DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Potential new TTK kinase inhibitors by the means of structure-based pharmacophore modeling and docking experiments“

verfasst von / submitted by

Katharina Schwartz

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

Magistra der Pharmazie (Mag.pharm.)

Wien, 2020 / Vienna, 2020

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. Gerhard F. Ecker
Mitbetreut von / Co-Supervisor:	Riccardo Martini, PhD

Kurzfassung

Das Multiple Myelom ist eine im Knochenmark entstehende Krebserkrankung die zu einer erhöhten Anzahl an Plasmazellen führt. Das Multiple Myelom ist eine unheilbare Krankheit, an der jährlich tausende von Menschen sterben. [1] Es gibt keine effektive Behandlung dieser Erkrankung, weshalb lediglich eine Steigerung der Lebensqualität möglich ist.

Neuere computergestützter Studien haben einen Signalweg identifiziert, welcher zu einer Hochrisikokrankheit führt und unter anderem die Proteinkinase TTK (auch MPS1 genannt) mit dualer Spezifizität beinhaltet. [2] Dieses Enzym stellt einen entscheidenden mitotischen Kontrollpunkt während der Zentrosomen Duplikation dar und führt zu Aneuploidie und letztendlich zum Zelltod. [3]

Ziel dieser Arbeit ist es, potentielle Inhibitoren der TTK Kinase mittels in-silico Methoden zu identifizieren. Zuerst wurden die verfügbaren TTK-Liganden Komplexe in der Proteindatenbank analysiert und anschließend ein virtuelles Screening-Verfahren implementiert, welches strukturbasiertes Pharmacophore Modeling, molekulare Docking-Simulationen, MM-GBSA-Berechnungen und manuelle Vergleiche von Posen kombiniert. Durch Clustering und Ähnlichkeitssuche wurde das Ergebnis des virtuellen Screenings verfeinert um unterschiedliche Scaffolds zu gewährleisten. Schließlich wurde eine Liste potentieller neuer TTK-Inhibitoren für weitere Experimente vorgeschlagen.

Schlagwörter: Kinase, TTK, Multiples Myelom, Inhibitoren

Abstract

Every year thousands of people die as a result of multiple myeloma. [1] This type of cancer is located in the bone marrow and leads to an increased amount of plasma cells and multiple myeloma is a terminal illness. Currently there is no effective treatment available, but there are possibilities to improve the patients quality of life.

Recent computational studies have identified a pathway that leads to a high-risk disease and includes the dual specificity protein kinase TTK (also called MPS1) as one of the factors. [2] This enzyme is a crucial mitotic checkpoint during the centrosome duplication and, if inhibited, leads to aneuploidy and ultimately to cell death. [3]

The aim of this thesis is to identify potential inhibitors of the TTK kinase using in-silico methods. Starting from the analysis of TTK-ligand complexes available on the Protein Data Bank, a virtual screening method that combines structure-based pharmacophore modeling, molecular docking simulations, MM-GBSA calculations and pose comparison was developed. To guarantee different scaffolds, a refinement of the virtual screening output with clustering and similarity search was done. Finally a list of potential new TTK inhibitors was proposed for further experiments.

Acknowledgement

Firstly, I would like to express my sincere gratitude to my advisor Univ.-Prof. Dr. Gerhard F. Ecker for the excellent support of my thesis. I could not have imagined a better advisor and mentor for my diploma thesis. He is a superior scientist and also an admirable group leader of the Pharmacoinformatic Research Group.

Very special thanks to my Co-Supervisor Riccardo Martini too, who was available all the time and gave me an outstanding support. Thank you for sharing your knowledge with me and for having such a nice time in the group.

I also want to thank other members of the "Pharminfomaniacs" group. The atmosphere in the group is really relaxed, so it is very pleasant to work there the whole day. All people are really cooperative and I appreciated all the help from the group.

Last but not least, I want to thank my whole family, my friends and also my boyfriend, who supported me with love and understanding. Without you, I could never have reached this current level of success.

Thank you all for your support.

Table of contents

1 Introduction

Studies of the American Cancer Society Inc. show that in 2019 more than 1.7 million new cancer cases are expected to be diagnosed in the United States of America only. About 606.880 Americans are expected to die of cancer in 2019 (Figure 1), which translates to about 1.660 deaths per day. Cancer is the second leading cause of death in the US, after heart diseases. [1]

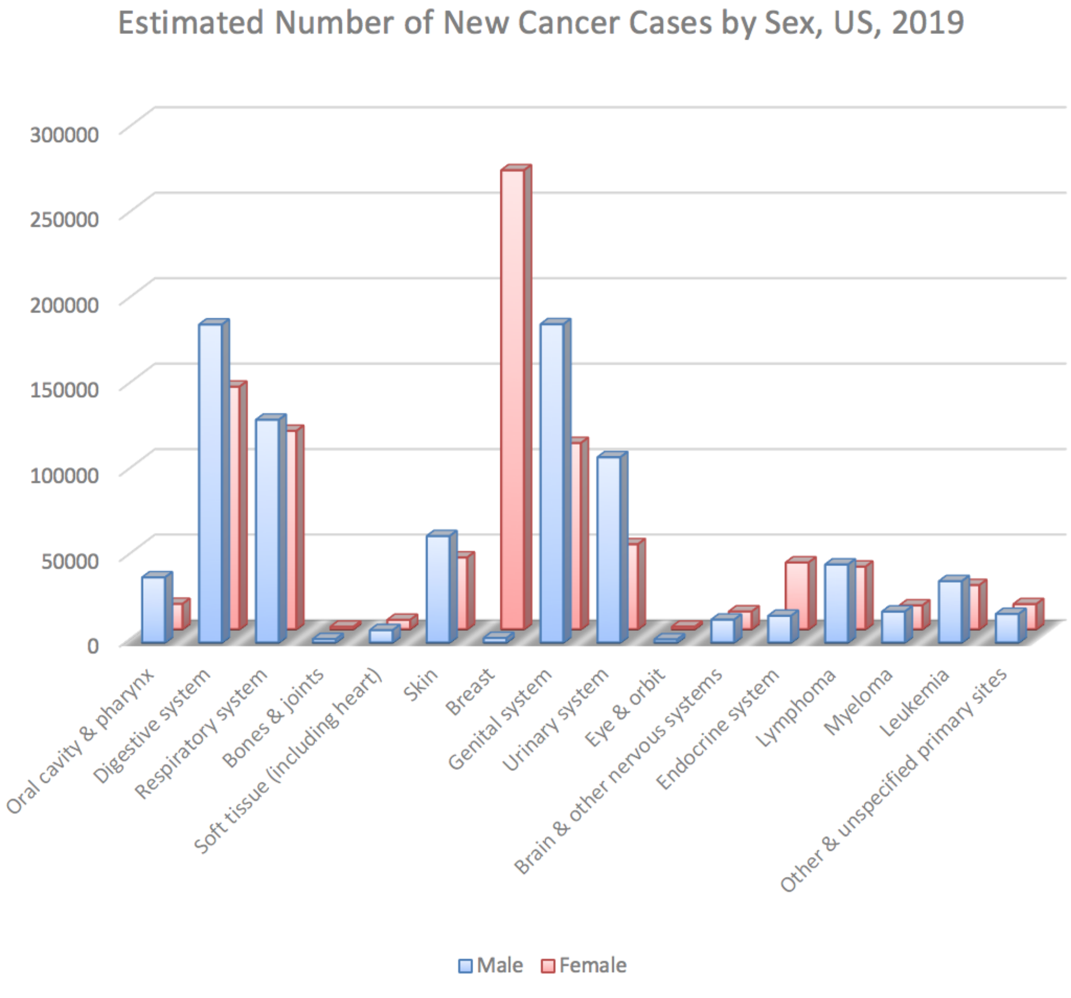


Figure 1: Estimated number of new cancer cases by sex, US 2019

Figure 1 shows the different types of cancer and the frequency of each type. Relating to myeloma, the estimated number of new cancer cases including male and female is about 32.100. In comparison to that, the digestive system, the most frequent type of cancer, has an count of 328.030 including male and female. Based on this counts, the digestive system cancer occurs 10 times more often compared to the myeloma. This disorder is uncommon, therefore this type of cancer has not been explored so far and thereby the treatment options are limited.

1.1 Multiple myeloma

Multiple myeloma (MM) is a rare cancer of plasma cells. It is important to understand the further development of a blood stem cell to a plasma cell (Figure 2). Normal plasma cells are in the bone marrow, a spongy tissue inside the bones. The bone marrow produces red blood cells along with white blood cells and platelets every day. Red blood cells are essential for the carriage of oxygen to supply every part of the body. A low count of red blood cells can lead to impaired physical capability. The function of the white blood cells is to protect against foreign bodies, such as viruses, bacteria and molds. A low count of white blood cells reduces the body's ability to fight diseases. The platelets play an important role in blood clotting to help to prevent bleeding. If the blood has a low number of platelets the risk for serious bleeding is increased.

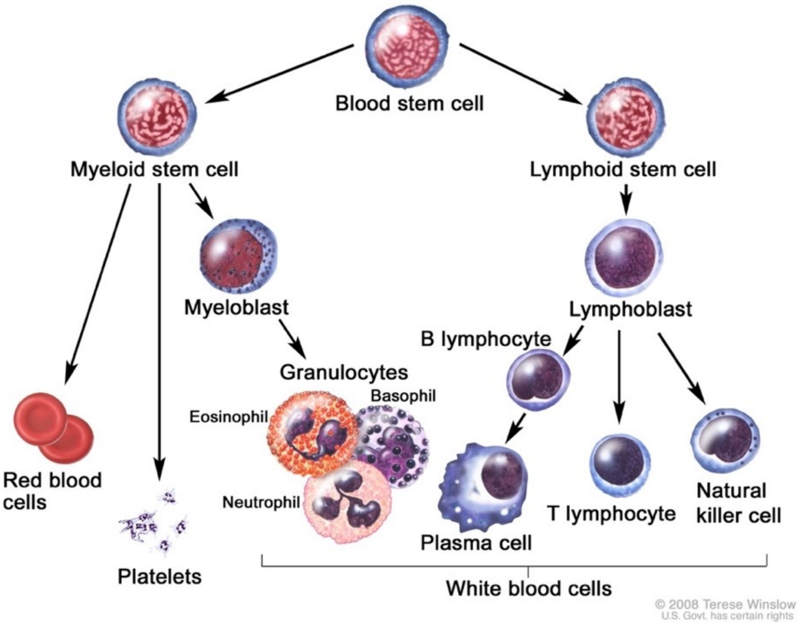


Figure 2: Blood cell development [3]

Normal plasma cells play an essential role for our immune system that is made from a network of various types of cells to protect the body from germs and other harmful substances.

Lymphocytes are one of the main types of white blood cells in the immune system and include T- and B-cells. They are located in many areas of the body, such as lymph nodes and the bone marrow. When B-cells react to an infection they change into plasma cells. When plasma cells become cancerous and begin to grow unregulated and form tumors in bones of the body, this is called multiple myeloma. [4]

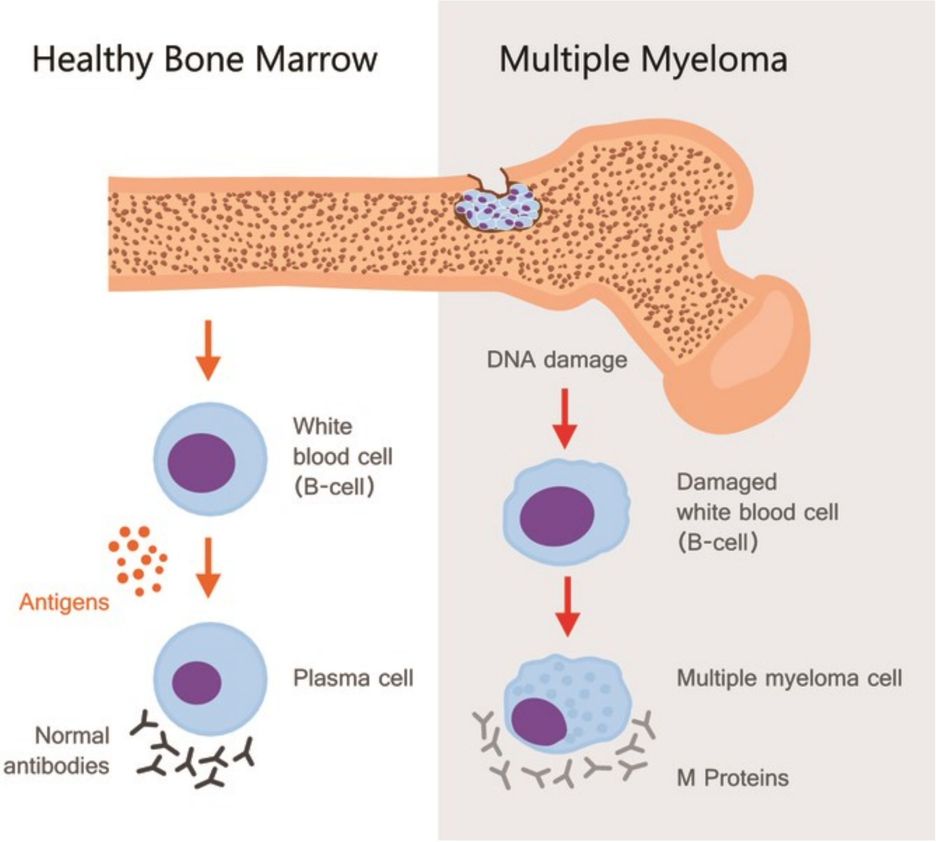


Figure 3: Development of M-Proteins [5]

Normal plasma cells produce antibodies to combat infections. Malignant plasma cells clone and accumulate in the bone marrow. The number of normal plasma cells is reduced, which can lead to a number of signs and symptoms. In most patients the malignant plasma cells secrete an antibody called the M- protein or M-spike (Figure 3). Their presence is mostly associated with MM and they are detectable in blood or urine. [6]

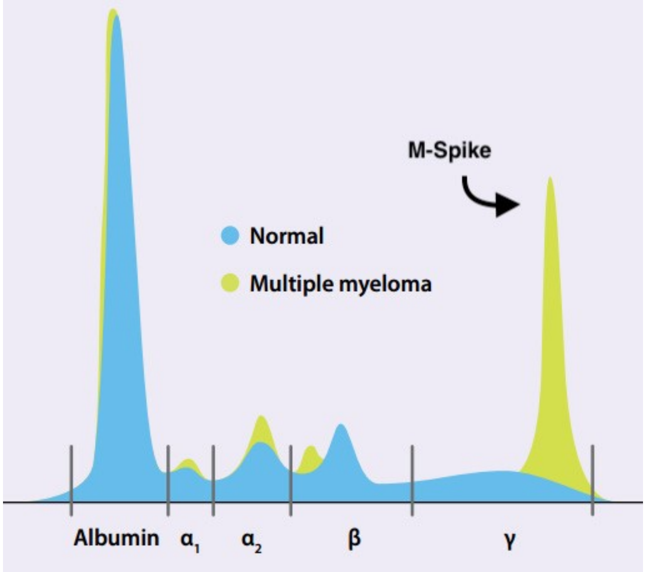


Figure 4: Serum Protein Electrophoresis (SPEP) [7]

Figure 4 shows an abnormal serum protein electrophoresis in a patient with multiple myeloma. SPEP is used to separate and identify the presence of M-Proteins. The M-Spike appears as a large peak on the graph in the gamma region and this is an indication of MM.

1.1.1 Epidemiology

Multiple myeloma is the second most common hematological malign disease with an incidence of six per 100.000 per year in the USA and Europe. The incidence of MM is up to three times higher within African Americans, therefore it is the most frequent hematological malignancy in this ethnic group. The risk of developing MM in old age is higher, therefore the average age at diagnosis is 69 years. 75% of patients being diagnosed above the age of 55 years and 66% of patients being men. With the development of effective therapeutic strategies and enhancements in supportive treatment, the median survival has raised from three to six years in the past two decades. [8]

1.1.2 Symptoms

Multiple myeloma symptoms may develop slowly over time. Oftentimes early-stage multiple myeloma is asymptomatic and symptoms does not appear until the serious illness reaches an advanced stage. Multiple myeloma symptoms can vary widely for each patient.

The most common symptoms are: [9]

- Renal failure
- Anemia
- Bone pain, often in the back or ribs
- Frequent infections
- Hypercalcemia
- Unintentional weight loss

There is also a rare type of MM called smoldering multiple myeloma (SMM). In this case no symptoms or signs occur. This type of MM is not a cancer but it can develop to a cancer, so it is important to get a regular medical checkup. Damage of bones or the kidneys can happen in a small amount.

1.1.3 Causes

Scientists still do not realize exactly what causes the development of malignant plasma cells, thus it is a genetically complex disorder. The following genetic changes in DNA can cause cancer: [10]

- **Mutation**: Cancer can be caused by mistakes or defects in the DNA that turn on oncogenes or turn off tumor suppressor genes. Studies have found that abnormalities of some oncogenes develop early in the course of plasma cell tumors
- **Duplication**: Normal human cells contain 46 chromosomes, some cancer cells may have extra chromosomes
- **Deletion**: Some cancer cells have all or a part of a chromosome missing. These leads to more aggressive myeloma cells and to resistance of the treatment
- **Translocation**: This means that a part of one chromosome has switched with a part of another chromosome. This happens in about 50% of all patients. Some of those higher-risk genetic changes are translocations involving chromosome 14 with chromosomes 4, 11 or 16 and deletion of a section of chromosome 17 [11]

1.1.4 Risk factors

Factors that may increase the risk of multiple myeloma:

- **Age**: The median age at diagnosis is 69 years [8]
- **Sex**: Male are more prone [12]
- **Race**: The risk is three times higher within African Americans [8]
- **Genetic predisposition**: Having a brother or mother with MM [12]

1.1.5 Complications

Typical complications of MM consist of: [12]

- **Frequent infections**: The immune system is often weakened or not fully functional
- **Bone problems**: Figure 5 displays osteoporosis, which leads to bone pain and thinning bones. Bones become less dense and so they are more likely to break
- **Reduced kidney function**: MM may cause kidney failure, therefore it is necessary to get dialysis or if the worst comes to the worst kidney transplantation
- **Anemia**: Multiple myeloma can lead to a significant reduction of red blood cells

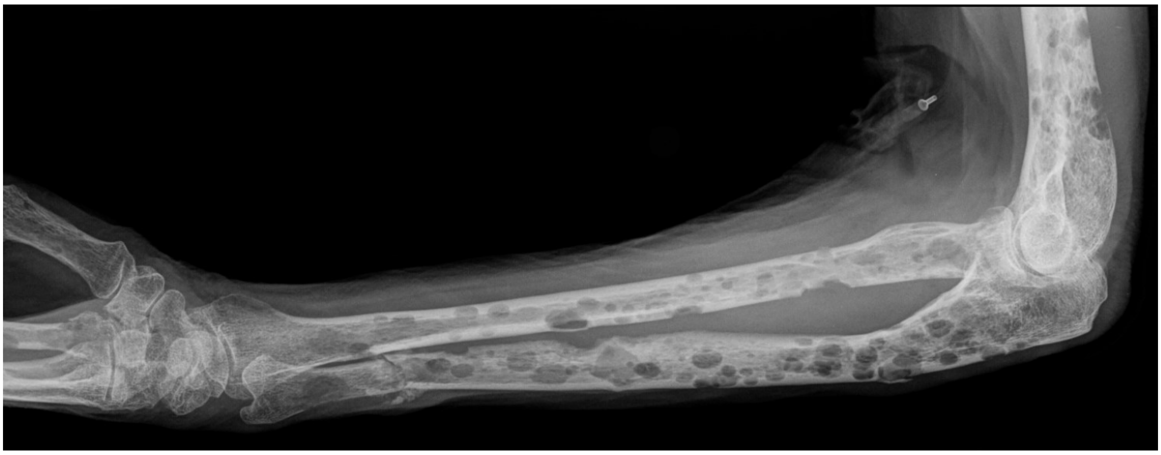


Figure 5: X-ray with multiple osteolytic lesions in the forearm [13]

1.1.6 Diagnosis

In most cases, multiple myeloma will not be diagnosed directly. Often the diagnosis happens accidentally by another medical investigation. Various methods can be applied to confirm the diagnosis of a suspected MM patient (Figure 6) : [14]

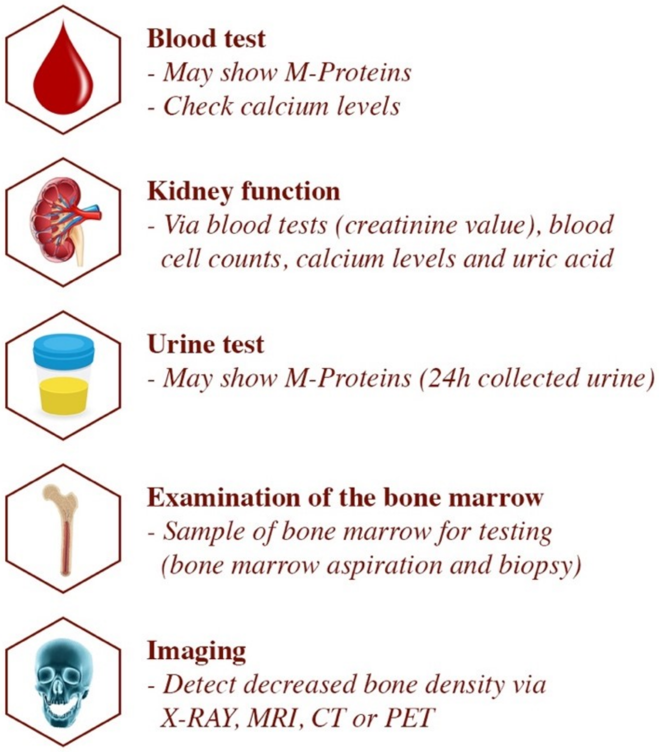

Blood test
- May show M-Proteins
- Check calcium levels
Kidney function
- Via blood tests (creatinine value), blood cell counts, calcium levels and uric acid
Urine test
- May show M-Proteins (24h collected urine)
Examination of the bone marrow
- Sample of bone marrow for testing (bone marrow aspiration and biopsy)
Imaging
- Detect decreased bone density via X-RAY, MRI, CT or PET

Figure 6: Different possibilities to diagnose MM

1.1.7 Treatments

Multiple myeloma is a type of blood cancer which is not curable yet but there are effective ways of treatment to improve life quality and slow down the progress of cancer growth. Patients diagnosed with MM today have a much greater life expectancy due to the newest development of medicines.

There are two kinds of treatments:

- **Non-intensive**: for older and less fit patients, lower doses
- **Intensive**: for younger and fitter patients, higher doses, stem cell transplantation (stem cells will be collected before the treatment or they will be donated by a suitable person). [14]

This cancer is often treated with a combination of drugs including alkylating chemotherapeutic agents, glucocorticoids, Thalidomide / Lenalidomide or Bortezomib, nitrogen-containing bisphosphonates together with a support therapy (antiemetic therapy, blood transfusions, dialysis and pain therapy). [15]

1.2 Multiple myeloma pathway

Ludwig et al. [2] have developed a computational model that helps researchers to focus on the most promising targets referring to multiple myeloma. The result from this model includes the identification of a pathway, driving a high-risk disease (progression or death within 18 months). [16] There are 17 genes detected as high risk drivers, which can be seen in Table 1.

CDK1	Unconfirmed in IFM / DFCI Experimental validation	RELA	Confirmed in IFM / DFCI
PLK4	Confirmed in IFM / DFCI Experimental validation	E2F1	Confirmed in IFM / DFCI
MELK	Confirmed in IFM / DFCI Experimental validation	TONSL	Confirmed in IFM / DFCI
TTK	**Confirmed in IFM / DFCI**	FOXM1	Confirmed in IFM / DFCI
BUB1B	Confirmed in IFM / DFCI	MYBL2	Confirmed in IFM / DFCI
PKMYT1	Confirmed in IFM / DFCI	WDHD1	Confirmed in IFM / DFCI
CREBL2	Confirmed in IFM / DFCI	NEK2	Unconfirmed in IFM / DFCI
ZBTB4	Confirmed in IFM / DFCI	AURKA	Unconfirmed in IFM / DFCI
PHF1	Confirmed in IFM / DFCI		

Table 1: 17 genes detected as high risk drivers in MM [16]

2 Aim of the thesis

The aim of this thesis is to identify potential inhibitors of the TTK kinase by using in silico methods. Based on the analysis of the TTK-ligand complexes available in the protein database, a virtual screening method was performed, which includes structure-based pharmacophore modeling, molecular docking simulations, MM-GBSA calculations and pose comparisons. To ensure diverse compounds, a refinement of the virtual screening result was carried out with clustering and similarity search. In a final step a list of potential new TTK inhibitors for additional experiments was recommended.

2.1 Scheme of the thesis

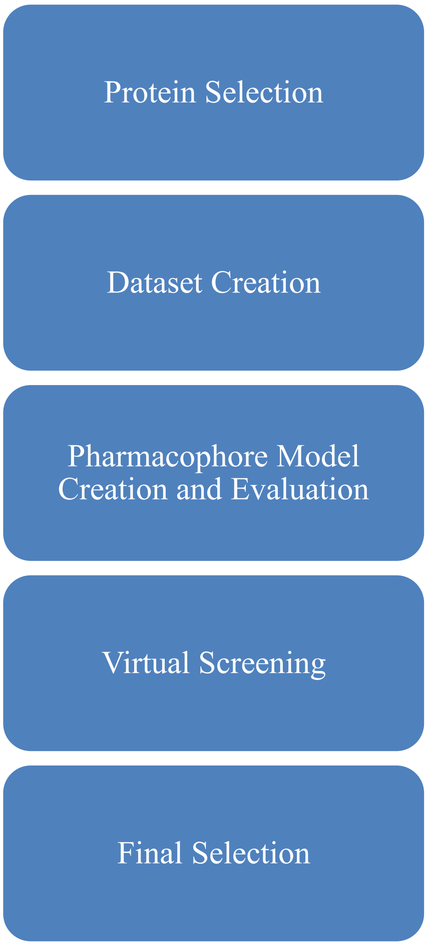


Figure 7: schematic overview of the thesis

3 Computational background

In the last few years, Computer-Aided Drug Design (CADD) has evolved into a powerful technology that is used in the development of new drug candidates. Using sophisticated *in silico* methods to accelerate the drug discovery process of new molecules is state of the art. [17] In this thesis several relevant databases and many essential techniques of CADD were used. In structure-based drug design, the first priority is to obtain the target structure before the receptor-ligand relationship can be explored. Therefore, protein structures can be easily retrieved from protein databases such as Protein Data Bank and Uniprot. An additional essential database for working in silico is ChEMBL. The database contains the required knowledge about the bioactivity of small molecules. ChEMBL has become a very valuable resource due to the large number of entries. Another highly useful database is the DUD-E database. It is one of the largest databases of decoys available to the public. With all these fundamental data from a variety of databases, the design of a pharmacophor model can be started.

3.1 Databases and online resources

3.1.1 Protein Data Bank

In 1970 an association of crystallographers founded the Research Collaborators for Structural Bioinformatics Protein Data Bank (RCSB PDB). The basis for this is the three-dimensional structural data acquisition of biomolecules including proteins and nucleic acids. Experimental methods such as X-RAY crystallography, NMR spectroscopy and electron microscopy make a significant contribution to the visualization of biological molecules. The database can easily be searched for the three-dimensional protein structure using different search options. [18] The identification of active substances by means of in silico is only possible with the aid of crystallography, so the PDB often forms the basis for further scientific research and development.

3.1.2 Uniprot

The Universal Protein Resource (UniProt) is a reliable, comprehensive, freely accessible, funded collection of protein sequences. [19] The aim of the database is to provide scientists with a resource of sequence and functional information available for proteins. The UniProt database contains over 60 million protein sequences with further information about the biological function of proteins. [20]

3.1.3 ChEMBL

ChEMBL is a large open bioactivity database containing approximately 15.504.603 activities, 12.482 targets, 1.879.206 compounds.
In addition, ChEMBL contains compounds in clinical development and patents. The data are taken from the scientific primary literature. The database aims to help scientists to improve their understanding of what constitutes an effective drug. [21]

3.1.4 DUD-E

The database of useful decoys-enhanced (DUD-E) is a tool for researches and contains decoys for virtual screening. DUD-E decoys are based on physical properties of ligands such as, molecular weight, calculated logP, number of rotatable bonds, and hydrogen bond donors and acceptors. The decoys were selected to be physically similar to ligands, which makes docking difficult, but topologically dissimilar to minimize the likelihood of actual binding. DUD-E uses 2-D similarity fingerprints to minimize the topological similarity between decoys and ligands. [22]

3.1.5 ChemMine

ChemMine is an online platform for the analysis and clustering of small molecules.
[23] Clustering of compounds according to structural or property similarity is a relevant approach and is also frequently used for diversity analysis. ChemMine offers three clustering algorithms consisting of hierarchical clustering, multidimensional scaling (MDS) and binning clustering.

3.2 Pharmacophore modeling

3.2.1 Introduction

The International Union of Pure and Applied Chemistry (IUPAC) defines a pharmacophore model as "*an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response*". [24]

A three-dimensional pharmacophore model can be defined as set of chemical features or functionalities aligned in a three-dimensional space.

The pharmacophore reduces a molecule to a collection of features at the 2D or 3D level. Any type of atom or functional group in a molecule can be converted to a pharmacophore feature. [25]

A pharmacophore is not equivalent to a specific molecule, but an abstract concept that reflects the shared molecular properties of the interaction with the receptor. [26]

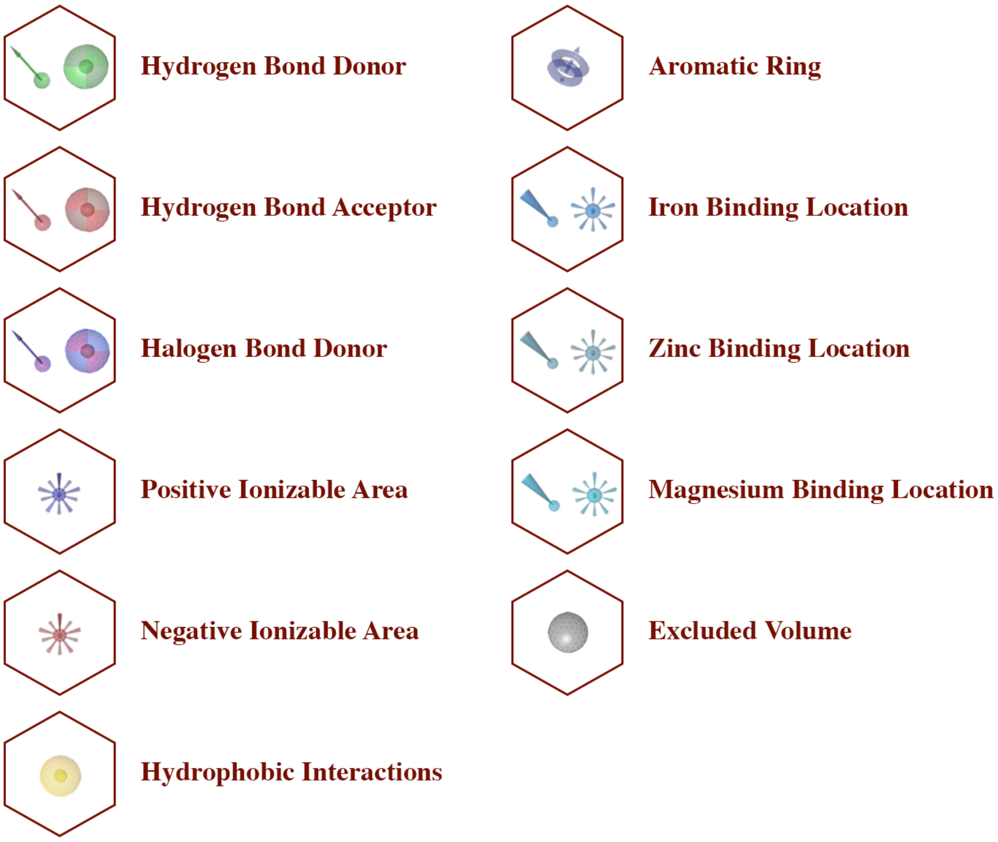


Figure 8: Basic chemical features of LigandScout

Depending on the situation there are two major approaches for pharmacophore modeling available to construct pharmacophore models, either the structure-based design or the ligand-based design.

The structure-based approach is valuable for *in silico* drug discovery to explore the bonding between a ligand and a biological structure. This model requires knowledge of the three-dimensional structure of the biological target. It involves an analysis of the possible interaction sites between the macromolecular target and the ligands. [24]

Ligand-based pharmacophore modeling has also become an important computational technique for drug development. In the absence of the macromolecular target structure, the ligand-based method is required. The ligand-based pharmacophore extracts common chemical properties from 3D structures of a number of known ligands and superimposes them. [27]

Commercial software products such as LigandScout, MOE and Phase are available for structure- and ligand-based pharmacophore modeling.

The tool of choice for evaluating the performance of the pharmacophore model is the ROC curve. A ROC curve illustrates the rate of active compounds on the x axis and the rate of decoys on the y axis. When following the dashed diagonal line, an insignificant classification model that cannot distinguish between decoys and actives is visualized.

The enrichment factor describes the number of active chemical compounds detected by using a specific pharmacophore model, as opposed to the amount hypothetically found when compounds are arbitrarily investigated. [28]

3.3 Docking

Docking has become a relevant and efficient tool of drug discovery and design. It shows the interaction between a small molecule and a protein. (Figure 9) There are two major reasons for performing docking studies. The first one is to predict the binding mode of a ligand into a receptor and the second is to obtain new hits in virtual screening. It is an optimization task to identify the ligand conformation bound to the target with the most favorable binding energy. Docking requires two operations: sampling conformations of the ligand in the active site of the protein and ranking these conformations via scoring functions afterwards. Docking is especially effective in reducing a collection of virtual compounds down to a feasible number to focus the researcher's attention on the most meaningful compounds.

Some of the most widely used docking programs are DOCK, Gold, FlexX, Glide and AUTODOCK.

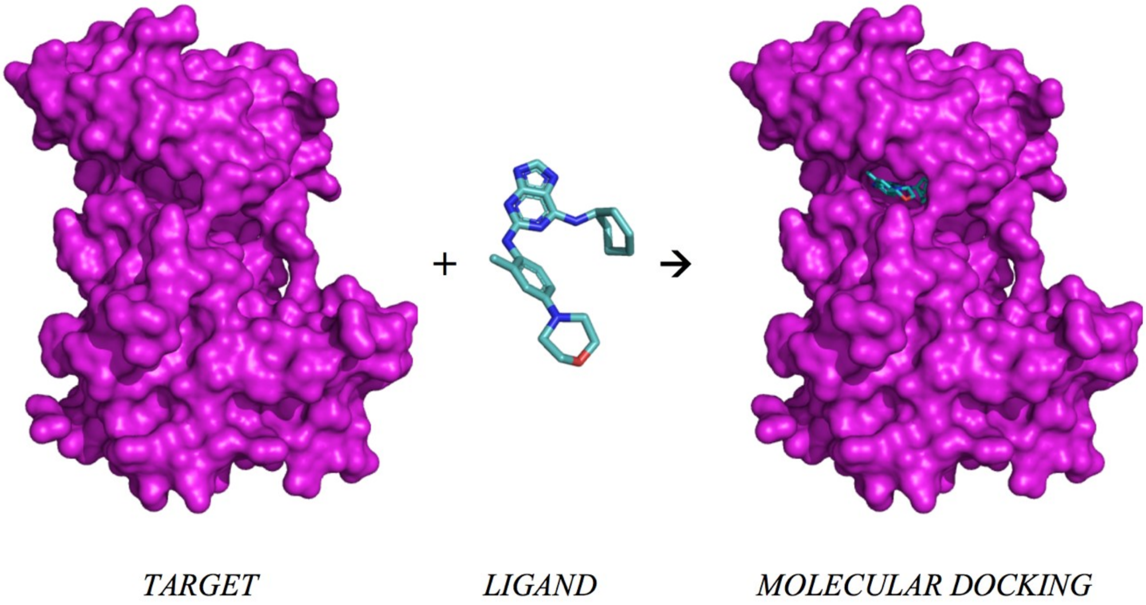


Figure 9: Schema of molecular docking (PDB: 5N7V)

3.3.1 Scoring functions

Scoring functions are one of the most important components in the evaluation of, for example, a docking run. Docking algorithms generate scores that allow ideally to distinguish between strong-binding and weak-binding molecules (according to their active site). A scoring function is used to rank ligand orientations/conformations according to their quality of binding. An ideal scoring function would rank the experimentally determined binding mode on top of all poses. [29]

3.4 Virtual screening

Virtual screening, also known as *in-silico* screening, is a cost-efficient, powerful and rapid technology for hit identification. It is commonly used in academic laboratories as well as in pharmaceutical companies. The objective of screening a large virtual database of chemical compounds is to identify a set of potential candidates via computer-aided methods, followed by synthesizing and experimental testing for their biological activity. There are various programs for virtual screening available, in this thesis LigandScout was used.

3.5 Software

3.5.1 LigandScout (Inte:ligand)

LigandScout is a powerful structure- and ligand-based program for generating pharmacophore models. It uses fast algorithms to perform alignments and to interpret ligand-macromolecule interactions. It extracts and interprets ligands and their macromolecular environment from PDB files and creates three-dimensional pharmacophore models. Fast and accurate *in silico* screening of compound libraries is integrated along with tools to analyze the performance of the models. [30]

3.5.2 PyMOL (Schrödinger Inc.)

PyMOL generates three-dimensional images of small molecules and biological macromolecules (e.g. proteins) and can visualize and analyze their structure. The user can select parts of the protein and change the color to highlight relevant structural features or to change the view of the protein. The advantage of PyMOL is that it is an open-source software.

3.5.3 Glide (Schrödinger Inc.)

Glide is used for ligand-receptor docking and offers a high-throughput virtual screening mode for efficient enrichment of millions of compound libraries, either with the standard precision mode (SP) or with the extra precision mode (XP). Glide searches for favorable interactions between one or numerous ligands and a receptor. Docking calculations require a meaningful and careful preparation before a docking operations is performed. Receptor preparation is carried out with the Protein Preparation Wizard and ligand preparation with the Ligand Preparation Wizard.

3.5.4 Knime Analytics Platform

Knime is an open source software workflow management system for creating data science applications and services. The software helps to understand data and to design data science workflows. In KNIME, the user can build workflows, which are made up of nodes and connections in a simple way, and the data can be transferred between the individual nodes.

A workflow usually starts with a node that reads data from a data source that is usually text files. Imported data is stored in an internal table-based format consisting of columns with a certain data type and a variable number of rows conforming to the column specifications. These data tables are sent along the connections to other nodes that modify, transform, model, or visualize the data. [31]

There are many opportunities to process data:

- Create visual workflows with an intuitive user interface
- 2000 available nodes for precise workflow design
- Model each step of your analysis, control the flow of data and ensure your work is always up to date
- Open and combine simple text formats (CSV, PDF, XLS, XML, etc.) or unstructured data types (images, documents, molecules, etc.)
- Aggregate, sort, filter, and join data either on your local machine, in-database, or in distributed big data environments
- Visualize data with classic (bar chart, scatter plot) as well as advanced charts (heat map) and customize them to your needs [32]

4 Results and discussion

4.1 Phylogenetic tree

A phylogenetic tree shows similarities between different species. It is capable to construct a phylogenetic tree based on a hypothesis consisting on available information.

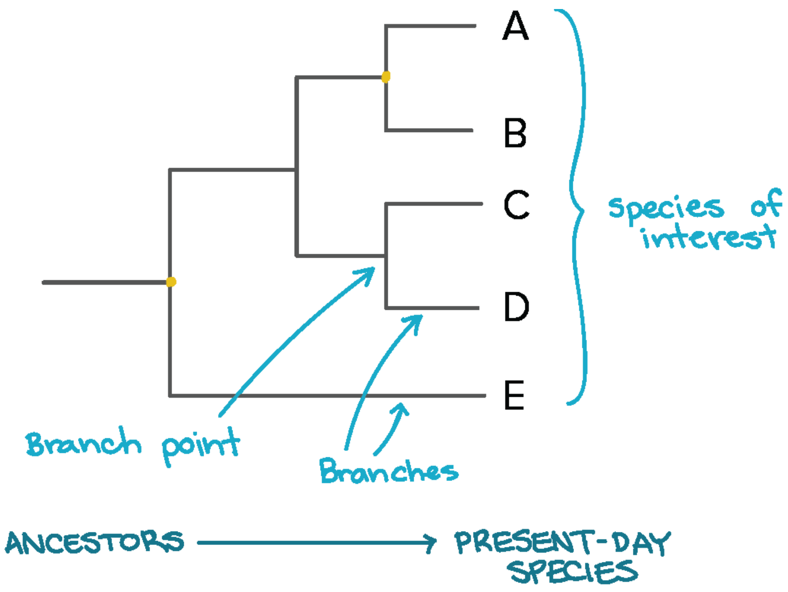


Figure 7: Organization of a phylogenetic tree [33]

The pattern of branch-connections represents how the species evolved in the tree from a series of common ancestors. Each branch point represents the division of a single group into two descendant groups. Each horizontal line in our phylogenetic tree represents a series of ancestors, leading up to the species at its end. Whether A and B are closer related than C and D cannot be determined. The aim of creating a phylogenetic tree is to compare and analyze many characteristics of species. These characteristics can include for example biochemical pathways, DNA and protein sequences. [33]

Figure 10 and Table 2 shows the relationships between the 17 gene targets. The phylogenetic tree was created with Knime. The phylogenetic tree shows the number of patents for each gene (purple), the count of the proteins that are a target for drugs (light blue), how many diseases are related to this gene (red) and the number of available activity data in ChEMBL (green).

Database	Property	Count
SureChEMBL	Patents	± 350
DrugBank	Protein as TARGET for drugs	± 2
DisGeNET	Diseases	± 1
ChEMBL	Activity values present in the ChEMBL for given protein	± 2300

Table 2: Additional information for Figure 10

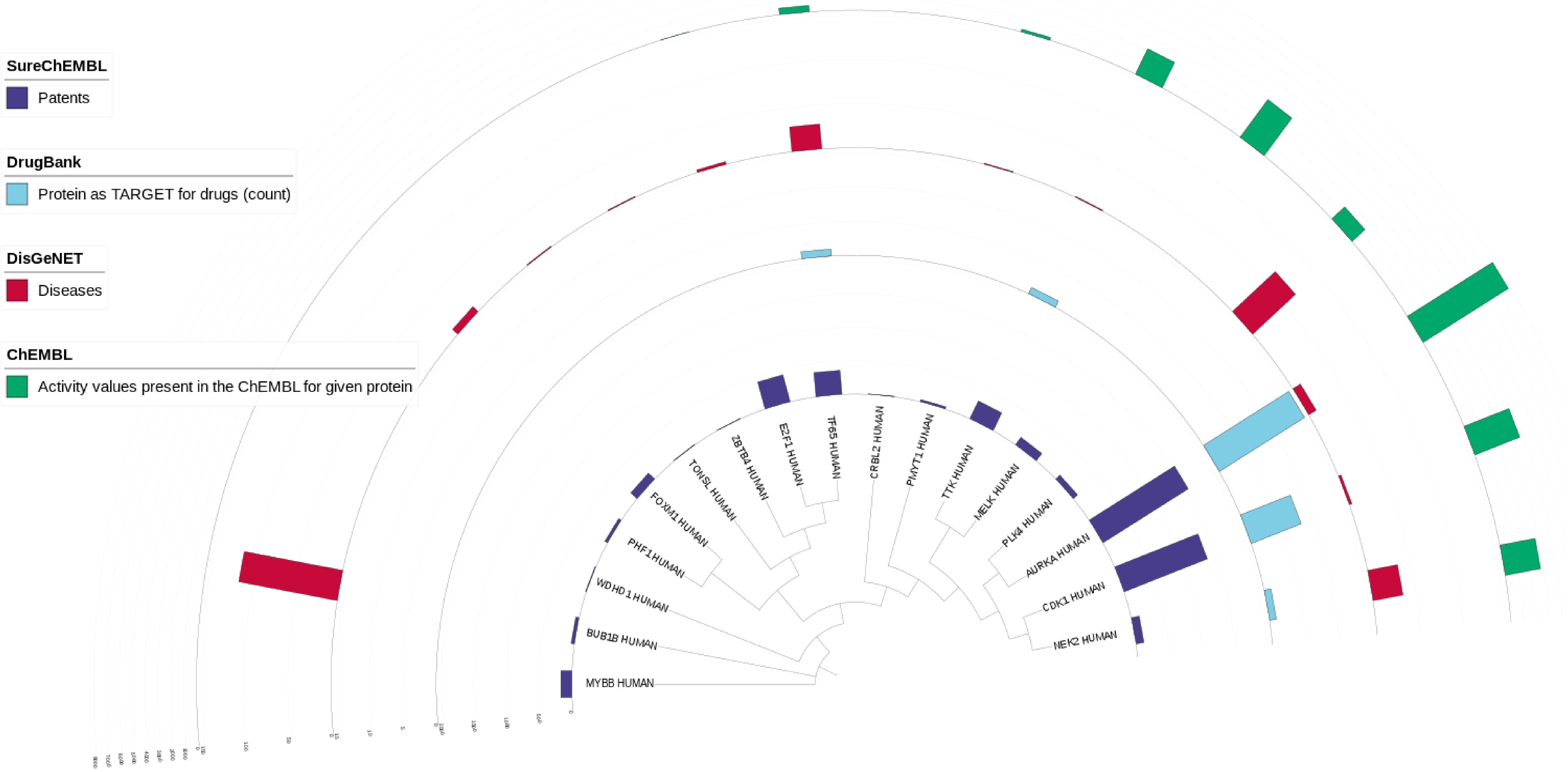


Figure 10: Phylogenetic tree of the involved genes in the multiple myeloma pathway

4.2 PDB extraction

In this project, the number of crystal structures available in the protein database for each protein of the pathway was investigated. These were automatically retrieved starting with the Uniprot ID, which is an unique identifier for every protein in the database, and using a Knime workflow available in house.

Table 3 illustrates the number of available PDB structures for each Uniprot ID according to the 17 genes of the multiple myeloma pathway (paragraph 1.2).

UNIPROT ID	PROTEIN NAME	NUMBER OF PDBs
O00444	PLK4	10
O14965	AURKA	142
O43189	PHF1	6
O60566	BUB1B	8
O75717	WDHD1	4
P06493	CDK1	5
P33981	**TTK**	**64**
P51955	NEK2	25
Q96HA7	TONSL	1
Q01094	E2F1	5
Q04206	RELA	10
Q08050	FOXM1	1
Q14680	MELK	29
Q99640	PKMYT1	9

Table 3: Number of PDBs correlated to the Uniprot ID

4.3 Target selection

The next step, after the PDB extraction, was to select one target out of the 17 genes (Figure 11).

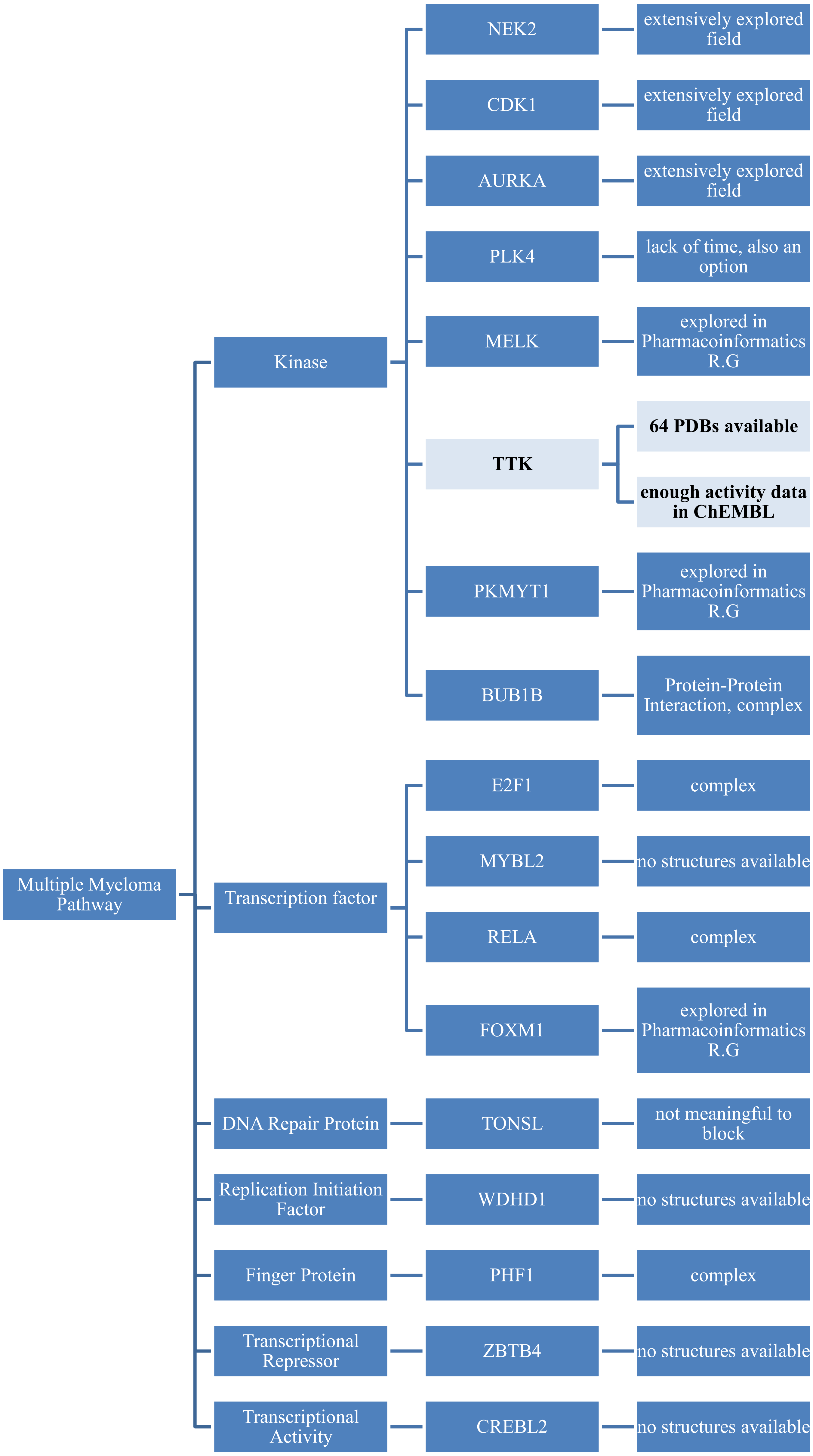


Figure 11: Target selection

4.3.1 TTK

The monopolar spindle 1 (MPS1, also known as TTK) is a kinase that regulates the biological activity of proteins. TTK is a mitotic target and plays a crucial role in the transition from metaphase to anaphase. The kinase is one of the central components of the Spindle Assembly Control Point (SAC). TTK is up-regulated in various tumor types. [34]

It represents a relevant potential target for new therapeutic agents in cancer treatment. Therefore, there are numerous small molecule inhibitors in development or in clinical trials. By inhibiting TTK kinase activity, cells leave mitosis prematurely and die. [35]

Figure 12 represents the overall structure of the human TTK catalytic domain. The N-terminal small lobe consists of a standard five-stranded β sheet (turquoise) and a αC helix (turquoise). Besides there is an additional β-strand (turquoise) at the N-terminus of the small lobe. The large lobe consists of a β sheet of two strands, namely β6 and β7 (purple) close by the small lobe, seven α helices (blue), the catalytic loop (purple) and the activation loop (yellow). Although the activation loop is partially disordered (dotted lines), the N-terminus and C-terminal P+1 (yellow + blue) loop assume well-defined conformation. [36]

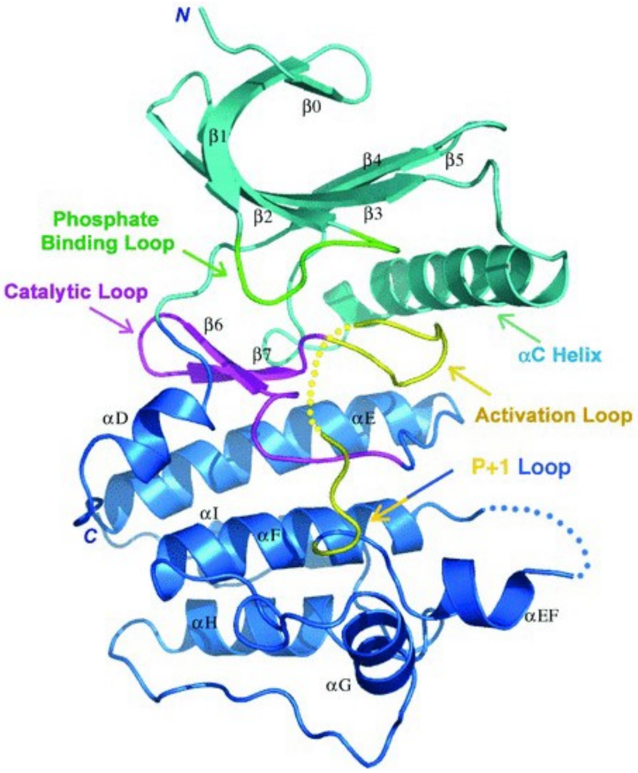


Figure 12: Structure of the human TTK [37]

4.3.2 TTK selective inhibitors

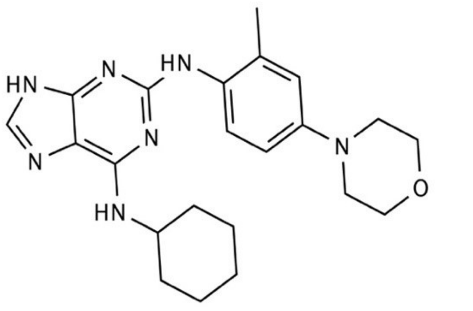
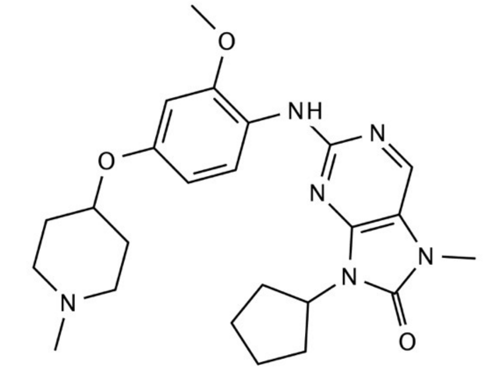
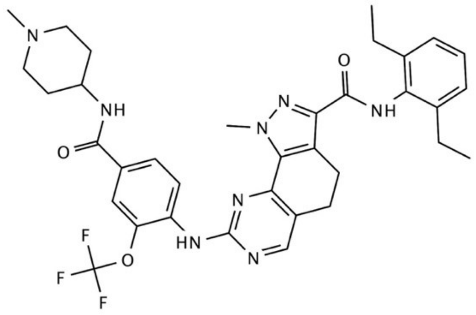
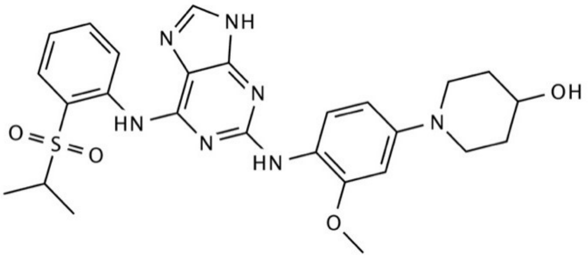
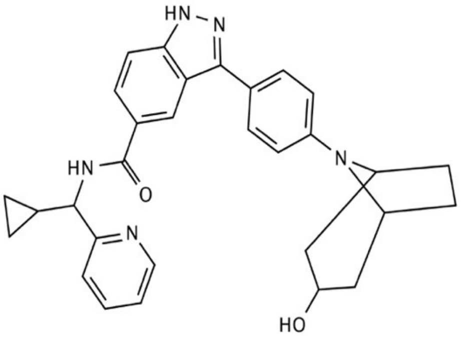
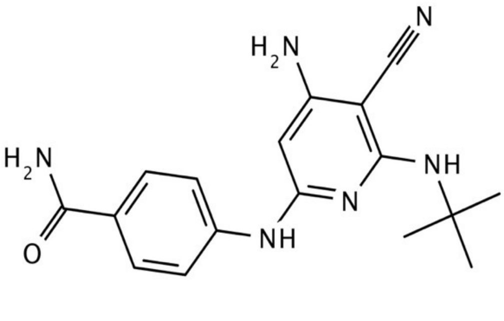
The efficacy of a kinase inhibitor is determined by its selectivity. The potency of a kinase inhibitor for its target can be measured as an IC_{50} value.

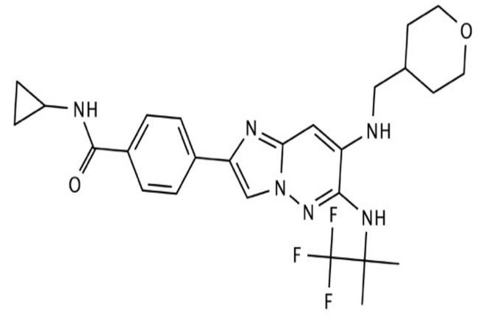
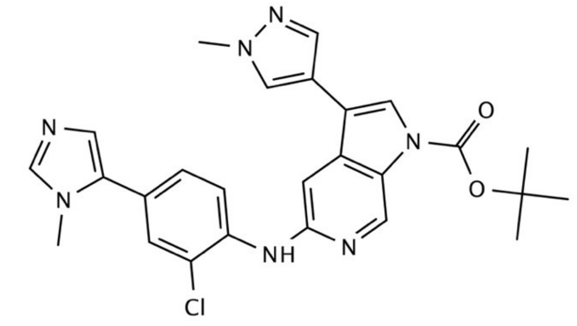
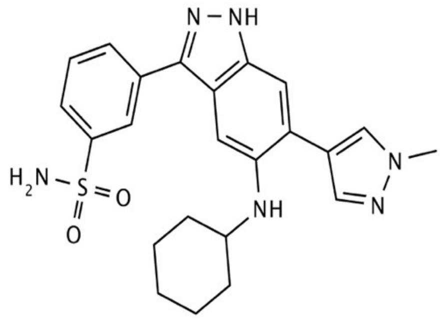
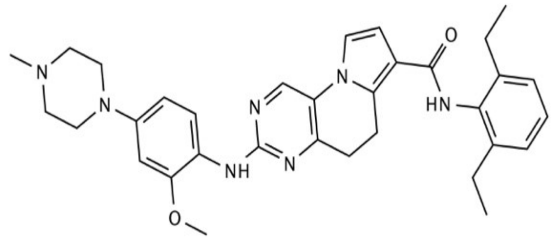
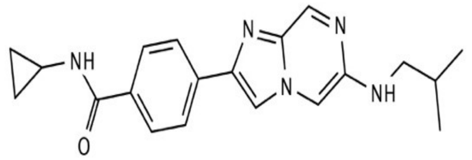
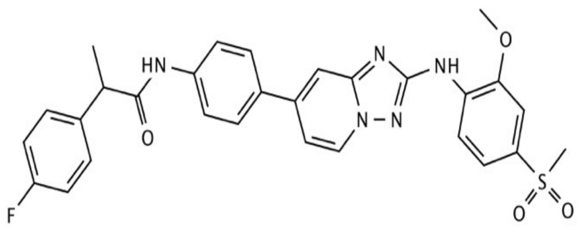
Table 4 displays the measured IC_{50} values for TTK inhibitors collected from different papers. The table also includes several IC_{50} values for other kinases, which were retrieved from the ChEMBL database. These molecules are highly selective, as their activity differs significantly as compared to other kinases.

INHIBITOR NAME	IC_{50} TTK (nM)	IC_{50} PLK1 (nM)	IC_{50} AURORA-A (nM)	IC_{50} AURORA-B (nM)	IC_{50} YES (nM)	IC_{50} PLK1 (nM)
AZ3146	5,2				34499	
MPI-0479605	1,9	>5000	>10 000	>5000		
NMS-P715	1,3	>10 000	>30 000			
MPS-IN-3	15					
CFI-401870	3,1		1300	1700		
CHEMBL2089255	6,4		>100 000			
CHEMBL3410084	2,8					
CCT251455	3		>40 000			
CHEMBL2380586	10,4					>5000
NTRC 0066-0	0,90					
MPS-BAY2B	6,7					
BAY 1161909	2,4					
BAY 1217389	1,0					

Table 4: Measured IC_{50} values of selective TTK inhibitors[38][39][40][41]

Table 5 contains thirteen selective TTK inhibitors, which are shown in two dimensional representation generated by LigandScout.

	
Myrexis compound MPI-0479605 [42]	AstraZeneca's AZ3146 [42]
	
Nerviano compound NMS-P715 [42]	MPS-IN-3 [42]
	
CFI-401870 [42]	Shionogi compound CHEMBL2089255 [42]

	
Shionogi compound CHEMBL3410084 [42]	CCT251455 [43]
	
CHEMBL2380586 [43]	NTRC 0066-0 [38]
	
Mps-BAY2b [38]	BAY 1161909 [38]

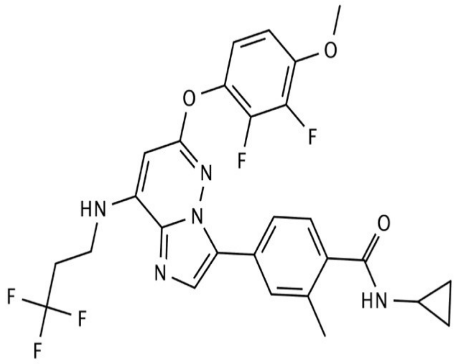
 O O F F N N HN N O HN F F F	
BAY 1217389 [27]	

Table 5: Selective Inhibitors

4.4 Crystal selection and generation of the shared pharmacophore model

The first step was to create a structure-based pharmacophore model based on crystal structures from the RCSB Protein Data Bank.

As described in 4.2, an *in house* Knime workflow was used to obtain 64 files of TTK (Uniprot ID P33981). After downloading the files, these were filtered according to the following criteria:

- Resolution < 2.5Å
- IC_{50} of the cocrystallized ligand < 10nM
- Cocrystallized with selective inhibitor

Based on such criteria, the following crystal structures were selected for structure-based pharmacophore modeling (Table 6):

- 4C4J
- 5AP0
- 5AP7
- 5N7V

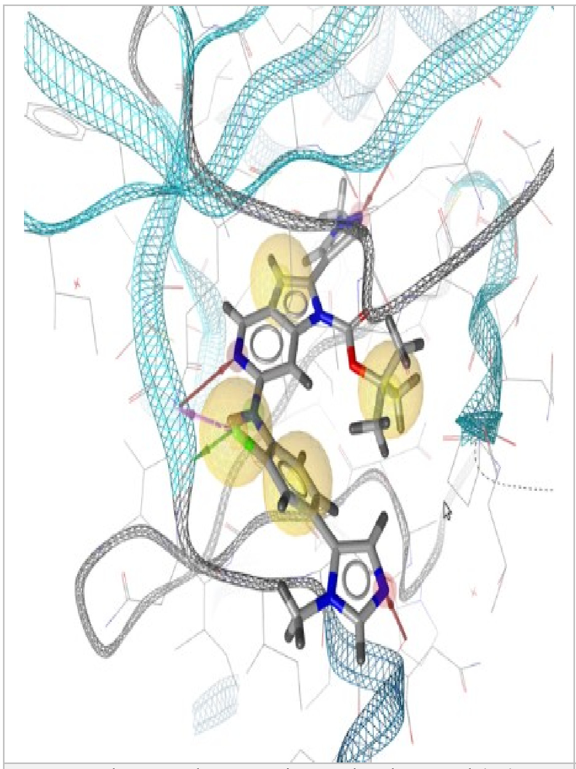
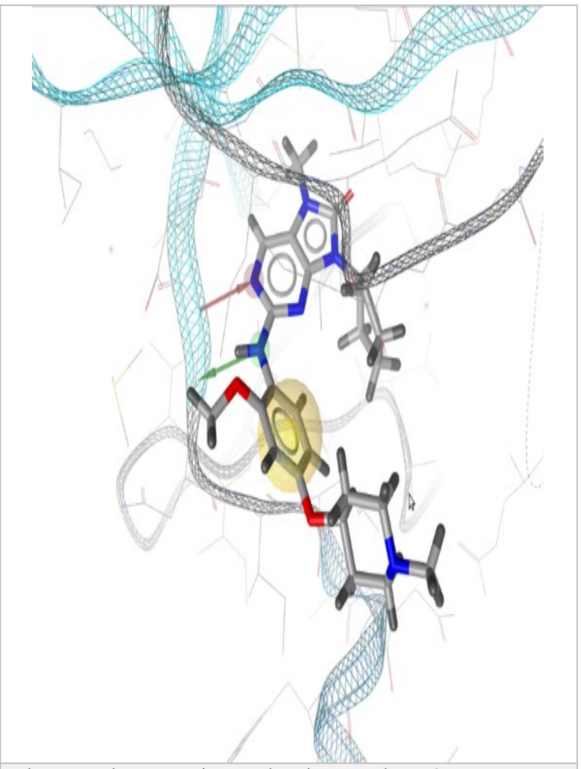
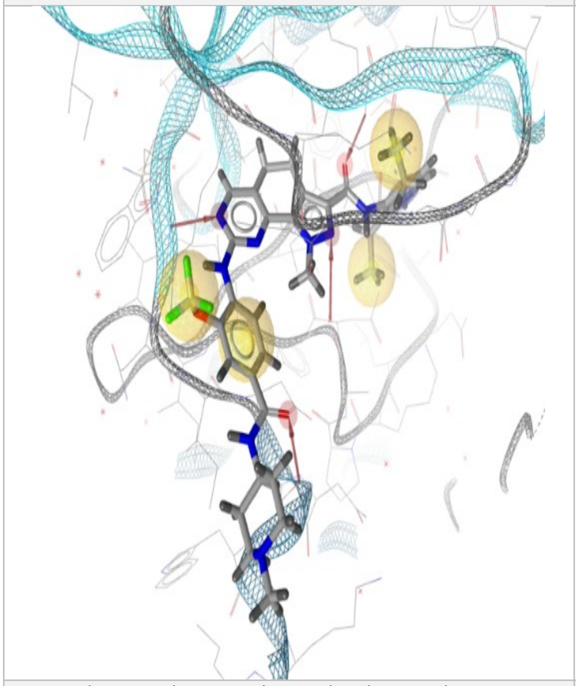
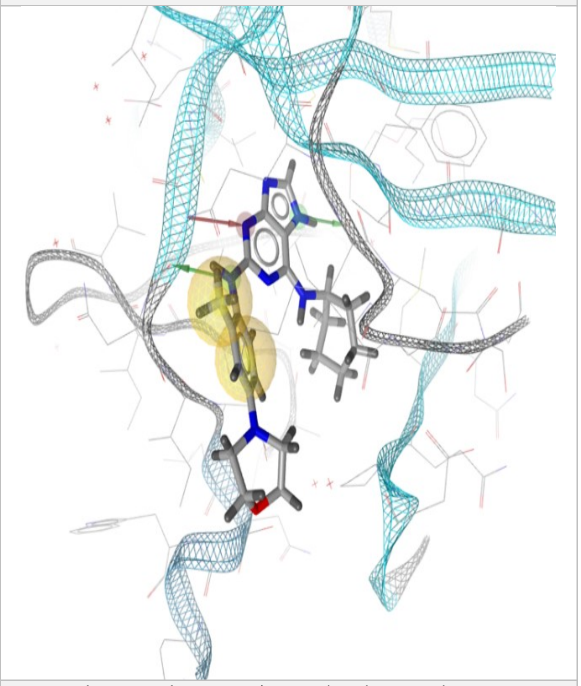
	
Pharmacophore superimposed to the crystal 4C4J	Pharmacophore superimposed to the crystal 5AP0
	
Pharmacophore superimposed to the crystal 5AP7	Pharmacophore superimposed to the crystal 5N7V

Table 6: Four pharmacophores

Using LigandScout 4.3, four three-dimensional structure-based pharmacophore models were generated in respect of the chosen crystal structures. These models were aligned by reference points (Figure 13) to create a shared feature pharmacophore.

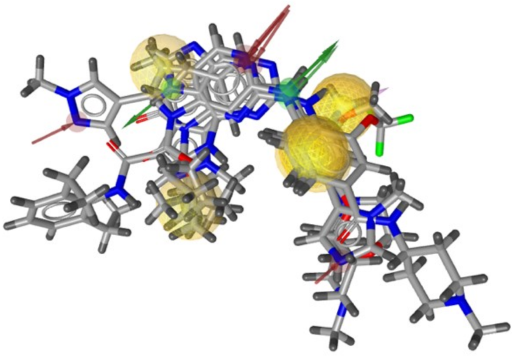


Figure 13: Alignment by reference points performed by LigandScout

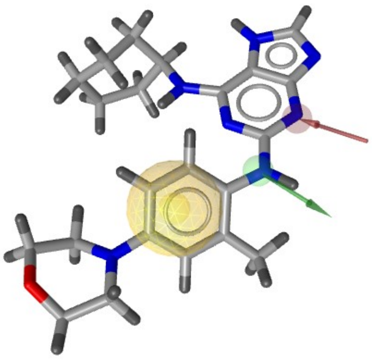


Figure 14: Minimum shared pharmacophore features with the ligand of 5N7V

During the generation of pharmacophores, LigandScout analyzes the shape of the active site and places excluded volume spheres in positions that are sterically claimed by the protein environment. This process ensures that the compounds obtained in virtual screening match the steric requirements of the active site while drastically increasing selectivity. The pharmacophore model with the minimum shared pharmacophore features (Figure 14) has been modified by adding a second hydrophobic area, a couple of hydrogen bond acceptors (one of them is optional) and another hydrogen bond donor as an optional feature. Therefore, the combination of the four pharmacophores shown in Figure 15 consists of two lipophilic areas represented as spheres (yellow), one hydrogen bond donor (green) and two hydrogen bond acceptors (red) as directional vectors. The model contains also one hydrogen bond donor (green) and one hydrogen bond acceptor (red) as optional vectors. In addition to the chemical features, 26 excluded volume spheres (grey) were added to increase the selectivity of the pharmacophore model (Figure 16).

The hydrogen bond donor (green) forms an interaction with the carboxyl group of the residue Gly 605. The hydrogen bond acceptor (red) forms an interaction with the secondary amine of the residue Gly 605. The optional hydrogen bond donor (dotted line, green) forms an interaction with the carboxyl group of the residue Glu 603. Besides there may also be interactions with the residues Lys 553 and Ser 611 if a different ligand is present.

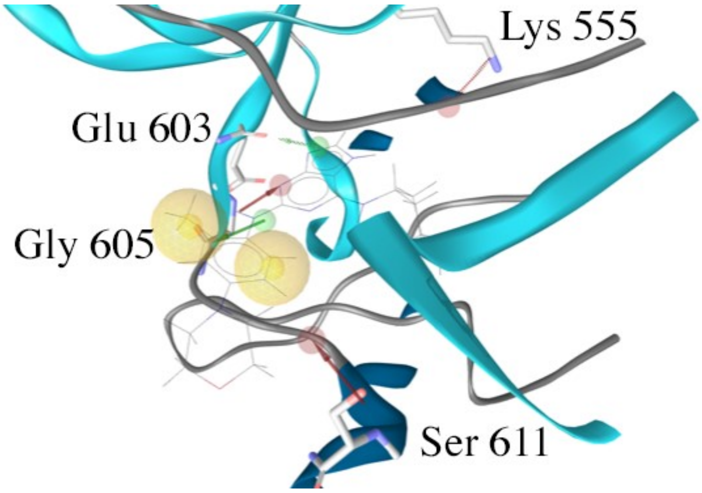


Figure 15: Generated pharmacophore (four PDBs)

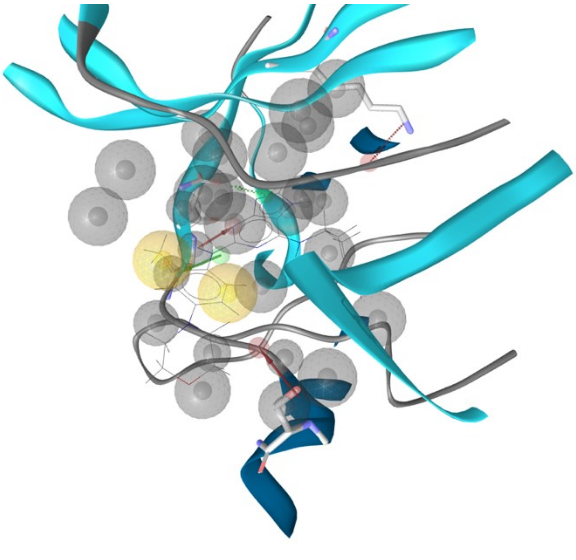


Figure 16: Generated pharmacophore with excluded volume spheres

4.5 Evaluation procedure of the pharmacophore model

4.5.1 Dataset preparation

In order to evaluate the pharmacophore model a series of screening databases were generated. Compounds tested on TTK and with a bioactivity value expressed as IC_{50} were downloaded from the ChEMBL database (ChEMBL ID: ChEMBL3983). Data points were collected in a Microsoft Excel spreadsheet. A Knime workflow (Figure 17) was used to divide the collected database between active and inactive molecules. Molecules with activity lower than 10nM were considered active, inactive when such a value was reported to be greater than 1000nM. For the screening process, a multi-conformational database was generated containing a maximum of 200 confirmations per ligand. The remaining compounds were checked for duplicates. In the end, two different screening databases were created. One dataset consisted of 477 active compounds and the other of 26 inactive compounds.

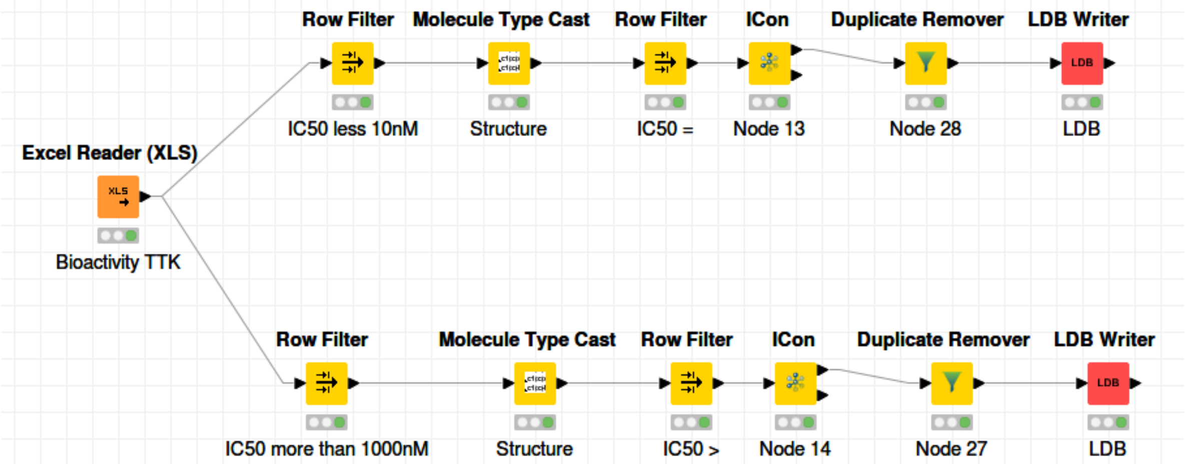


Figure 17: Knime workflow bioactivity data TTK

Figure 18 illustrates the bioactivity data downloaded from the ChEMBL database. The data can be divided into four IC_{50} groups visualizing the distribution of the compounds.

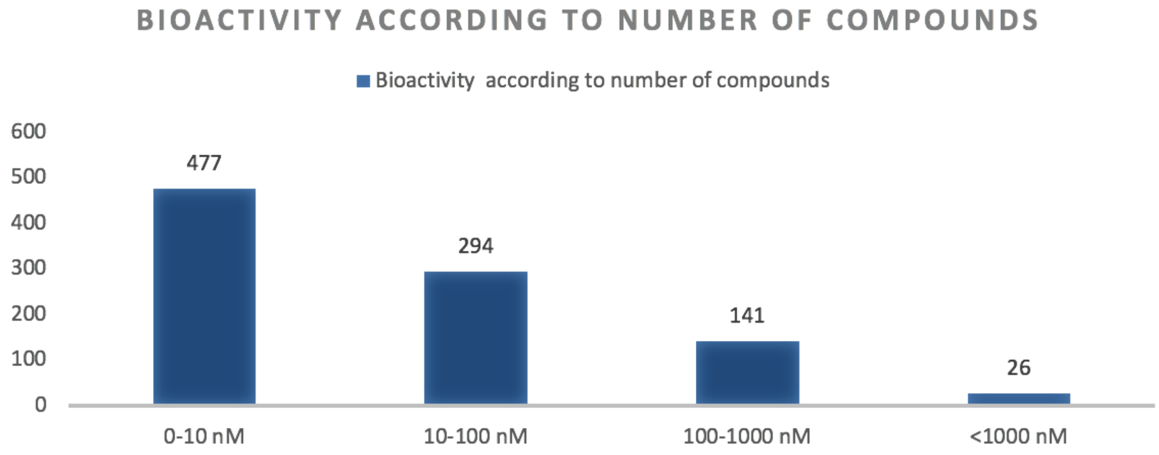


Figure 18: Distribution of bioactivity data

Considering the low number of inactive molecules retrieved, the DUD-E (Enhanced Database of Useful Decoys) database was used to obtain kinase specific decoys. Potential duplicates were eliminated using the duplicate removal node inside the aforementioned Knime workflow (Figure 19). In addition, the Icon node was inserted to generate a three-dimensional multiconformational database for virtual screening. This database contained 35.091 molecules after processing.

These decoys were added to the inactive molecules and used for the subsequent evaluation of computational models. The ratio of active/inactive-decoy compounds is above the suggested 1/40.

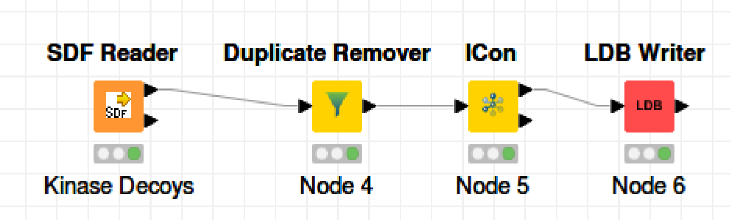


Figure 19: Knime Workflow Decoy

4.5.2 Pharmacophore screening of ChEMBL and DUD-E

The validation of the structure-based pharmacophore model included the screening of the prepared database which contained 477 active and 26 inactive compounds from ChEMBL and decoys of the DUD-E database with a total of 35.091 compounds. The 35.594 compounds were used for a Receiver Operating Characteristics (ROC) curve and the calculation of an enrichment factor. The structure-based pharmacophore model was used to screen the ChEMBL and DUD-E database. Figure 20 shows the result of the screening that led to the identification of 34 hits.

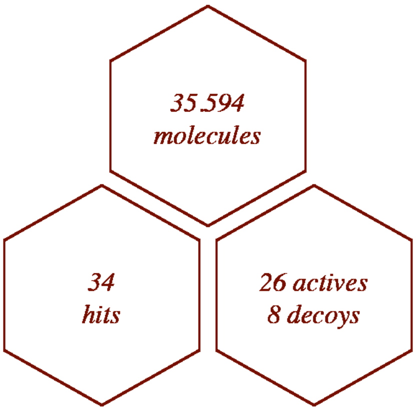


Figure 20: Screening actives / inactives / decoys

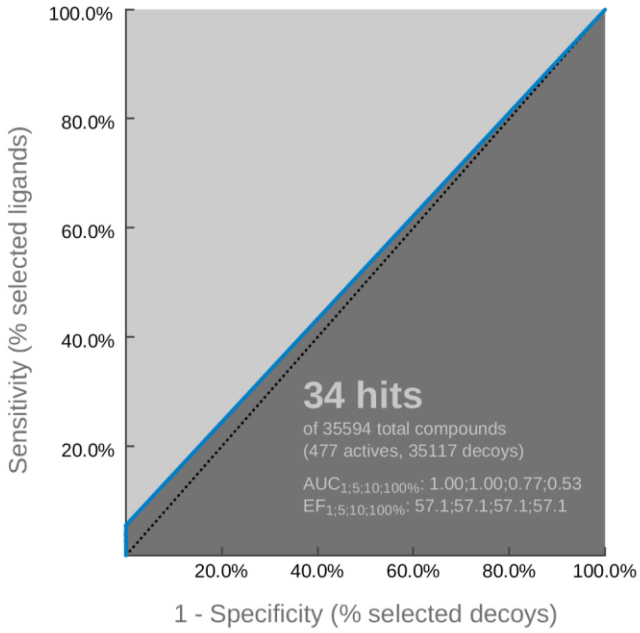


Figure 21: ROC curve validation of the three-dimensional structure-based pharmacophore model using 477 active compounds and 35.091 decoys

Figure 21 shows the ROC curve, the enrichment factor (EF) and the area under the curve (AUC). The ROC curve was used to validate the structure-based pharmacophore model. The ROC curve represents the sensitivity versus specificity. The screening of the database with the selected pharmacophore model using LigandScout generated a ROC curve that is above the diagonal of the random hits, indicating that the model is able to distinguish active from inactive compounds. The virtual screening included 477 active molecules and 35 117 decoys from the extended database of useful decoys (DUDE). The pharmacophore obtained 26 of 477 active compounds from the ChEMBL database. The area under the curve (AUC) was measured as 1.0, 1.0, 0.77 and 0.53 in the upper 1, 5, 10 and 100% of the screened database. Additionally, the enrichment factor (EF) was calculated as 57.1 in the upper 1, 5, 10 and 100% of the screened database. Altogether, the validation study has shown that the chosen pharmacophore model has a certain selectivity among the active substances and the decoys in the database.

The 26 active substances of the screening result were analyzed to determine to what extent they exhibit highly selective activity for TTK or not. Table 7 below contains data from the ChEMBL database showing that the compounds are highly selective for TTK according to IC50, whereas they are not selective for any other kinase.

CHEMBL ID	*TARGET*	*IC_{50}*	*IC_{50}*	*IC_{50}*	*IC_{50}*	*IC_{50}*	*IC_{50}*
		TTK	*AURORA-A*	*AURORA-B*	*CDK 2 / CYLCIN A*	*NEK2*	*PLK1*
2047943	PLK1, Aurora-A, TTK, Aurora-B	1.8 nM, 5 nM, 4 nM	>10000 nM	>5000 nM			>5000 nM
2047951	TTK	1 nM					
2047952	TTK	1 nM					
2047956	TTK	1 nM					
3109933	PLK1, Aurora-A, TTK, NEK2, CDK2/Cyclin A	6 nM	23000 nM		1900 nM	>10000 nM	>10000 nM
3109936	TTK, Aurora-A, CDK2/Cyclin A	7 nM	>100000 nM		5500nM		
3109938	TTK, Aurora-A, CDK2/Cyclin A	4 nM	17000 nM		680 nM		
3109939	TTK, Aurora-A, CDK2/Cyclin A	5 nM	53000 nM		750 nM		
3109972	TTK, Aurora-A, Aurora-B, CDK2/Cyclin A	8 nM	>100000 nM	>100000 nM	4200 nM		
3808471	TTK, CDK2/Cyclin A	8 nM, 10 nM					
3905145	TTK	8 nM					
3907428	TTK	3 nM					
3908898	TTK	5 nM					
3910420	TTK	10 nM					
3924132	TTK	3 nM					
3928170	TTK	1 nM					
3932702	TTK	1 nM					
3936850	TTK	4 nM					
3942820	TTK	2 nM					
3943079	TTK	6 nM					
3943432	TTK	2 nM					
3944328	TTK	5 nM					
3946383	TTK	2 nM					
3959503	TTK	10 nM					
3967693	TTK	3 nM					
3979350	TTK	4 nM					

Table 7: Bioactivity of active compound

4.6 Receptor selection for Docking

In the next step, the four chosen protein structures were prepared using the Protein Preparation Wizard of the Schrödinger suite with standard settings. All water molecules were deleted because none of them were considered as structural water. PEG and DMSO were also deleted. With the aid of the ligand preparation part of LigPrep, the chosen PDB entries as well as the hitlist of the screening were then prepared with default settings for the following docking procedure. The cocrystallized ligands of the PDB entries were docked into each receptor grid (PDB IDs: 4C4J, 5AP0, 5AP7, 5N7V) using Schrödinger Glide with default settings. This was done to determine a suitable receptor with a significant docking score and a reasonable pose of the ligands. After comparing the docking scores and poses of four PDB entries, two (5N7V and 5APO) were selected for further research. The two preselected PDB entries showed high docking values, besides the poses of the cocrystallized ligands were nicely reproduced.

Finally, the PDB entry 5N7V was selected for the following reasons:

- The structure of the protein is more complete
- Good docking scores of the ligands into the receptor
- Good reproduction of the ligand crystallographic pose

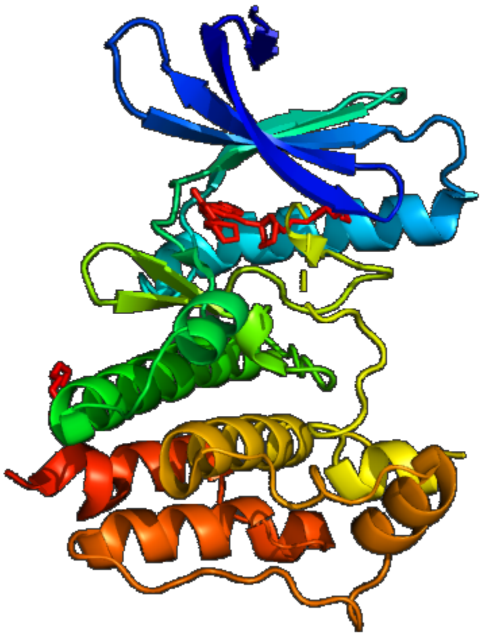


Figure 22: PDB entry 5N7V in PyMOL

Figure 22 reproduces the structure of the 5N7V receptor in complex with a ligand.

4.6.1 Docking of ChEMBL and DUD-E

Subsequentially, the dataset of molecules compiled from ChEMBL and DUD-E was docked into the selected receptor (PDB ID: 5N7V). Two docking methods were used: SP (Single Precision) and XP (Extra Precision). Then, the results of the two docking methods were compared and ultimately it was decided to use the SP, since this method produced results much faster and there were hardly any differences between the poses generated with SP and XP. The same default docking settings were applied. The objective of docking the molecules into a receptor was to get suitable scoring functions for further research.

4.7 Consensus scoring

Consensus scoring combines multiple scoring functions and strongly reduces the number of false positives. Therefore, consensus scoring can improve hit rates. It has been proven to be more robust and accurate than single scoring. The best way is to use a moderate number of scoring functions such as three or four. [44] I chose four scoring functions (Figure 23) to achieve a higher precision.

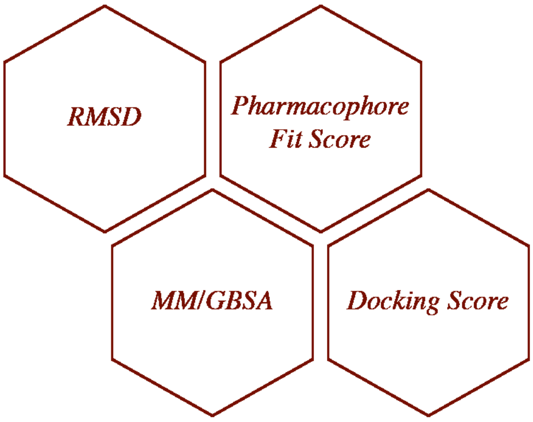


Figure 23: Applied scoring functions

4.7.1 Docking score

In pharmaceutical drug discovery, molecular docking is typically used to predict the favored orientation of two molecules to provide a stable complex. The prediction of the binding affinity between two molecules is defined by the scoring function. Docking can also be seen as a key-lock mechanism, where the correct orientation of a key can open the lock. [25] The docking score was calculated with a program called Schrödinger's Glide. Glide uses a series of hierarchical filters to search for possible positions of the ligand in the active region of the receptor. The preparation procedure is important to achieve precise docking with Glide. [45]

4.7.2 Pharmacophore fit score

The pharmacophore fit score measures the geometric fit of the features of a molecule to the three-dimensional structure-based pharmacophore model. A higher fit score indicates a better fit to the pharmacophore model. A lower pharmacophore fit score indicates that the features cannot be matched. [46] LigandScout was used to calculate the pharmacophore fit score.

4.7.3 RMSD

The RMSD is the most commonly used parameter to measure the similarity between two superimposed compounds. For this reason a Knime workflow that calculates the RMSD of two different ligand poses was created. The compounds submitted to the workflow were the ligand poses resulting from the pharmacophore screening with LigandScout, and the docked poses of Maestro. Only heavy atoms were considered for the “inplace” RMS alignment inside the RMSD Schrödinger Knime node. Figure 24 shows the implemented workflow. First the SDF reader node loads molecules from sdf files for further processing. The Sorter node sorts the lines by user-defined criteria.

Afterwards the Molecule-to-MAE node converts Smiles, Mol2, SD or PDB to Maestro. The Chunking Loop Start node is the beginning of the loop, where you can define the frequency of the loop execution. Next, the RMSD node calculates the RMSD between the ligand pose resulting from the pharmacophore screening with LigandScout, and the docked pose of Maestro. The loop end node is used to indicate the end of a workflow and collects the output by linking the incoming tables line by line. [47] The output of the implemented Knime workflow was checked for accuracy with PyMOL.

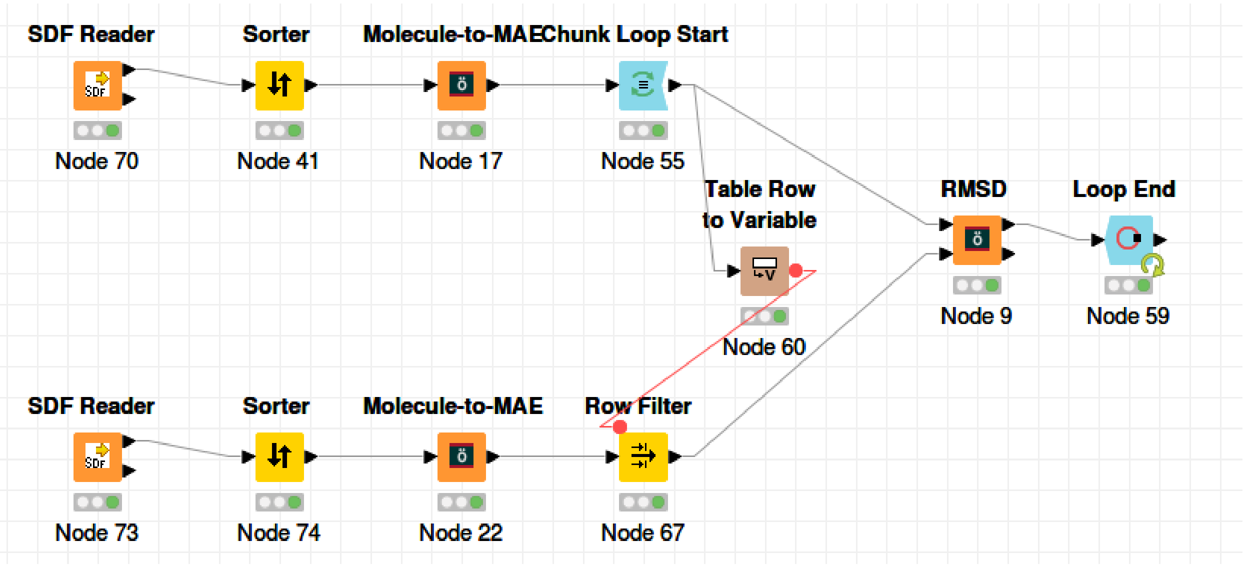


Figure 24: RMSD calculation

4.7.4 MM/GBSA

The MM/GBSA approach (Molecular Mechanics Generalized Born Surface Area) is a popular way to estimate the free energy of the binding of small ligands to biological macromolecules. This method improves the result of virtual screening and docking and calculates the binding-free energy of a ligand within a protein. [48] The standard settings in Maestro were used for the calculation.

4.8 Rankings for the ChEMBL and DUD-E dataset

The following tables show the results of the four applied scoring functions from the ChEMBL and DUD-E database. The vales are NOT sorted according to the score.

ID	TITLE	MMGBSA DG BIND	ACTIVE/DECOY
34	Row10855	-36,67	decoy
28	Row117	-47,86	active
17	Row120	-59,62	active
33	Row12355	-37,97	decoy
3	Row133	-71,96	active
11	Row176	-65,17	active
32	Row19338	-40,27	decoy
27	Row20063	-48,54	decoy
29	Row20300	-45,79	decoy
21	Row204	-55,59	active
16	Row213	-61,48	active
5	Row217	-69,74	active
15	Row24	-61,69	active
30	Row2409	-44,96	decoy
2	Row250	-74,82	active
24	Row254	-54,43	active
26	Row27	-49,12	active
14	Row3210	-61,9	decoy
6	Row337	-68,64	active
7	Row35	-68,26	active
10	Row386	-65,21	active
18	Row401	-58,52	active
22	Row402	-55,4	active
23	Row409	-54,95	active
13	Row419	-62,49	active
4	Row435	-69,77	active
31	Row462	-44,02	active
20	Row497	-56,19	active
19	Row529	-57,75	decoy
8	Row62	-66,82	active
9	Row65	-65,43	active
25	Row78	-51,11	active
12	Row8	-63,31	active
1	Row86	-75,61	active

Table 8: MM/GBSA calculation

ID	TITLE	DOCKING SCORE	ACTIVE/DECOY
26	Row10855	-7,58	decoy
29	Row117	-6,93	active
14	Row120	-8,446	active
27	Row12355	-7,526	decoy
2	Row133	-9,929	active
7	Row176	-9,696	active
24	Row19338	-7,62	decoy
20	Row20063	-8,148	decoy
23	Row20300	-7,8	decoy
33	Row204	-6,297	active
11	Row213	-9,052	active
9	Row217	-9,52	active
22	Row24	-7,948	active
30	Row2409	-6,757	decoy
5	Row250	-9,812	active
18	Row254	-8,203	active
34	Row27	-5,809	active
28	Row3210	-7,093	decoy
6	Row337	-9,734	active
10	Row35	-9,151	active
4	Row386	-9,813	active
17	Row401	-8,224	active
19	Row402	-8,161	active
13	Row409	-8,792	active
16	Row419	-8,259	active
1	Row435	-10,013	active
31	Row462	-6,51	active
25	Row497	-7,614	active
21	Row529	-8,14	decoy
8	Row62	-9,645	active
12	Row65	-8,936	active
32	Row78	-6,346	active
15	Row8	-8,401	active
3	Row86	-9,913	active

Table 9: Docking score calculation

ID	TITLE	PHARMACOPHORE SCORE	ACTIVE/DECOY
19	Row10855	56,21	**decoy**
6	Row117	65,95	active
1	Row120	66,14	active
20	Row12355	56,08	**decoy**
4	Row133	66	active
18	Row176	56,31	active
25	Row19338	56,04	**decoy**
13	Row20063	64,7	**decoy**
21	Row20300	56,08	**decoy**
11	Row204	65,46	active
31	Row213	55,34	active
15	Row217	56,64	active
12	Row24	65	active
34	Row2409	54,63	**decoy**
33	Row250	54,63	active
17	Row254	56,52	active
3	Row27	66,02	active
24	Row3210	56,04	**decoy**
29	Row337	55,57	active
32	Row35	54,81	active
10	Row386	65,7	active
30	Row401	55,38	active
9	Row402	65,75	active
14	Row409	56,8	active
2	Row419	66,14	active
28	Row435	55,61	active
7	Row462	65,92	active
5	Row497	65,99	active
23	Row529	56,04	**decoy**
27	Row62	55,65	active
26	Row65	55,97	active
8	Row78	65,78	active
16	Row8	56,55	active
22	Row86	56,06	active

Table 10: Pharmacophore score calculation

ID	TITLE	RMSD	ACTIVE/DECOY
28	Row10855	8,25	**decoy**
21	Row117	4,17	active
23	Row120	4,77	active
32	Row12355	11,75	**decoy**
2	Row133	1,35	active
14	Row176	2,06	active
34	Row19338	13,06	**decoy**
27	Row20063	7,87	**decoy**
30	Row20300	9,65	**decoy**
24	Row204	5,28	active
8	Row213	1,58	active
6	Row217	1,54	active
12	Row24	1,85	active
26	Row2409	7,31	**decoy**
7	Row250	1,56	active
25	Row254	5,55	active
20	Row27	3,83	active
31	Row3210	10,46	**decoy**
16	Row337	2,64	active
9	Row35	1,58	active
5	Row386	1,53	active
17	Row401	2,65	active
22	Row402	4,76	active
11	Row409	1,59	active
18	Row419	3,07	active
4	Row435	1,47	active
19	Row462	3,67	active
15	Row497	2,26	active
29	Row529	9,39	**decoy**
3	Row62	1,45	active
1	Row65	1,34	active
33	Row78	11,8	active
13	Row8	1,97	active
10	Row86	1,58	active

Table 11: RMSD calculation

4.9 RBR for the screening database (ChEMBL and DUD-E)

With all the compounds having four different scores (Docking Score, MM/GBSA, Pharmacophore Fit Score and RMSD), a Rank by Rank resorting was applied to improve the individual results. In the end, all compounds were ranked according to the average rank among all four scoring functions.

In the following table below you can see the combination of the four rankings (fictive values) into a Rank by Rank list.

PHARMACOPHORE SCORE	DOCKING RANK	MM-GBSA RANK	RMSD RANK	R.B.R
MOL A: 8	MOL A: 3	MOL A: 10	MOL A: 5	MOL A: 6,5
MOL B: 3	MOL B: 5	MOL B: 7	MOL B: 8	MOL B: 5,75
MOL C: 2	MOL C: 5	MOL C: 8	MOL C: 12	MOL C: 6,75

Table 12: Explanation of the RBR using an example

After several combinations the RBR with four scoring functions was the best one, rank 1-23 were only active compounds. The first decoy appears at rank 24. The table below demonstrates that active substances have a higher rank than decoys. Only 3 out of 23 active substances are ranked lower than the active compounds above. There were no inactive compounds retrieved out of the ChEMBL database.

RBR 4	TITLE	ACTIVE/DECOY
2,75	Row133	active
7,25	Row386	active
8,75	Row217	active
9	Row86	active
9,25	Row435	active
11,5	Row62	active
11,75	Row250	active
12	Row65	active
12,25	Row419	active
12,5	Row176	active
13,75	Row120	active

14	Row8	active
14,25	Row337	active
14,5	Row35	active
15,25	Row24	active
15,25	Row409	active
16,25	Row497	active
16,5	Row213	active
18	Row402	active
20,5	Row401	active
20,75	Row27	active
21	Row117	active
21	Row254	active
21,75	Row20063	**decoy**
22	Row462	active
22,25	Row204	active
23	Row529	**decoy**
24,25	Row3210	**decoy**
24,5	Row78	active
25,75	Row20300	**decoy**
26,75	Row10855	**decoy**
28	Row12355	**decoy**
28,75	Row19338	**decoy**
30	Row2409	**decoy**

Table 13: Rank by Rank including four scoring functions

4.10 Vendors database

In Knime, a database of commercially available compounds from several vendors has been created with a special focus on libraries for kinases. All the molecules were standardized and then converted into three-dimensional conformational chemical compounds using the advanced algorithm *idbgen* of LigandScout 4.3. The prepared library contains 87151 unique molecules from eight different suppliers. (Figure 25)

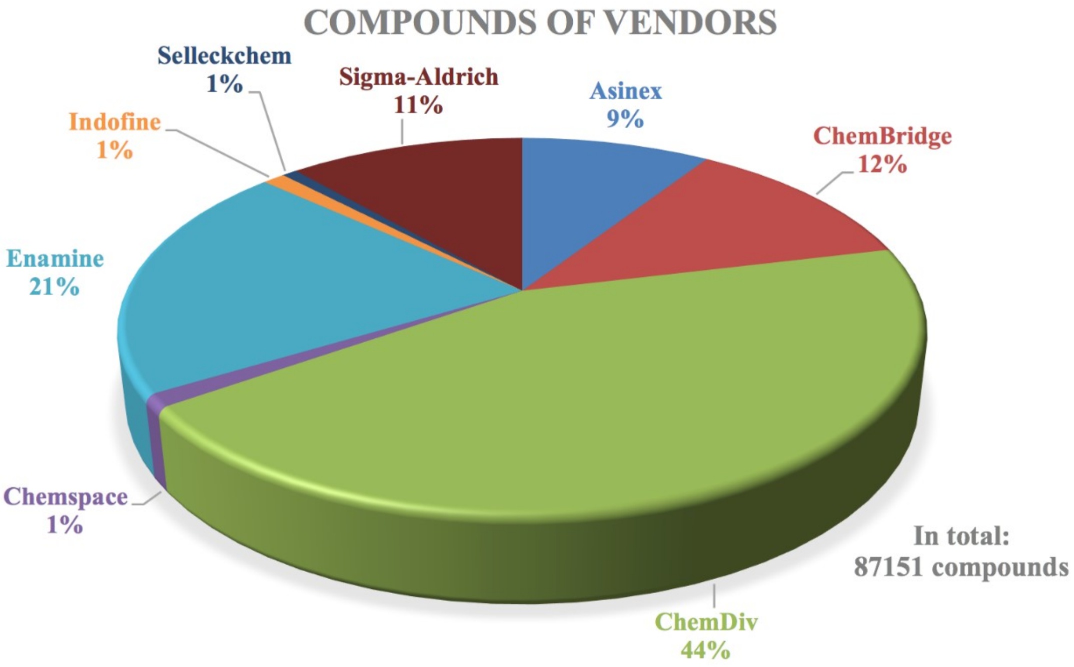


Figure 25: Percentage composition of vendors

Table 14 shows the used settings to standardize the vendors database with the help of an in house Knime workflow.

SETTINGS	VALUE
KEEP ALL STRUCTURES	No
FILTER BY WEIGHT	No
CLEAR STEREO?	No
IF KEEP STEREO: STANDARDIZE STEREO?	No
KEEP SALTS OR SOLVENT MOLECULES?	No
MOLECULAR WEIGHT	1000
KEEP MIXTURES OF MOLECULES?	No
KEEP MOLECULES WITH NONORGANIC ATOMS?	No

Table 14: Standardizer for vendors database

4.11 Pharmacophore model modification and screening

Too few hits were found in the screening of the vendors database, therefore it was necessary to refine the generated pharmacophore model. Therefore the h-bond acceptor was set to optional. Figure 26 shows the final structure-based pharmacophore model.

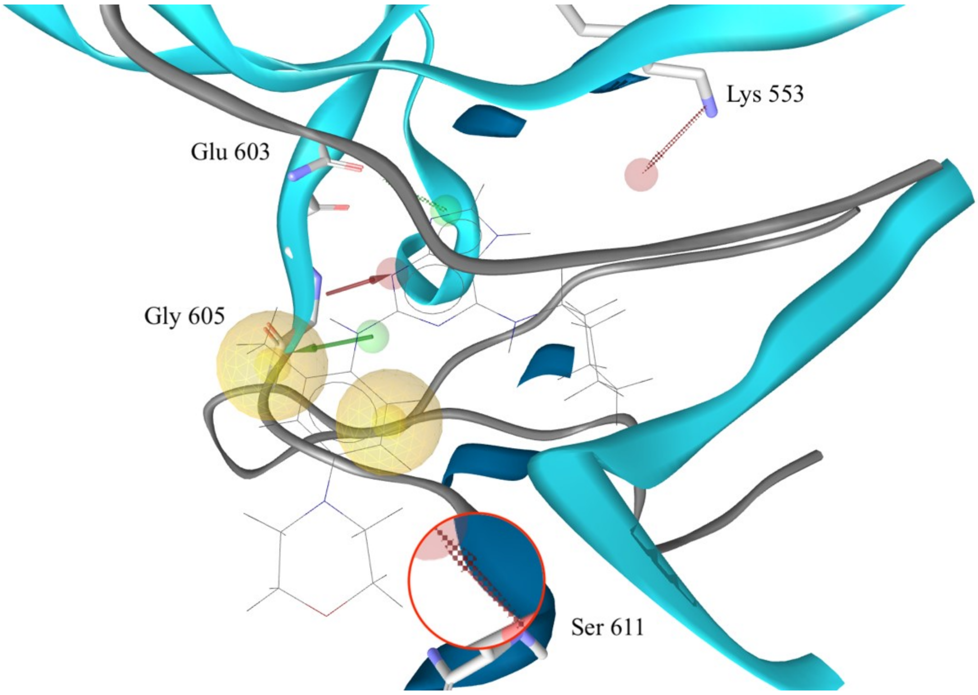


Figure 26: Refined pharmacophore superimposed to the crystal 5N7V

The pharmacophore consisted of two lipophilic areas (yellow), one hydrogen bond donor (green) and one hydrogen bond acceptor (red). The model also consisted of one hydrogen bond donor and two hydrogen bond acceptors as optional vectors (dotted lines). The feature marked in red shows the refinement of the pharmacophore, which led to significantly more hits because it is now marked as an optional feature. The refined pharmacophore model was used to screen the vendors library. Figure 27 shows screening results of the vendors database.

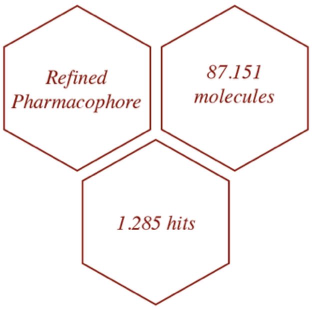


Figure 27: Screening results of Vendors database

The modified pharmacophore model was used to find molecules with the required features from the vendors database. The screening resulted in 1285 hits from 87151 commercially available compounds (1,47%). The analysis showed that the pharmacophore is still highly selective, although the optional hydrogen bond acceptor has been modified.

4.12 Ranking the Vendors database

The docking software Glide was again used to dock the 1.285 hits out of the vendors database into the 5N7V receptor. Due to the large amounts of data, not all scoring functions can be listed in this thesis.

After calculating all scores using the scoring functions, the top 100 highest RBR ranked compounds of the vendors database where selected (see Annex B). Figure 24 shows the implemented workflow for calculating the RMSD between the offered compounds of the vendors in LigandScout and Maestro.

The output of the implemented Knime workflow was checked for accuracy with PyMOL.

4.13 Binning clustering with ChemMine

Clustering is a widely used technique to divide a large hitlist of compounds into small groups with high similarity and is very useful on large datasets. As a result, I used the 100 highest ranked compounds from Rank by Rank and clustered them using ChemMine.

Binning clustering was performed with a Tanimoto coefficient of 0.6. The compounds with a similar or greater value were clustered into groups, using a single linkage rule for cluster joining. This gave as a result 49 different clusters for the 100 compounds.

A representative compound with the best rank by rank was chosen for each cluster. As an example, the clusters in Figure 28 and Figure 29 (e.g. cluster 1 and 2) show the similarity of the compounds in each cluster. Cluster 1 and cluster 2 show that the chemical structure of the compounds per cluster are very homogeneous. With the help of ChemMine, a meaningful clustering was implemented. Binning clustering has made it easier to select suitable compounds.

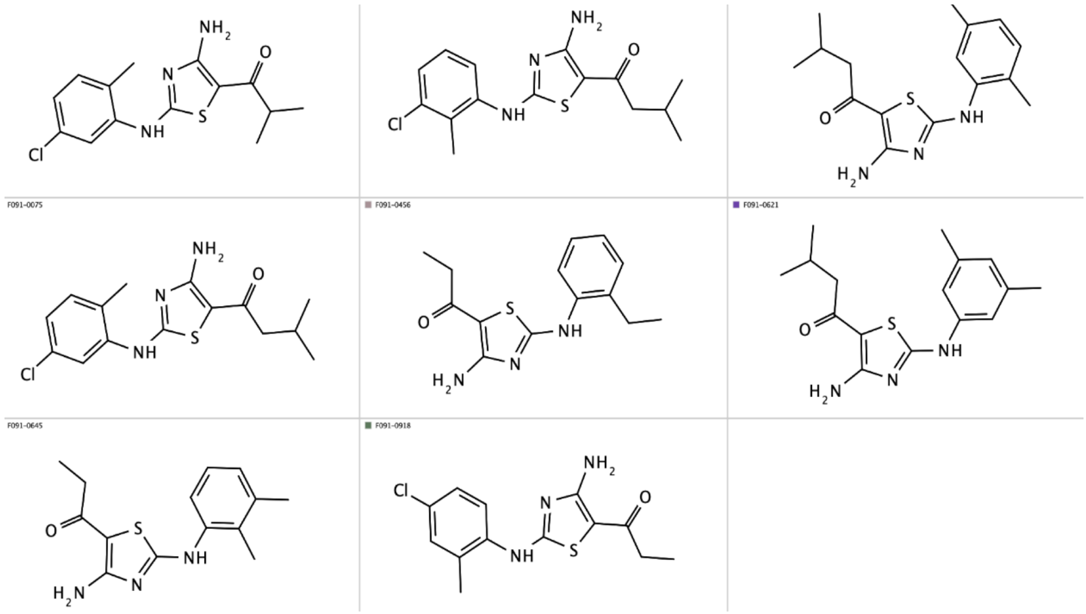


Figure 28: Cluster 1

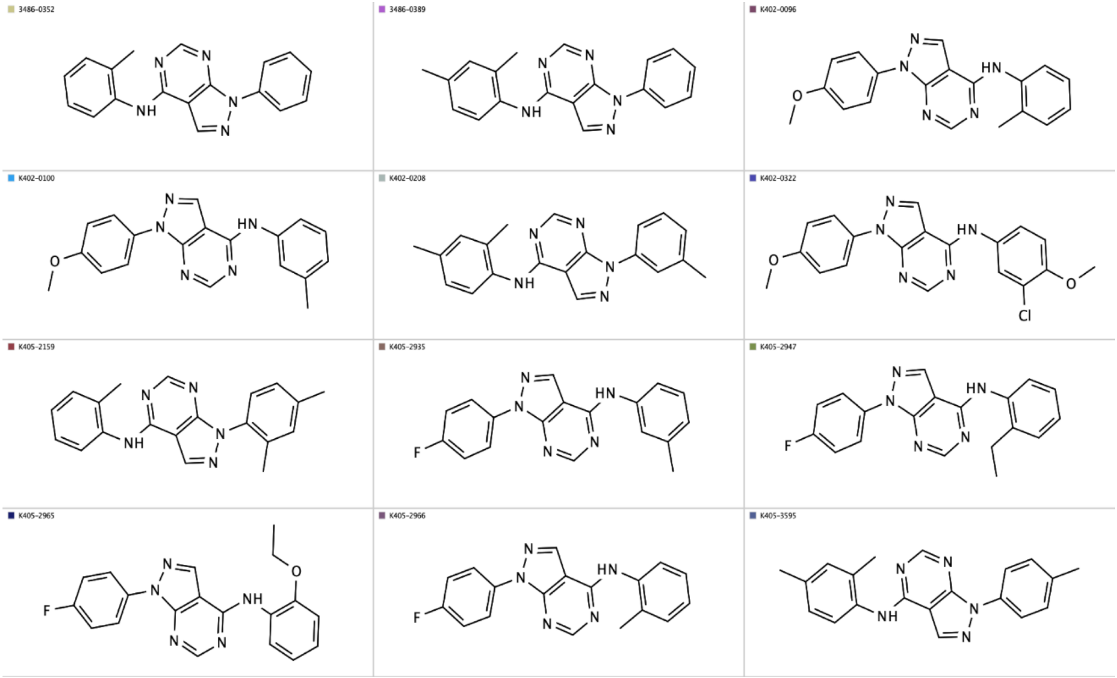


Figure 29: Cluster 2

4.14 Visual inspection

A representative compound for each cluster, which I obtained from the binning clustering with ChemMine, was selected. The compounds with the highest rank in each group were visually inspected. As a criteria, both the ligand pose as well as the interactions from the docking experiment (Glide), should have shown strong similarity with the poses retrieved with the pharmacophore screening (LigandScout).

The examples below show, that the poses and interactions of the compounds are similar in both LigandScout and Maestro, increasing the confidence in the procedure.

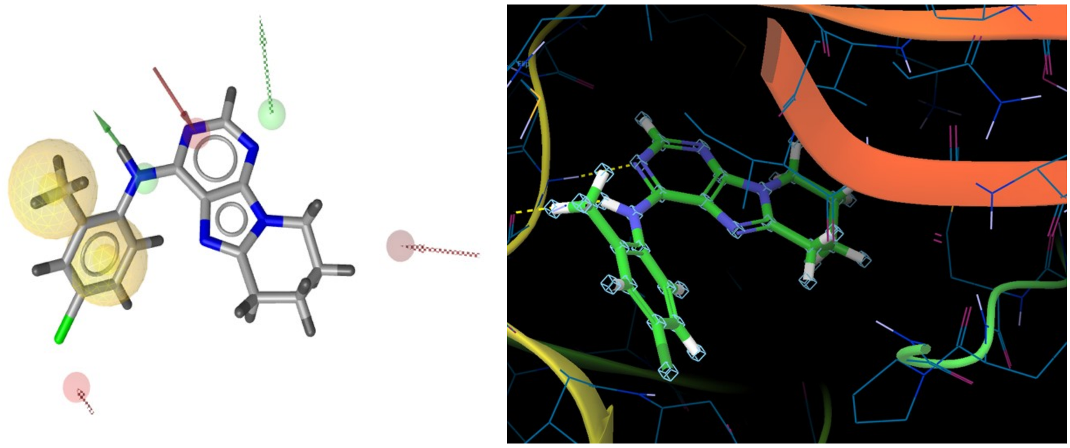


Figure 30: Comparison of compound E715-0971 in LigandScout (left) and Maestro (right)

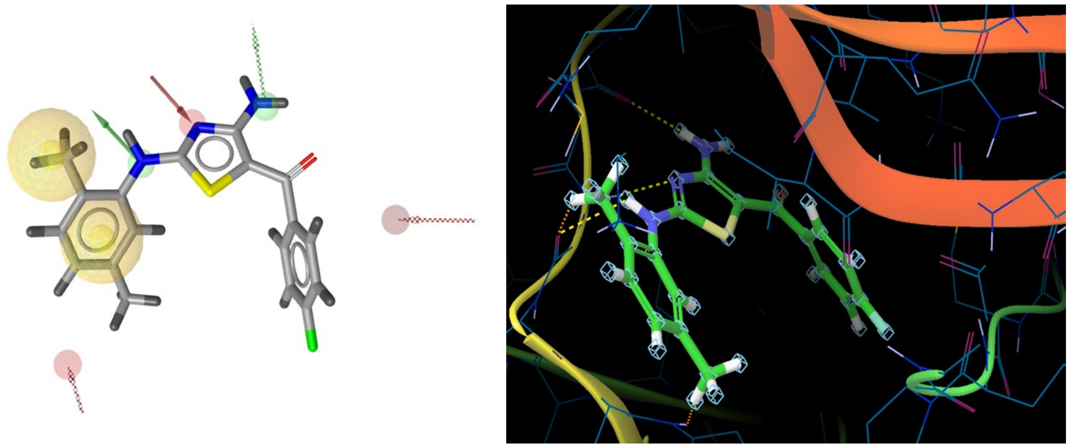


Figure 31: Comparison of compound F091-1168 in LigandScout (left) and Maestro (right)

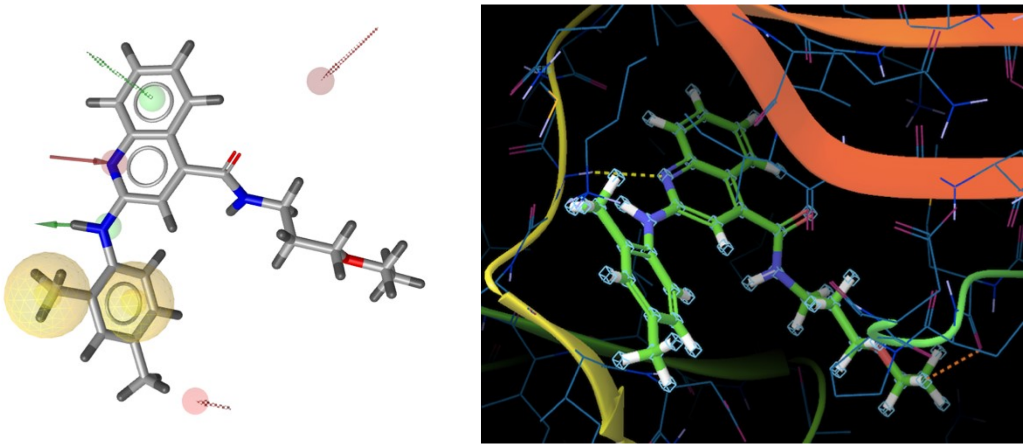


Figure 32: Comparison of compound C387-1105 in LigandScout (left) and Maestro (right)

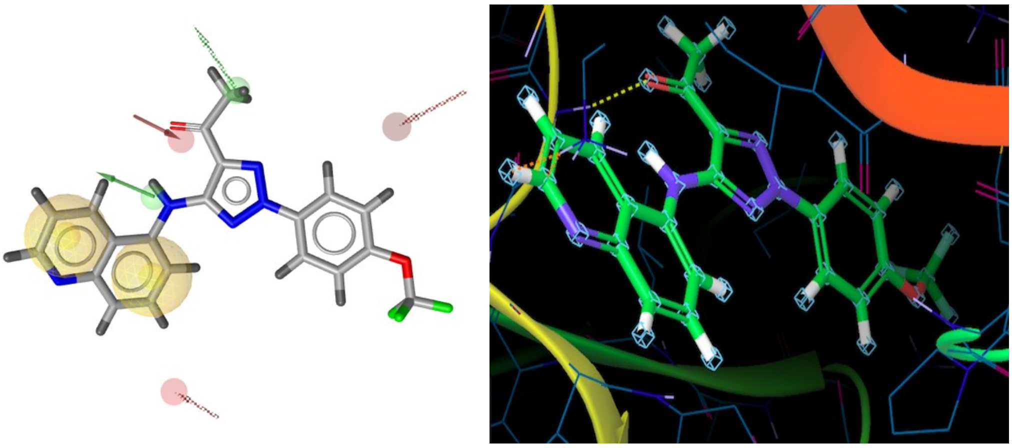


Figure 33: Comparison of compound D153-0152 in LigandScout (left) and Maestro (right)

4.15 Solubility and permeability predictions

The compounds resulting from the vendors library screening were evaluated for their drug-like behavior using Schrödinger QikProp, since nearly 40% of drug candidates in clinical trials fail due to poor ADME (absorption, distribution, metabolism, and excretion) properties. Standard ADME properties were used. Special attention was paid to the QPlogS as well as to the QPPCaco property. The QPlogS descriptor predicts the water solubility, therefore the solubility of a compound. The recommended range should be between – 6.5 to + 0.5. [49] The results are within an acceptable range, four values deviate very slightly.

The QPPCaco descriptor predicts the apparent Caco-2 cell permeability in nm/sec of a compound. Caco-2 cells are a model for the gut-blood barrier, so they directly reflect the ability of a molecule to permeate the gut-blood barrier. Compounds display good permeability if values > 500 are reached. If the values are < 25, this suggests that the cell permeability is poor. [49] 20 of 23 compounds had QPPCaco values of over 500, which means that cell permeability is excellent. The three compounds with a value below 500 are still in a good range and there should be no problem with cell permeability.

NAME LIGANDSCOUT	PRODUCT ID VENDOR	QPLOGS
C387-0954	C387-0954	-6,621
C387-1105	C387-1105	-6,592
D042-0041	D042-0041	-4,413
D153-0152	D153-0152	-6,516
D367-0335	D367-0335	-4,825
D463-0064	D463-0064	-5,917
D727-0702	D727-0702	-4,996
E715-0971	E715-0971	-5,198
F047-0446	F047-0446	-5,905
F091-0077	F091-0077	-4,685
F091-1168	F091-1168	-5,215
F532-0990	F532-0990	-4,382
J013-1406	J013-1406	-7,116
K405-2966	K405-2966	-5,516
L656-0064	L656-0064	-5,862
M008-8339	M008-8339	-5,267
ROW13081	Z990847972	-5,477
ROW13258	Z909612044	-3,565
ROW13542	Z729071454	-4,903
ROW15273	Z1647890721	-3,666
ROW3845	Z1611173508	-3,772
ROW4557	Z101408880	-3,602
ROW8325	Z1173566458	-4,274

Table 15: QplogS property of final compounds

NAME LIGANDSCOUT	PRODUCT ID VENDOR	QPPCACO
C387-0954	C387-0954	1968,714
C387-1105	C387-1105	2573,034
D042-0041	D042-0041	452,22
D153-0152	D153-0152	1612,694
D367-0335	D367-0335	1286,276
D463-0064	D463-0064	4872,783
D727-0702	D727-0702	1411,799
E715-0971	E715-0971	3005,963
F047-0446	F047-0446	1739,637
F091-0077	F091-0077	797,744
F091-1168	F091-1168	595,014
F532-0990	F532-0990	404,987
J013-1406	J013-1406	964,967
K405-2966	K405-2966	2724,083
L656-0064	L656-0064	3072,677
M008-8339	M008-8339	1868,416
ROW13081	Z990847972	1306,512
ROW13258	Z909612044	852,014
ROW13542	Z729071454	2869,197
ROW15273	Z1647890721	380,46
ROW3845	Z1611173508	3423,858
ROW4557	Z101408880	1128,939
ROW8325	Z1173566458	814,977

Table 16: QPPCaco property of final compounds

4.16 Chemical diversity

It is often worthless to test many compounds with high similarity, as similar compounds generally have similar bioactivity. For this reason, MACCS fingerprints were calculated and the Chart View node in Knime (Figure 34) was used to graphically illustrate the diversity of the substances.

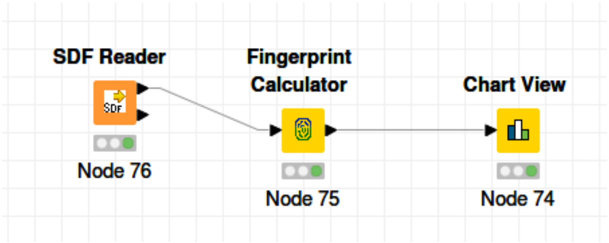


Figure 34: Diversity check executed by Knime

Heatmaps are excellent for visualizing large amounts of datasets and are used to highlight compounds with similar structures because they are displayed as areas of similar color.
The heatmap in Figure 37 shows the structural diversity of the 23 selected compounds using MACCS fingerprints and the chart view node in Knime. The graphical representation of the result was according to the requirements and there were only a few compounds that are slightly similar, e.g. C387-1105 and C387-0954. (Figure 35) Furthermore the chemical structures of L656-0064 and F047-0446 (Figure 36) appeared to be slightly similar, as you can see here in the heatmap. The preset color gradient sets the lowest value in the heatmap to dark blue, the highest value and thus the highest similarity to a bright green. The brighter the shade of green, the more similar the objects are.

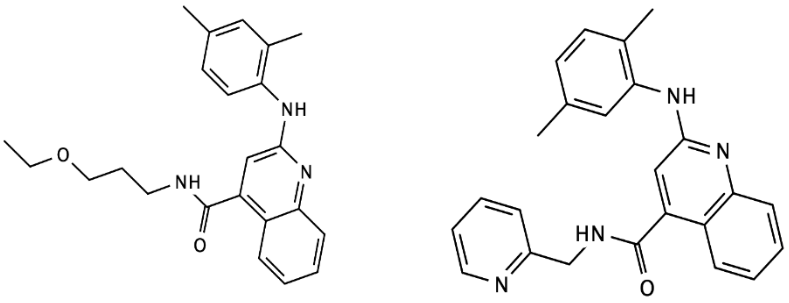


Figure 35: Comparison of compound C387-1105 (left) and C387-0954 (right)

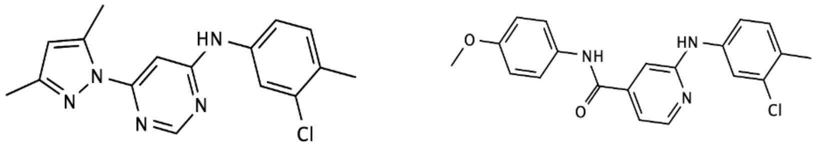


Figure 36: Comparison of compound L656-0064 (left) and F047-0446 (right)

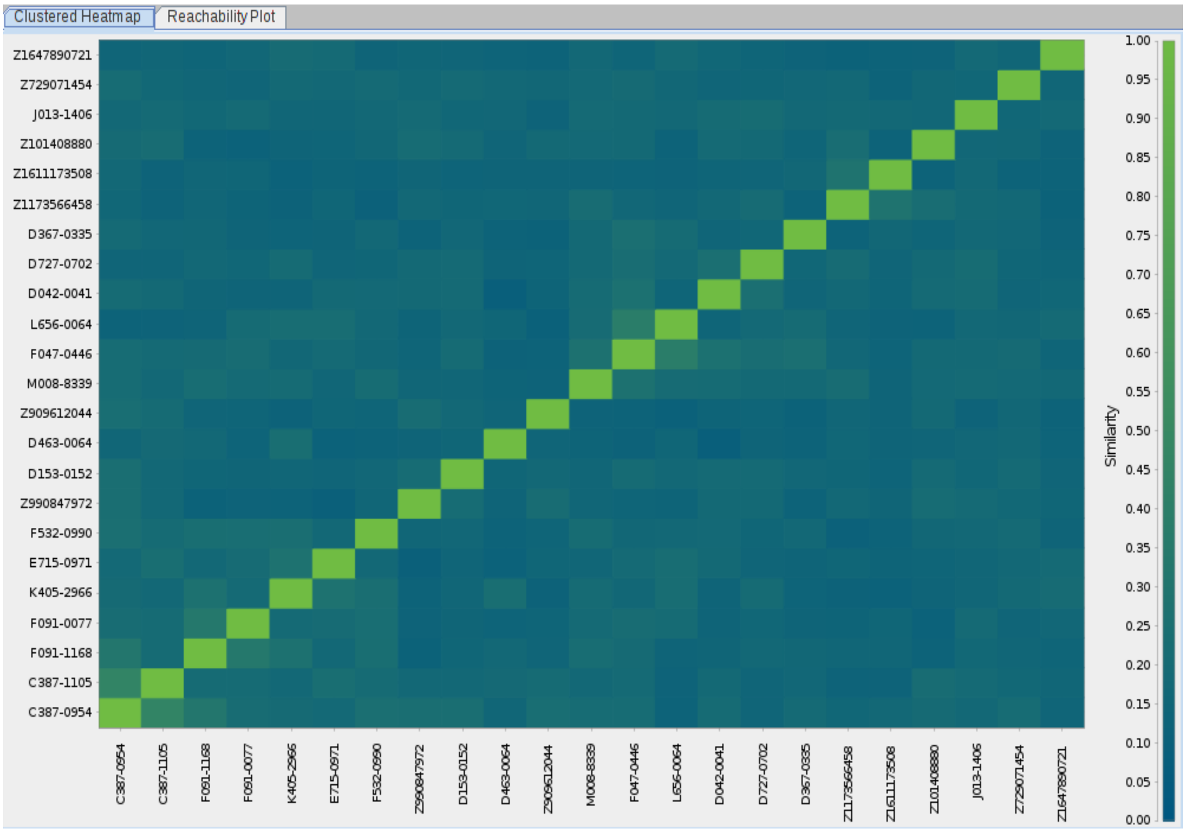


Figure 37: Heatmap of final compounds executed by Knime

4.17 PAINS check

The last step was the verification of PAINS (Pan-assay interference compounds) via an *in house* Knime workflow. They are classes of chemical compounds that can interfere with bioassays and often give false positive results in high-throughput screens. More than 450 compound classes have been designated as PAINS to date. Classified PAINS are typically small reactive or otherwise liable molecules that are contained as substructures in larger compounds. [50] The 23 final compounds were successfully transferred through a Knime workflow in the format of a csv file. The csv file is containing the 23 compounds represented as canonical SMILES. Using the PAINS method, it was ensured for each compound that there should be no interaction with the bioactivity assay. (Table 17)

NAME	2D STRUCTURE	PAINS	PROCESSED
C387-0954	Cc1ccc(C)c(Nc2cc(C(=O)NCc3ccccn3)c3ccccc3n2)c1	0	True
C387-1105	CCOCCCNC(=O)c1cc(Nc2ccc(C)cc2C)nc2ccccc12	0	True
D042-0041	COc1ccc(N(Cc2cc3cccc(C)c3[nH]c2=O)S(C)(=O)=O)cc1	0	True
D153-0152	CC(=O)c1nn(-c2ccc(OC(F)(F)F)cc2)nc1Nc1cccc2ncccc12	0	True
D367-0335	Cc1ccc(Nc2ncc3c(n2)CC(C)CC3=O)cc1C	0	True
D463-0064	CCCNc1cc(C)nc2c(-c3ccc(F)cc3)c(C)nn12	0	True
D727-0702	COc1ccc(OCc2nn3c(-c4cc5ccccc5[nH]4)nnc3s2)cc1	0	True
E715-0971	Cc1cc(Cl)ccc1Nc1ncnc2c1nc1n2CCCC1	0	True
F047-0446	COc1ccc(NC(=O)c2ccnc(Nc3ccc(C)c(Cl)c3)c2)cc1	0	True
F091-0077	Cc1ccc(Cl)cc1Nc1nc(N)c(C(=O)C(C)C)s1	0	True
F091-1168	Cc1ccc(C)c(Nc2nc(N)c(C(=O)c3ccc(F)cc3)s2)c1	0	True
F532-0990	CCc1nc(-c2ccc(NC(=O)Nc3ccccc3C)cc2)no1	0	True
J013-1406	CCCCOc1ccc(C(=O)C2CC(=O)Nc3c2c(C)nn3-c2cccc(C)c2C)cc1OC	0	True
K405-2966	Cc1ccccc1Nc1ncnc2c1cnn2-c1ccc(F)cc1	0	True
L656-0064	Cc1cc(C)n(-c2cc(Nc3ccc(C)c(Cl)c3)ncn2)n1	0	True
M008-8339	COc1ccc(NC(=O)c2sc(Nc3ncccc3C)nc2C)cc1OC	0	True
ROW13081	Cc1nn(-c2ccccc2)c(C)c1CNc1nc2ccccc2c(=O)[nH]1	0	True
ROW13258	O=C1NCCCCC1Nc1nc(C(F)(F)F)nc2ccccc12	0	True
ROW13542	Cc1nn(C(C)C)c(C)c1Nc1cnc2ccccc2n1	0	True
ROW15273	Cc1noc(C)c1CC(=O)Nc1ncnc2[nH]c(C)c(C)c12	0	True
ROW3845	COCCc1ccc2oc(NCc3scnc3C)nc2c1	0	True
ROW4557	CCN(Cc1nc2cc(OC)c(OC)cc2c(=O)[nH]1)C(=O)c1ccoc1C	0	True
ROW8325	COc1cc2nc(NCc3scnc3C)nc(N)c2cc1OC	0	True

Table 17: PAINS results

5 Conclusion and outlook

As a result of this study 23 compounds were selected using different in silico methods and proposed for biological testing as potential TTK inhibitors. In detail, after the selection of the appropriate crystal structures available on the Protein Data Bank, a structure-based pharmacophore model was developed. Its performances were evaluated using compounds tested on TTK present on the ChEMBL database. Molecular docking studies were performed in Glide and the hit compounds were further ranked with the combination of four different scoring functions. The procedure includes a refinement of the virtual screening output with comparing of docking poses, clustering, similarly search to ensure good diversity coverage and also solubility and permeability prediction was used as final filter. After final visual inspection the 23 best compounds were selected out of 1285 hits. The list was refined according to the availability of the molecules from the vendors. The potential inhibitors of TTK kinase were biologically tested at 10 µM by CEREP. [51] Unfortunately this first iteration of the project didn't give the wanted outcome, by showing no biological activity for the selected molecules. However, one of the compounds was tested also at 100 µM and found having a weak activity against the target. Such a molecule might be considered as a new starting point for the second iteration of the project in which new compounds will be selected taking into consideration the work done within this thesis.

6 References

[1] American Cancer Society Inc., “Cancer Facts & Figures 2019,” 2019.

[2] L. Furchtgott *et al.*, “Multiple Myeloma Drivers of High Risk and Response to Stem Cell Transplantation Identified By Causal Machine Learning: Out-of-Cohort and Experimental Validation,” *Blood*, vol. 130, no. Suppl 1, pp. 3029–3029, Dec. 2017.

[3] S. T. Pachis and G. J. P. L. Kops, “Leader of the SAC: molecular mechanisms of Mps1/TTK regulation in mitosis,” *Open Biol*, vol. 8, no. 8, Aug. 2018, doi: 10.1098/rsob.180109.

[4] A. C. S. Inc., “What Is Multiple Myeloma?,” 28-Feb-2018. [Online]. Available: https://www.cancer.org/cancer/multiple-myeloma/about/what-is-multiple-myeloma.html. [Accessed: 31-Jan-2019].

[5] National Institute of Health Research, “Multiple myeloma.” [Online]. Available: https://www.nihr.ac.uk/research-and-impact/making-a-difference/multiple-myeloma.htm. [Accessed: 30-Mar-2019].

[6] Benjamin Parsons, DO, “UNDERSTANDING MULTIPLE MYELOMA AND LABORATORY VALUES.”

[7] “fig1.2.jpg 513×444 Pixel.” [Online]. Available: https://bpac.org.nz/BT/2011/July/img/fig1.2.jpg. [Accessed: 16-Apr-2019].

[8] C. Röllig, S. Knop, and M. Bornhäuser, “Multiple myeloma,” *The Lancet*, vol. 385, no. 9983, pp. 2197–2208, May 2015, doi: 10.1016/S0140-6736(14)60493-1.

[9] Dana-Farber Cancer Institute, “Multiple Myeloma: Signs and Symptoms,” 21-Mar-2017. [Online]. Available: https://blog.dana-farber.org/insight/2017/03/signs-and-symptoms-of-multiple-myeloma/. [Accessed: 15-Apr-2019].

[10] “What Causes Multiple Myeloma?” [Online]. Available: https://www.cancer.org/cancer/multiple-myeloma/causes-risks-prevention/what-causes.html. [Accessed: 31-Jan-2019].

[11] Dana-Farber Cancer Institute, “What is the Role of Genetic Analysis in Multiple Myeloma?,” *Dana-Farber Cancer Institute*, 08-Jun-2018. [Online]. Available: https://blog.dana-farber.org/insight/2018/06/role-genetic-analysis-multiple-myeloma/. [Accessed: 18-Apr-2019].

[12] Mayo Clinic, “Multiple myeloma - Symptoms and causes.” [Online]. Available: https://www.mayoclinic.org/diseases-conditions/multiple-myeloma/symptoms-causes/syc-20353378. [Accessed: 31-Jan-2019].

[13] P. B. Jarreau, “An Ounce of Prevention for Multiple Myeloma Requires Data, Data, Data,” *Medium*, 18-Jul-2018. [Online]. Available: https://medium.com/lifeomic/an-ounce-of-prevention-for-multiple-myeloma-requires-data-data-data-21157ce26f38. [Accessed: 14-Apr-2019].

[14] Mayo Clinic, “Multiple myeloma - Diagnosis and treatment.” [Online]. Available: https://www.mayoclinic.org/diseases-conditions/multiple-myeloma/diagnosis-treatment/drc-20353383. [Accessed: 31-Jan-2019].

[15] W. Forth, D. Henschler und W. Rummel, *Allgemeine und spezielle Pharmakologie und Toxikologie*, 11th ed. .

[16] L. Furchtgott *et al.*, “Multiple Myeloma Drivers of High Risk and Response to Stem Cell Transplantation Identified By Causal Machine Learning: Out-of-Cohort and Experimental Validation,” *Blood*, vol. 130, no. Suppl 1, pp. 3029–3029, Dec. 2017.

[17] M. Hassan Baig *et al.*, “Computer Aided Drug Design: Success and Limitations,” *Current Pharmaceutical Design*, vol. 22, no. 5, pp. 572–581, Jan. 2016, doi:

10.2174/1381612822666151125000550.

[18] S. Parasuraman, “Protein data bank,” *J Pharmacol Pharmacother*, vol. 3, no. 4, pp. 351–352, Oct. 2012, doi: 10.4103/0976-500X.103704.

[19] The UniProt Consortium, “The Universal Protein Resource (UniProt),” *Nucleic Acids Research*, vol. 36, no. Database, pp. D190–D195, Dec. 2007, doi: 10.1093/nar/gkm895.

[20] The UniProt Consortium, “UniProt: the universal protein knowledgebase,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D158–D169, Jan. 2017, doi: 10.1093/nar/gkw1099.

[21] A. Gaulton *et al.*, “The ChEMBL database in 2017,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D945–D954, Jan. 2017, doi: 10.1093/nar/gkw1074.

[22] M. M. Mysinger, M. Carchia, John. J. Irwin, and B. K. Shoichet, “Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking,” *Journal of Medicinal Chemistry*, vol. 55, no. 14, pp. 6582–6594, Jul. 2012, doi: 10.1021/jm300687e.

[23] T. W. H. Backman, Y. Cao, and T. Girke, “ChemMine tools: an online service for analyzing and clustering small molecules,” *Nucleic Acids Research*, vol. 39, no. suppl, pp. W486–W491, Jul. 2011, doi: 10.1093/nar/gkr320.

[24] S.-Y. Yang, “Pharmacophore modeling and applications in drug discovery: challenges and recent advances,” *Drug Discovery Today*, vol. 15, no. 11–12, pp. 444–450, Jun. 2010, doi: 10.1016/j.drudis.2010.03.013.

[25] S. eswari Jujjavarapu, S. Dhagat, and M. Yadav, *Computer-Aided Design of Antimicrobial Lipopeptides as Prospective Drug Candidates*, 1st ed. Boca Raton, Florida : CRC Press, 2019.: CRC Press, 2019.

[26] Hartl Robert, “Pharmacophore.”

[27] S.-Y. Yang, “Pharmacophore modeling and applications in drug discovery: challenges and recent advances,” *Drug Discovery Today*, vol. 15, no. 11–12, pp. 444–450, Jun. 2010, doi: 10.1016/j.drudis.2010.03.013.

[28] M. Wieder, U. Perricone, T. Seidel, S. Boresch, and T. Langer, “Comparing pharmacophore models derived from crystal structures and from molecular dynamics simulations,” *Monatsh Chem*, vol. 147, pp. 553–563, 2016, doi: 10.1007/s00706-016-1674-1.

[29] S.-Y. Huang, S. Z. Grinter, and X. Zou, “Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions,” *Physical Chemistry Chemical Physics*, vol. 12, no. 40, p. 12899, 2010, doi: 10.1039/c0cp00151a.

[30] “LigandScout User Manual,” p. 143.

[31] M. R. Berthold *et al.*, “KNIME – The Konstanz Information Miner,” vol. 11, no. 1, p. 6.

[32] KNIME, “KNIME Analytics Platform.” [Online]. Available: https://www.knime.com/knime-software/knime-analytics-platform. [Accessed: 03-Feb-2019].

[33] “Phylogenetic trees,” *Khan Academy*. [Online]. Available: https://www.khanacademy.org/science/high-school-biology/hs-evolution/hs-phylogeny/a/phylogenetic-trees. [Accessed: 31-Jan-2019].

[34] P. Innocenti *et al.*, “Rapid Discovery of Pyrido[3,4- *d*]pyrimidine Inhibitors of Monopolar Spindle Kinase 1 (MPS1) Using a Structure-Based Hybridization Approach,” *Journal of Medicinal Chemistry*, vol. 59, no. 8, pp. 3671–3688, Apr. 2016, doi: 10.1021/acs.jmedchem.5b01811.

[35] M. D. Gurden *et al.*, “Naturally Occurring Mutations in the MPS1 Gene Predispose

Cells to Kinase Inhibitor Drug Resistance," *Cancer Research*, vol. 75, no. 16, pp. 3340–3354, Aug. 2015, doi: 10.1158/0008-5472.CAN-14-3272.

[36] "Structural and mechanistic insights into Mps1 kinase activation." [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2829362/. [Accessed: 04-Apr-2019].

[37] Wei Wang, Yuting Yang , Yuefeng Gao, "Structural and mechanistic insights into Mps1 kinase activation."

[38] J. C. M. Uitdehaag *et al.*, "Target Residence Time-Guided Optimization on TTK Kinase Results in Inhibitors with Potent Anti-Proliferative Activity," *Journal of Molecular Biology*, vol. 429, no. 14, pp. 2211–2230, Jul. 2017, doi: 10.1016/j.jmb.2017.05.014.

[39] Y. Liu *et al.*, "The Discovery of Orally Bioavailable Tyrosine Threonine Kinase (TTK) Inhibitors: 3-(4-(heterocyclyl)phenyl)-1 *H* -indazole-5-carboxamides as Anticancer Agents," *Journal of Medicinal Chemistry*, vol. 58, no. 8, pp. 3366–3392, Apr. 2015, doi: 10.1021/jm501740a.

[40] K. Kusakabe *et al.*, "Discovery of Imidazo[1,2- *b*]pyridazine Derivatives: Selective and Orally Available Mps1 (TTK) Kinase Inhibitors Exhibiting Remarkable Antiproliferative Activity," *Journal of Medicinal Chemistry*, vol. 58, no. 4, pp. 1760–1775, Feb. 2015, doi: 10.1021/jm501599u.

[41] K. Kusakabe *et al.*, "Indazole-Based Potent and Cell-Active Mps1 Kinase Inhibitors: Rational Design from Pan-Kinase Inhibitor Anthrapyrazolone (SP600125)," *Journal of Medicinal Chemistry*, vol. 56, no. 11, pp. 4343–4356, Jun. 2013, doi: 10.1021/jm4000215.

[42] P. Innocenti *et al.*, "Rapid Discovery of Pyrido[3,4- *d*]pyrimidine Inhibitors of Monopolar Spindle Kinase 1 (MPS1) Using a Structure-Based Hybridization Approach," *Journal of Medicinal Chemistry*, vol. 59, no. 8, pp. 3671–3688, Apr. 2016, doi: 10.1021/acs.jmedchem.5b01811.

[43] S. Naud *et al.*, "Structure-Based Design of Orally Bioavailable 1 *H* -Pyrrolo[3,2- *c*]pyridine Inhibitors of Mitotic Kinase Monopolar Spindle 1 (MPS1)," *Journal of Medicinal Chemistry*, vol. 56, no. 24, pp. 10045–10065, Dec. 2013, doi: 10.1021/jm401395s.

[44] P. S. Charifson, J. J. Corkery, M. A. Murcko, and W. P. Walters, "Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins," *Journal of Medicinal Chemistry*, vol. 42, no. 25, pp. 5100–5109, Dec. 1999, doi: 10.1021/jm990352k.

[45] R. A. Friesner *et al.*, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy," *Journal of Medicinal Chemistry*, vol. 47, no. 7, pp. 1739–1749, Mar. 2004, doi: 10.1021/jm0306430.

[46] M. Muchtaridi, H. Syahidah, A. Subarnas, M. Yusuf, S. Bryant, and T. Langer, "Molecular Docking and 3D-Pharmacophore Modeling to Study the Interactions of Chalcone Derivatives with Estrogen Receptor Alpha," *Pharmaceuticals*, vol. 10, no. 4, p. 81, Oct. 2017, doi: 10.3390/ph10040081.

[47] "KNIME Hub," *KNIME Hub*. [Online]. Available: https://hub.knime.com/. [Accessed: 15-Feb-2020].

[48] S. Genheden and U. Ryde, "The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities," *Expert Opinion on Drug Discovery*, vol. 10, no. 5, pp. 449–461, May 2015, doi: 10.1517/17460441.2015.1032936.

[49] Schrödinger, "QikProp User Manual," p. 50.

[50] S. Jasial, Y. Hu, and J. Bajorath, "How Frequently Are Pan-Assay Interference Compounds Active? Large-Scale Analysis of Screening Data Reveals Diverse Activity Profiles, Low Global Hit Frequency, and Many Consistently Inactive Compounds," *Journal*

of Medicinal Chemistry, vol. 60, no. 9, pp. 3879–3886, May 2017, doi: 10.1021/acs.jmedchem.7b00154.

[51] "CEREP Laboratories France," *Eurofins Scientific*. [Online]. Available: https://www.eurofins.com/contact-us/worldwide-interactive-map/france/eurofins-cerep-france/. [Accessed: 10-Feb-2020].

7 Appendix

7.1 Lists

7.1.1 List of Figures

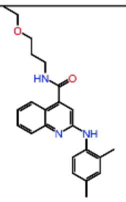
7.1.2 List of Tables

7.1.3 List of Abbreviations

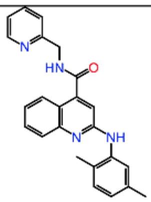
ADME	ABSORPTION DISTRIBUTION METABOLISM EXCRETION
ATP	ADENOSINE TRIPHOSPHATE
CADD	COMPUTER AIDED DRUG DESIGN
CT	COMPUTED TOMOGRAPHY
DMSO	DIMETHYL SULFOXIDE
DNA	DEOXYRIBONUCLEIC ACID
DUD - E	DATABASE OF USEFUL DECOYS – ENHANCED
FDA	FOOD AND DRUG ADMINISTRATION
HTS	HIGH – THROUGHPUT SCREENING
I.V	INTRAVENOUS
IC_{50}	HALF MAXIMAL INHIBITORY CONCENTRATION
IFM / DFCI	INTERGROUPE FRANCOPHONE MYELOMA / DANA - FARBER CANCER INSTITUTE
IUPAC	INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY
MACCS	MOLECULAR ACCESS SYSTEM
MDS	MOLECULAR DYNAMICS SIMULATION
MM	MULTIPLE MYELOMA
MM/GBSA	MOLECULAR MECHANICS / GENERALIZED BORN SURFACE AREA
MPS1	MONOPOLAR SPINDLE 1
MRI	MAGNETIC RESONANCE IMAGING
NMR	NUCLEAR MAGNETIC RESONANCE
PAINS	PAN –ASSAY INTERFERENCE COMPOUNDS

PDB	PROTEIN DATA BANK
PEG	POLYETHYLENE GLYCOL
PET	POSITRON EMISSION TOMOGRAPHY
RBR	RANK BY RANK
RMSD	ROOT - MEAN - SQUARE – DEVIATION
S.C	SUBCUTANEOUS
SAC	SPINDLE ASSEMBLY CHECKPOINT
SMM	SMOLDERING MULTIPLE MYELOMA
SPEP	SERUM PROTEIN ELECTROPHORESIS
TTK	TYROSINE THREONINE KINASE

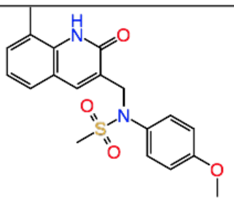
ANNEX A: 23 compounds for experimental testing



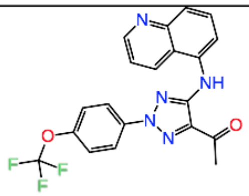
title: C387-1105



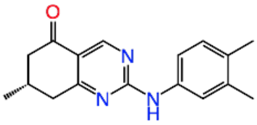
title: C387-0954



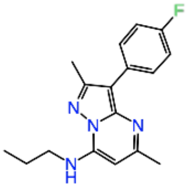
title: D042-0041



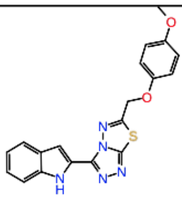
title: D153-0152



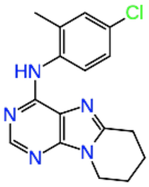
title: D367-0335



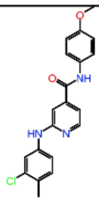
title: D463-0064



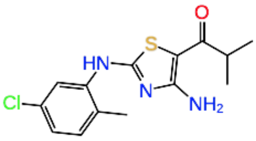
title: D727-0702



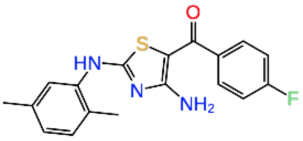
title: E715-0971



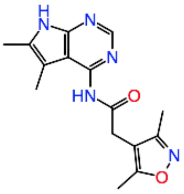
title: F047-0446



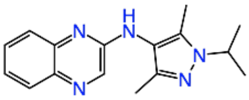
title: F091-0077



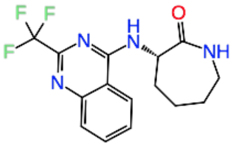
title: F091-1168



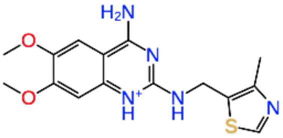
title: Row15273



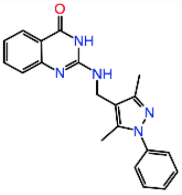
title: Row13542



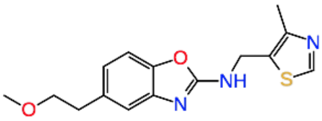
title: Row13258



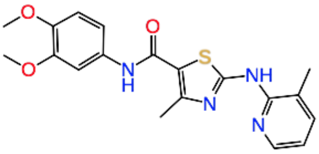
title: Row8325



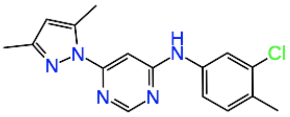
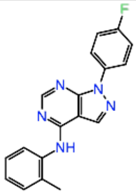
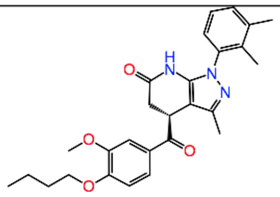
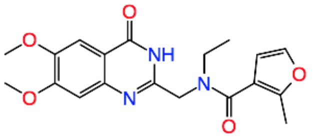
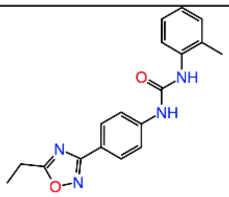
title: Row13081



title: Row3845



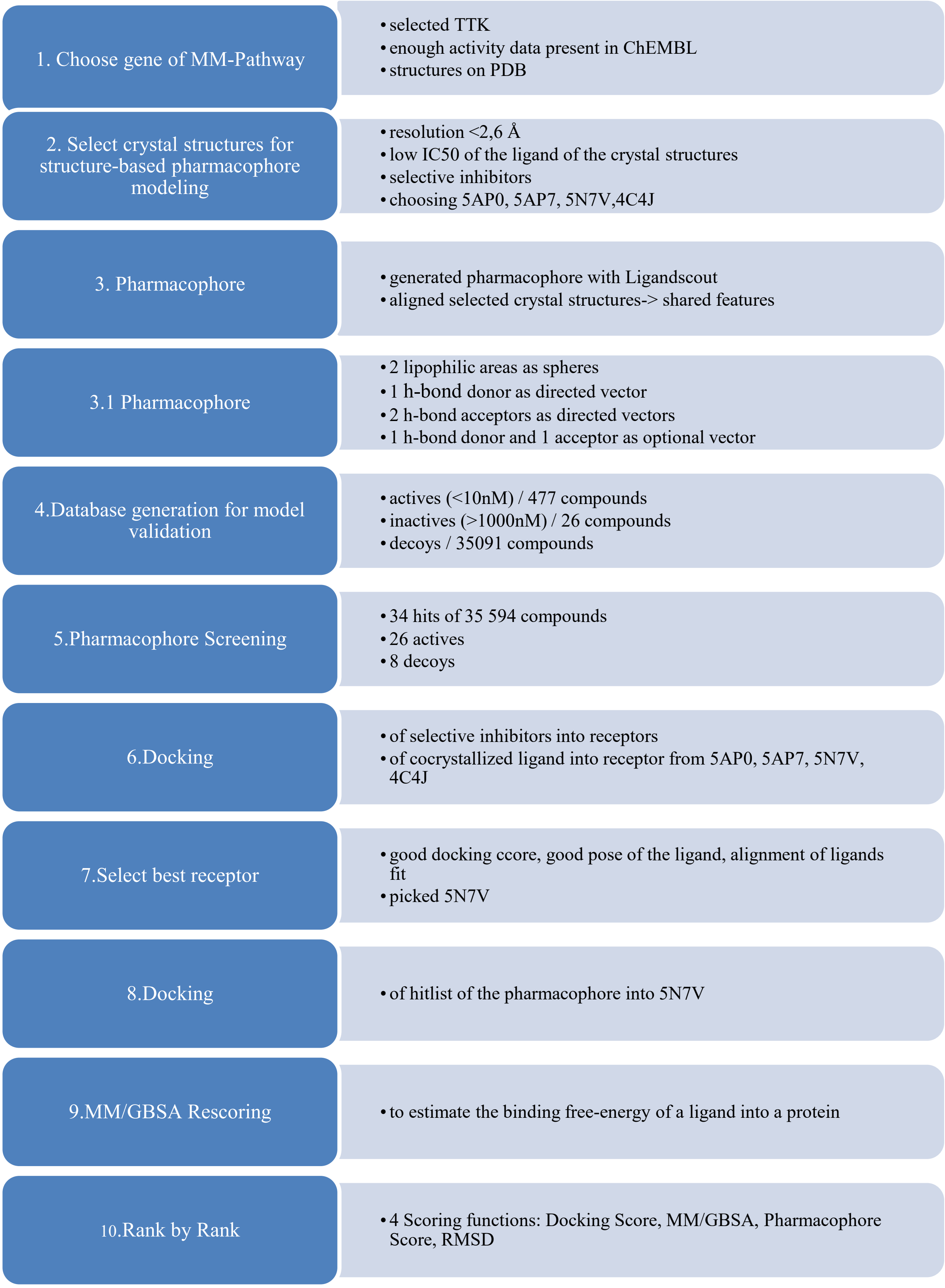
title: M008-8339

 title: L656-0064	 title: K405-2966	 title: J013-1406
 title: Row4557	 title: F532-0990	

ANNEX B: 23 compounds - vendors information

NAME	VENDOR	PRODUCT ID
C387-0954	ChemDiv	C387-0954
C387-1105	ChemDiv	C387-1105
D042-0041	ChemDiv	D042-0041
D153-0152	ChemDiv	D153-0152
D367-0335	ChemDiv	D367-0335
D463-0064	ChemDiv	D463-0064
D727-0702	ChemDiv	D727-0702
E715-0971	ChemDiv	E715-0971
F047-0446	ChemDiv	F047-0446
F091-0077	ChemDiv	F091-0077
F091-1168	ChemDiv	F091-1168
F091-1168	ChemDiv	F091-1168
J013-1406	ChemDiv	J013-1406
K405-2966	ChemDiv	K405-2966
L656-0064	ChemDiv	L656-0064
M008-8339	ChemDiv	M008-8339
ROW13081	Enamine	Z990847972
ROW13258	Enamine	Z909612044
ROW13542	Enamine	Z729071454
ROW15273	Enamine	Z1647890721
ROW3845	Enamine	Z1611173508
ROW4557	Enamine	Z101408880
ROW8325	Enamine	Z1173566458

ANNEX C: Workflow overview



1. Choose gene of MM-Pathway

- selected TTK
- enough activity data present in ChEMBL
- structures on PDB

2. Select crystal structures for structure-based pharmacophore modeling

- resolution <2,6 Å
- low IC50 of the ligand of the crystal structures
- selective inhibitors
- choosing 5AP0, 5AP7, 5N7V,4C4J

3. Pharmacophore

- generated pharmacophore with Ligandscout
- aligned selected crystal structures-> shared features

3.1 Pharmacophore

- 2 lipophilic areas as spheres
- 1 h-bond donor as directed vector
- 2 h-bond acceptors as directed vectors
- 1 h-bond donor and 1 acceptor as optional vector

4.Database generation for model validation

- actives (<10nM) / 477 compounds
- inactives (>1000nM) / 26 compounds
- decoys / 35091 compounds

5.Pharmacophore Screening

- 34 hits of 35 594 compounds
- 26 actives
- 8 decoys

6.Docking

- of selective inhibitors into receptors
- of cocrystallized ligand into receptor from 5AP0, 5AP7, 5N7V, 4C4J

7.Select best receptor

- good docking ccore, good pose of the ligand, alignment of ligands fit
- picked 5N7V

8.Docking

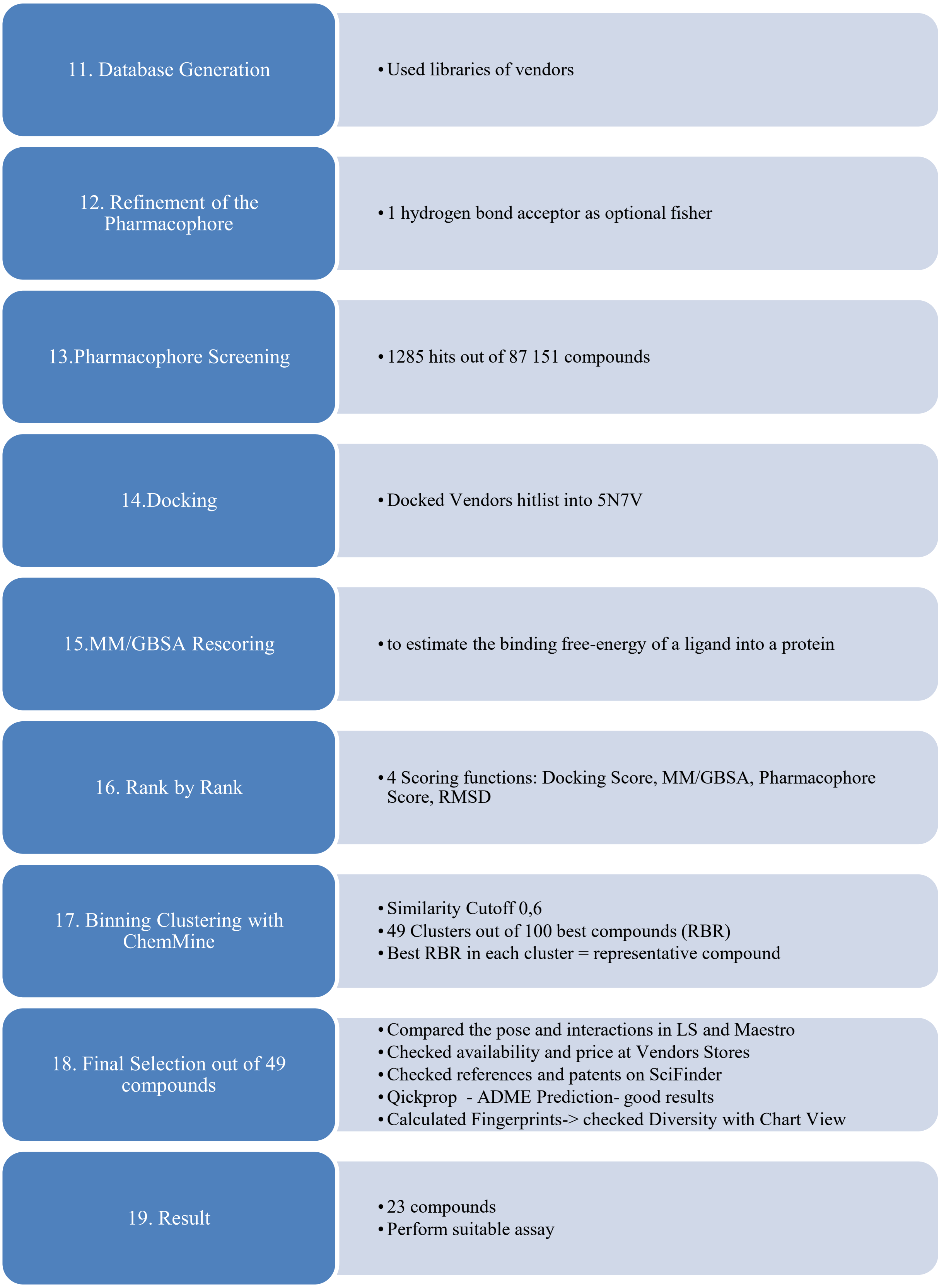
- of hitlist of the pharmacophore into 5N7V

9.MM/GBSA Rescoring

- to estimate the binding free-energy of a ligand into a protein

10.Rank by Rank

- 4 Scoring functions: Docking Score, MM/GBSA, Pharmacophore Score, RMSD



11. Database Generation

- Used libraries of vendors

12. Refinement of the Pharmacophore

- 1 hydrogen bond acceptor as optional fisher

13.Pharmacophore Screening

- 1285 hits out of 87 151 compounds

14.Docking

- Docked Vendors hitlist into 5N7V

15.MM/GBSA Rescoring

- to estimate the binding free-energy of a ligand into a protein

16. Rank by Rank

- 4 Scoring functions: Docking Score, MM/GBSA, Pharmacophore Score, RMSD

17. Binning Clustering with ChemMine

- Similarity Cutoff 0,6
- 49 Clusters out of 100 best compounds (RBR)
- Best RBR in each cluster = representative compound

18. Final Selection out of 49 compounds

- Compared the pose and interactions in LS and Maestro
- Checked availability and price at Vendors Stores
- Checked references and patents on SciFinder
- Qickprop - ADME Prediction- good results
- Calculated Fingerprints-> checked Diversity with Chart View

19. Result

- 23 compounds
- Perform suitable assay

ANNEX D: 100 best ranked compounds of the rank by rank (vendors library)

100 BEST RBR	NAME	NEW NAME	VENDOR
58,5	F091-1168	F091-1168_1	ChemDiv
61,25	K402-0096	K402-0096_1	ChemDiv
76,25	K405-2966	K405-2966_1	ChemDiv
84,75	K405-2965	K405-2965_1	ChemDiv
85	K405-2947	K405-2947_1	ChemDiv
91,25	F091-0068	F091-0068_1	ChemDiv
117	F091-0918	F091-0918_1	ChemDiv
131	K405-2935	K405-2935_1	ChemDiv
155,5	D042-0041	D042-0041_1	ChemDiv
157,75	F091-1010	F091-1010_1	ChemDiv
162	3486-0389	3486-0389_1	ChemDiv
163,25	F091-0645	F091-0645_1	ChemDiv
164,25	K405-3595	K405-3595_1	ChemDiv
166,5	E715-0971	E715-0971_1	ChemDiv
170,75	C387-0954	C387-0954_1	ChemDiv
172,5	F091-0456	F091-0456_1	ChemDiv
173,5	J013-1406	J013-1406_1	ChemDiv
176	F091-0077	F091-0077_1	ChemDiv
176	K405-2159	K405-2159_1	ChemDiv
178,5	3486-0352	3486-0352_1	ChemDiv
179,75	Row15273	Row15273_1	Enamine
180,75	Row15274	Row15274_1	Enamine
186	BDE 32453687	BDE 32453687_1	Asinex
187,75	E715-0803	E715-0803_1	ChemDiv
188,75	E715-0743	E715-0743_1	ChemDiv
188,75	F091-0075	F091-0075_1	ChemDiv
202,25	F550-3891	F550-3891_1	ChemDiv
204	F532-0990	F532-0990_1	ChemDiv
204,75	E715-0727	E715-0727_1	ChemDiv
213,5	E715-0850	E715-0850_1	ChemDiv
228,25	Row15271	Row15271_1	Enamine
229,5	S7807	S7807_1	Selleckchem
238	LAS 57281172	LAS 57281172_1	Asinex
239,5	F091-0920	F091-0920_1	ChemDiv
242,75	F091-0092	F091-0092_1	ChemDiv
242,75	Row7636	Row7636_1	Enamine
243	K402-0208	K402-0208_1	ChemDiv
244,75	62927947	62927947_1	ChemBridge
249,25	D463-0064	D463-0064_1	ChemDiv
254,75	Row13081	Row13081_1	Enamine
255	F091-0247	F091-0247_1	ChemDiv
259,5	E715-0927	E715-0927_1	ChemDiv
260	D393-0086	D393-0086_1	ChemDiv
261	E679-0771	E679-0771_1	ChemDiv
262,5	Row4623	Row4623_1	Enamine
262,75	Row17707	Row17707_1	Enamine
265,75	E679-0616	E679-0616_1	ChemDiv

265,75	L656-0064	L656-0064_1	ChemDiv
273,25	F091-0227	F091-0227_1	ChemDiv
274,75	Row3845	Row3845_1	Enamine
275,25	Row15276	Row15276_1	Enamine
278,5	C387-1105	C387-1105_1	ChemDiv
279,5	F091-0897	F091-0897_1	ChemDiv
282,5	Row15772	Row15772_1	Enamine
285	F091-0647	F091-0647_1	ChemDiv
288,25	F091-1031	F091-1031_1	ChemDiv
288,75	F091-0575	F091-0575_1	ChemDiv
290,75	Row4557	Row4557_1	Enamine
292	Row13258	Row13258_1	Enamine
292,75	J013-1389	J013-1389_1	ChemDiv
294	D463-0097	D463-0097_1	ChemDiv
294,25	S8523	S8523_1	Selleckchem
294,75	Row12162	Row12162_1	Enamine
295,5	F091-0264	F091-0264_1	ChemDiv
296	M008-8339	M008-8339_1	ChemDiv
297,25	F091-1180	F091-1180_1	ChemDiv
300,75	F091-0222	F091-0222_1	ChemDiv
305,5	F091-1021F	F091-1021F_1	ChemDiv
306,75	Row3393	Row3393_1	Enamine
308,5	K402-0322	K402-0322_1	ChemDiv
309,75	D367-0335	D367-0335_1	ChemDiv
310,25	E679-0804	E679-0804_1	ChemDiv
312,75	D153-0152	D153-0152_1	ChemDiv
316	Row8325	Row8325_1	Enamine
317	F091-1040	F091-1040_1	ChemDiv
318	K402-0100	K402-0100_1	ChemDiv
319,75	Row10181	Row10181_1	Enamine
320,5	E679-0598	E679-0598_1	ChemDiv
321,75	F254-3113	F254-3113_1	ChemDiv
322,25	Row13542	Row13542_1	Enamine
322,75	F360-0652	F360-0652_1	ChemDiv
326,5	S7709	S7709_1	Selleckchem
327,5	F091-0322	F091-0322_1	ChemDiv
327,75	D463-0090	D463-0090_1	ChemDiv
327,75	F873-0470	F873-0470_1	ChemDiv
328,25	65209451	65209451_1	ChemBridge
328,5	F091-1235	F091-1235_1	ChemDiv
328,5	Row14366	Row14366_1	Enamine
328,75	F532-3260	F532-3260_1	ChemDiv
330,5	F873-0492	F873-0492_1	ChemDiv
331,75	F091-0621	F091-0621_1	ChemDiv
332	60494006	60494006_1	ChemBridge
332,25	F047-0446	F047-0446_1	ChemDiv
332,75	D727-0702	D727-0702_1	ChemDiv
333	E679-0605	E679-0605_1	ChemDiv
335,25	Row4100	Row4100_1	Enamine
336	E715-0919	E715-0919_1	ChemDiv
336,5	8137-0130	8137-0130_1	ChemDiv
336,5	97912743	97912743_1	ChemBridge
338,25	LAS 57279675	LAS 57279675_1	Asinex