1741325
2	CHEMBL1348445		75.82	active	AC50	12589.25	nM	CHEMBL1348445	CHEMBL1741325
3	CHEMBL1096182		75.73	active	IC50	18000.00	nM	CHEMBL1096182	CHEMBL1110981
4	CHEMBL584261		74.94	active	IC50	10000.00	nM	CHEMBL584261	CHEMBL1046296
5	CHEMBL1379160		73.11	active	AC50	15848.93	nM	CHEMBL1379160	CHEMBL1741325
6	CHEMBL1447768		73.08	active	AC50	10000.00	nM	CHEMBL1447768	CHEMBL1741325
7	CHEMBL1594134		72.79	decoy	AC50		Unspecified	CHEMBL1594134	CHEMBL1741325
8	CHEMBL2178247		66.43	active	IC50	10000.00	nM	CHEMBL2178247	CHEMBL2186605
9	CHEMBL1992293		66.29	active	AC50	14125.38	nM	CHEMBL1992293	CHEMBL1741325
10	CHEMBL2204469		65.94	active	IC50	16000.00	nM	CHEMBL2204469	CHEMBL2209448
11	CHEMBL1760809		65.83	active	IC50	26000.00	nM	CHEMBL1760809	CHEMBL1764826
12	CHEMBL1403583		65.81	active	AC50	10000.00	nM	CHEMBL1403583	CHEMBL1741325
13	CHEMBL1668243		65.79	active	IC50	18000.00	nM	CHEMBL1668243	CHEMBL1670713
14	CHEMBL2012958		65.78	active	IC50	10000.00	nM	CHEMBL2012958	CHEMBL2014810
15	CHEMBL2219648		65.67	decoy	AC50		Unspecified	CHEMBL2219648	CHEMBL1741325
16	CHEMBL1080787		65.58	active	IC50	30000.00	nM	CHEMBL1080787	CHEMBL1112317
17	CHEMBL1359326		65.37	active	AC50	15848.93	nM	CHEMBL1359326	CHEMBL1741325
18	CHEMBL1568369		65.29	decoy	AC50		Unspecified	CHEMBL1568369	CHEMBL1741325
19	CHEMBL575966		65.08	active	IC50	10000.00	nM	CHEMBL575966	CHEMBL1040537
20	CHEMBL1607062		65.05	decoy	AC50		Unspecified	CHEMBL1607062	CHEMBL1741325
21	CHEMBL520276		65.02	active	IC50	14000.00	nM	CHEMBL520276	CHEMBL963888
22	CHEMBL1600653		64.99	active	AC50	15848.93	nM	CHEMBL1600653	CHEMBL1741325
23	CHEMBL555563		64.74	active	IC50	10000.00	nM	CHEMBL555563	CHEMBL1954252
24	CHEMBL369379		64.55	decoy	AC50		Unspecified	CHEMBL369379	CHEMBL1741325
25	CHEMBL3627895		64.48	active	IC50	19820.00	nM	CHEMBL3627895	CHEMBL3630991
26	CHEMBL2392691		64.45	active	IC50	25000.00	nM	CHEMBL2392691	CHEMBL2393701
27	CHEMBL1308621		64.36	active	AC50	10000.00	nM	CHEMBL1308621	CHEMBL1741325
28	CHEMBL1329786		64.28	decoy	AC50		Unspecified	CHEMBL1329786	CHEMBL1741325
29	CHEMBL270902		64.28	active	IC50	10000.00	nM	CHEMBL270902	CHEMBL951259
30	CHEMBL3699081		64.27	active	IC50	13000.00	nM	CHEMBL3699081	CHEMBL3706057
31	CHEMBL3192772		64.24	active	AC50	14125.38	nM	CHEMBL3192772	CHEMBL1741325
32	CHEMBL1742180		64.19	active	AC50	14125.38	nM	CHEMBL1742180	CHEMBL1741325
33	CHEMBL1459301		64.19	active	AC50	11220.18	nM	CHEMBL1459301	CHEMBL1741325
34	CHEMBL1400491		64.17	decoy	AC50		Unspecified	CHEMBL1400491	CHEMBL1741325
35	CHEMBL171828		64.16	active	Ki	20000.00	nM	CHEMBL171828	CHEMBL661229
36	CHEMBL1495854		64.06	decoy	AC50		Unspecified	CHEMBL1495854	CHEMBL1741325
37	CHEMBL1495665		64.04	decoy	AC50		Unspecified	CHEMBL1495665	CHEMBL1741325
38	CHEMBL1448325		63.85	decoy	AC50		Unspecified	CHEMBL1448325	CHEMBL1741325
39	CHEMBL3819521		63.84	active	IC50	10000.00	nM	CHEMBL3819521	CHEMBL3821616
40	CHEMBL1329392		63.77	decoy	AC50		Unspecified	CHEMBL1329392	CHEMBL1741325
41	CHEMBL1344162		60.44	active	AC50	10000.00	nM	CHEMBL1344162	CHEMBL1741325
42	CHEMBL1571639		60.24	decoy	AC50		Unspecified	CHEMBL1571639	CHEMBL1741325
43	CHEMBL1580156		60.02	active	AC50	12589.25	nM	CHEMBL1580156	CHEMBL1741325
44	CHEMBL476302		54.83	active	IC50	17000.00	nM	CHEMBL476302	CHEMBL982037
45	CHEMBL1439808		53.87	active	AC50	11220.18	nM	CHEMBL1439808	CHEMBL1741325
46	CHEMBL1304990		53.66	active	AC50	11220.18	nM	CHEMBL1304990	CHEMBL1741325
47	CHEMBL1491709		53.62	active	AC50	12589.25	nM	CHEMBL1491709	CHEMBL1741325
48	CHEMBL1741376		53.57	active	AC50	10000.00	nM	CHEMBL1741376	CHEMBL1741325
49	CHEMBL1470129		53.56	active	AC50	12589.25	nM	CHEMBL1470129	CHEMBL1741325
50	CHEMBL1343857		53.51	active	AC50	12589.25	nM	CHEMBL1343857	CHEMBL1741325
51	CHEMBL1590550		53.43	decoy	AC50		Unspecified	CHEMBL1590550	CHEMBL1741325
52	CHEMBL1352188		53.40	decoy	AC50		Unspecified	CHEMBL1352188	CHEMBL1741325
53	CHEMBL3209485		53.38	active	AC50	10000.00	nM	CHEMBL3209485	CHEMBL1741325
54	CHEMBL576138		53.36	active	IC50	10000.00	nM	CHEMBL576138	CHEMBL1040537
55	CHEMBL1311305		53.36	active	AC50	11220.18	nM	CHEMBL1311305	CHEMBL1741325
56	CHEMBL1549431		53.35	decoy	AC50		Unspecified	CHEMBL1549431	CHEMBL1741325
57	CHEMBL1346850		53.32	active	AC50	12589.25	nM	CHEMBL1346850	CHEMBL1741325
58	CHEMBL3699082		53.31	active	IC50	42000.00	nM	CHEMBL3699082	CHEMBL3706057
59	CHEMBL3353881		53.29	active	IC50	17000.00	nM	CHEMBL3353881	CHEMBL3373572
60	CHEMBL1872483		53.21	active	IC50	30000.00	nM	CHEMBL1872483	CHEMBL3404119
61	CHEMBL1602892		53.13	active	AC50	11220.18	nM	CHEMBL1602892	CHEMBL1741325
62	CHEMBL1345025		53.03	active	AC50	10000.00	nM	CHEMBL1345025	CHEMBL1741325