universität wien

Masterarbeit / Master's Thesis

Titel der Masterarbeit / Title of the Master's Thesis

Structure based analysis of the sodium/glucose co-transporter 2 (SGLT2)

verfasst von / submitted by

Ajouan Mazoudji, BSc

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

Magister pharmaciae (Mag. Pharm.)

Wien 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 605
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Pharmazie
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Gerhard Ecker

Acknowledgements

The last six months have been a challenging and exiting time and gave me the opportunity to gain such an amount of knowledge and so many unique experiences that it could last a lifetime. One of these outlook changing events was this thesis that will be presented throughout the following pages.

For this reason, I would like to extend my deepest gratitude to my supervisor Univ.-Prof. Mag. Dr. Gerhard Ecker, whose guidance and feedback made this thesis possible.

I am also grateful to the members of the Pharmacoinformatics Research Group and my fellow master's students for their advice and help to grow intellectually but also for their moral support. Thanks should also go to Josef Magerl for proofreading this thesis

Lastly, I'd like to thank my family and friends for their support, and Megrican who encouraged me throughout this time (and also introduced me to Python).

Table of Contents

1 Introduction

Diabetes mellitus is a serious and chronic condition and represents one of the most significant global health emergencies of the 21st century. In 2021, 537 million adults were estimated to live with diabetes which represents 10.5% of the world's adult population and this number is expected to rise to 643 million (11.3%) by 2030 (International Diabetes Federation, 2021).

Despite the existence of a multitude of therapeutical options for the treatment of type 2 diabetes, which accounts for about 90% of all diabetes cases, many people experience failure in glycaemic control and adverse effects such as weight gain and hypoglycaemia (Clar et al., 2012; International Diabetes Federation, 2021). Additionally, most of the existing treatments are dependent on the insulin production and thus often require an increase of the dose as the disease advances and insulin production declines (Whaley et al., 2012). Therefore, in recent years, it has become apparent that there is a need of new therapeutics with reduced undesirable side effects and no insulin dependence in order to maximize the patient's quality of life (Bhattacharya et al., 2020; Clar et al., 2012).

SGLT2 inhibitors are a drug family that have been recently introduced onto the market, show a unique mechanism of action and have been mainly used as second-line therapeutics for the treatment of type 2 diabetes (Clar et al., 2012).

The introduction of new drugs to the market is a tedious and costly process which involves the investment of up to 1.8 billion US dollars over the course of 10-15 years (Macalino et al., 2015; Paul, 2010). In order to reduce costs and time-consuming tasks involved in the preliminary stage of drug discovery, computer-aided drug discovery (CADD) has become an essential part of the process (Macalino et al., 2015). One of the most-often used tools in CADD has been molecular docking, a structure-based *in silico* method which has been successfully implemented for the development and improvement of new drugs several times (Sethi et al., 2020). Examples include the improvement of inhibitory activity of Aurora Kinase A inhibitors and the design of new Cyclooxygenase inhibitors (Park et al. 2018; Dadashpour et al., 2015).

With this background in mind, the aim of this thesis is to provide an *in-silico*-based method, mainly driven by the use of molecular docking, for the assessment of the

activity of potential SGLT2 inhibitors that have not yet been tested *in vitro*. Furthermore, over the course of this study the methods have been applied to a series of compounds with unknown activity with the objective to support the development of new SGLT2 inhibitors.

1.1 Diabetes and Sodium-Dependent Glucose Co-Transporters (SGLT)

Diabetes mellitus is characterized by raised levels of blood glucose which is caused by a decrease or lack of the production of the hormone insulin or the lack of the body's ability to effectively use the produced insulin (International Diabetes Federation, 2021). According to the American Diabetes Association (ADA), diabetes can be classified into the four following general categories:

> " 1. Type 1 diabetes (due to autoimmune β-cell destruction, usually leading to absolute insulin deficiency)
>
> 2. Type 2 diabetes (due to a progressive loss of β-cell insulin secretion frequently on the background of insulin resistance)
>
> 3. Gestational diabetes mellitus (GDM) (diabetes diagnosed in the second or third trimester of pregnancy that was not clearly overt diabetes prior to gestation)
>
> 4. Specific types of diabetes due to other causes, e.g., monogenic diabetes syndromes (such as neonatal diabetes and maturity-onset diabetes of the young [MODY]), diseases of the exocrine pancreas (such as cystic fibrosis and pancreatitis), and drug- or chemical-induced diabetes (such as with glucocorticoid use, in the treatment of HIV/AIDS, or after organ transplantation)." (American Diabetes Association, 2017)

Among these types of diabetes mellitus, type 2 is the one that accounts for the overwhelming majority of cases (90%) and is therefore the one which requires the most attention (International Diabetes Federation, 2021). Type 2 diabetes is a multisystemic and progressive disease and must be treated by using a multifactorial therapy, which often involves the combination of multiple drugs at once (Shubrook et al., 2015). Most of the currently used treatments need insulin production in order to

provide therapeutic benefits. In recent years there has been a novel development in the field of diabetes type 2 therapy, as the emergence of the kidney as a treatment target provided the possibility of avoiding a dependency on insulin production (Clar et al., 2012; Shubrook et al., 2015).

The sodium-dependent glucose transporters (SGLT) are a family of three proteins which function either as sugar transporters (SGLT1 and SGLT2) or as a sensor (SGLT3) (Scheepers et al., 2004) and are expressed throughout the body, such as the intestine, the kidney, and specific regions of the brain (Wright et al., 2011; Shubrook et al., 2015). Glucose reabsorption in the kidney is one of the factors involved in maintaining a delicate balance keeping a physiologically healthy plasma concentration (Shubrook et al., 2015). Glucose reabsorption in the kidney is mediated by SGLT1 and SGLT2: The majority of this reabsorption is carried out in the first part of the proximal tubule where 90% of the filtered glucose is removed from the filtrate, while the remaining 10% are reabsorbed by SGLT1 in the later parts of the proximal tubule (Wright et al., 2011).

A naturally occurring mutation of SLC5A2, which is the gene encoding the protein SGLT2, leads to a defect protein and significant glycosuria. Thus the kidney, more specifically SGLT2 has been introduced as a novel target in the therapy of type 2 diabetes: SGLT2 inhibitors have been developed to mimic the effect of SLC5A2 mutation and inhibit the reabsorption of glucose (Clar et al., 2012).

Even though SGLT1 is also involved in the reabsorption of glucose in the kidney, a selective inhibition of SGLT2 may be of benefit as it is apparent that patients with non-functioning SGLT1 are suffering from gastrointestinal complications like severe diarrhoea (Shubrook et al., 2015).

1.2 Proteins and their ligands

In the following two chapters (1.3 and 1.4) the molecules that are used as SGLT2 inhibitors (ligands) and the protein SGLT2 itself will be further analysed and described. However, in order to achieve a deeper understanding of the underlying principles of the discussed topics it is important to have an overview of the behaviour of protein-ligand interactions.

Molecular recognition is the process of ligand interaction with a protein in order to form a specific complex and refers to a set of phenomena which may be described as being controlled by specific noncovalent interactions of the ligand with the amino acids of the protein (Gellman, 1997; Du et al., 2016). Much of the current knowledge of this process stems from high-resolution crystal structures and time-resolved spectroscopy, which has revealed that molecular recognition is a very dynamic event, in which both the ligand and the protein can change their conformation in order to bind (Morando et al., 2016).

The dynamic properties of the protein have been considered in the two currently most relevant hypotheses regarding the binding of ligands to their proteins (induced fit hypothesis and conformational selection hypothesis) and have replaced the older lock-and-key hypothesis where it is assumed that both the ligand and the protein are rigid and that their binding surfaces should match perfectly (Du et al., 2016).

Since the protein-ligand-solvent system is a thermodynamic system, the association between proteins and ligands is dictated by thermodynamic rules and only occurs when the change of Gibbs free energy (ΔG) is negative (Du et al., 2016). As the Gibbs free energy can be divided into its enthalpic and its entropic contributions, the relationship between the binding enthalpy (ΔH) and the Gibbs free energy can be represented by the following equation:

$\Delta G = \Delta H - T\Delta S$ (Du et al., 2016)

The binding enthalpy represents the energy change of a system upon the binding of a ligand to a protein. In a non-strict sense, ΔH is usually treated as the changes in energy that are resulting from the establishment of above-mentioned noncovalent interactions. Examples for such interactions are hydrogen-bonds, ion pairs, pi-pi-stacking, van der Waals contacts, and polar and apolar interactions (Du et al., 2016).

In order to compute the binding affinities of ligand-protein interactions, docking programs like Glide, which was employed for this thesis, use scoring functions (Friesner et al., 2004). These scoring functions are generally measuring the strength of the noncovalent interactions and are usually simplified to allow faster computational calculations (Du et al., 2016).

1.3 SGLT2 inhibitors

The history of SGLT2 inhibitors traces back to more than a hundred years ago, when phlorizin, which is a member of the chalcone class of organic compounds (figure 1), was isolated from the bark of the apple tree and identified to have a glycosuric effect (Ehrenkranz et al., 2005).

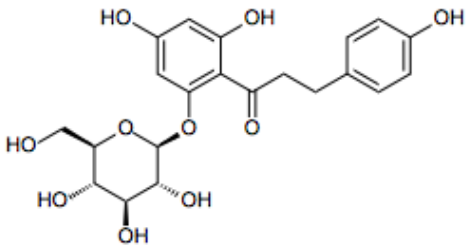
However, due to its low selectivity for SGLT2, it was observed that phlorizin administration led to frequent gastrointestinal side effects. Another limitation that was apparent and prevented its clinical use is its degradation by glucosidase in the intestine, where it is degraded to phloretin (Ehrenkranz et al., 2005).

The process of searching and finding a potent and selective SGLT2 inhibitor, which is not subject to degradation in the intestine, resulted in the development of several C-glycoside compounds that were introduced onto the market in the first half of the last decade, among them dapagliflozin, canagliflozin and empagliflozin. These inhibitors are characterized by a glucose nucleus, a C-glycosidic moiety at position C1 of the sugar and two aromatic rings (figure 2) (Bhattacharya et al., 2020; Cai et al., 2015).

The overall structure activity relationships of various SGLT2 inhibitors can be summarized as follows:

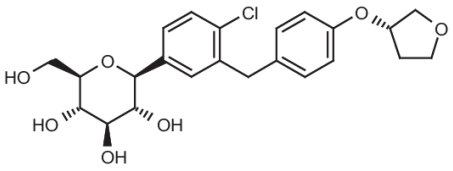
- The glucose moiety can only be substituted with different groups at positions C4 and C6 without losing activity. Substitutions with oxime at position C6 and with a strong electronegative group at C4 show the potential to enhance activity. A reversion of the configuration, especially at positions C1 and C5, leads to a diminishing of activity. Changing the hexose to a pentose can decrease the activity.
- The proximal benzene ring is essential for activity and the substitution with other (hetero)aryl groups may decrease the activity. Ortho and para substitutions may enhance the activity; a chlorine group is the most favourable substitution at the para position while an ether group is the most favourable substitution at ortho position.
- The methylene bridge is essential for activity and elongation could lead to decreased activity.

- The distal benzene ring is not essential for activity and other aryl groups can replace it in order to achieve better activity. Substitutions are only possible at para position and a small alkoxy group may enhance activity (Bhattacharya et al., 2020).

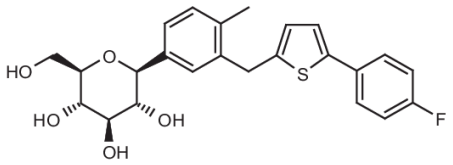


Phlorizin

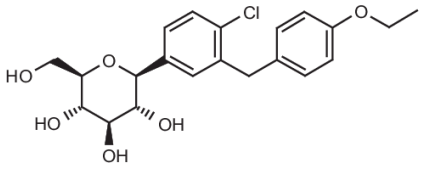
Figure 1: Chemical structure of Phlorizin which was the lead compound for the development of the modern potent SGLT2 inhibitors ("Phlorizin", n.d.).



Empagliflozin



Canagliflozin



Dapagliflozin

Figure 2: Chemical structure of SGLT inhibitors, usually characterized by their glucose moieties, proximal benzene rings substituted at the para position and a distal benzene ring which may be replaced by different heteroaryls (Chrysant, 2017).

In recent developments non-selective SGLT inhibitors have also been introduced to the market, despite previous concerns about the possibility of gastrointestinal side effects. However, long-term studies regarding their efficiency and safety are still needed (Tsimihodimos et al., 2018).

1.4 SGLT2 and its binding site

This thesis focuses on the human sodium-dependent glucose co-transporter 2 and the effects of its inhibition. Therefore, it is necessary to understand and analyse the

structure of the protein and its binding site. Recently, a cryogenic electron microscopy (cryo-EM) structure of SGLT2 co-crystallized with empagliflozin at its binding site was published for the first time, which provided a better understanding of the structure and the location of its binding site (Niu et al., 2021).

1.4.1 Overall and binding site structure

In humans, SGLT2 is existing in a complex with the membrane protein MAP17: an essential auxiliary subunit of SGLT2 which can enhance the activity of SGLT2 over a hundred-fold, and interacts with transmembrane helix 13 (Coady et al., 2016; Niu et al., 2021). The presence of MAP17 does not change the amount of expressed SGLT2, instead it is hypothesized that the interaction between MAP17 and SGLT2 changes the conformation of the co-transporter, allowing it to transport more glucose (Coady et al., 2016).

Overall, the SGLT2 transporter consists of 14 transmembrane helices and possesses a core that resembles the core structure of LeuT, which is a procaryotic neurotransmitter sodium transporter. The binding site is formed by amino acids of the transmembrane helices TM1, TM2, TM6, and TM10. In the cryo-EM structure it can be observed that the glucoside-group of the co-crystallized empagliflozin resides in the sugar-binding site where it forms a number of hydrogen bonds (Niu et al., 2021).

The main structural differences between the established SGLT2 inhibitors are characterized by their long hydrophobic aglycone tail at position C1 of their sugar, which, as seen in the co-crystallized structure of empagliflozin, branches from the sugar binding site and extends towards the extracellular side where it stays in the external vestibule. The two aromatic rings of empagliflozin show stacking with histidine 80 and phenylalanine 98 (Niu et al., 2021).

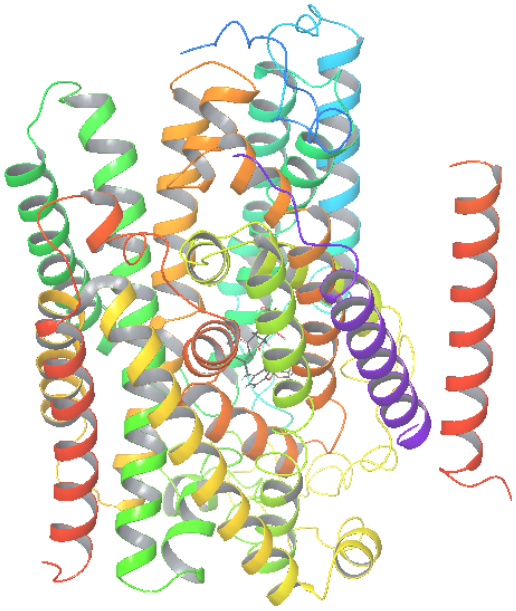


Figure 3: Overall structure of SGLT2. The structure was simplified by reducing it to a ribbon representation of the protein backbone in order to allow a better overview over the 14 transmembrane helices and the essential auxiliary transmembrane protein MAP17. MAP17 can be seen at the furthest right position as a red helix. At the very centre of the protein, the binding site with empagliflozin is located. This depiction of SGLT2 was generated by using Maestro, which is part of the Schrödinger Software (Schrödinger Release 2021-1: Maestro, 2021).

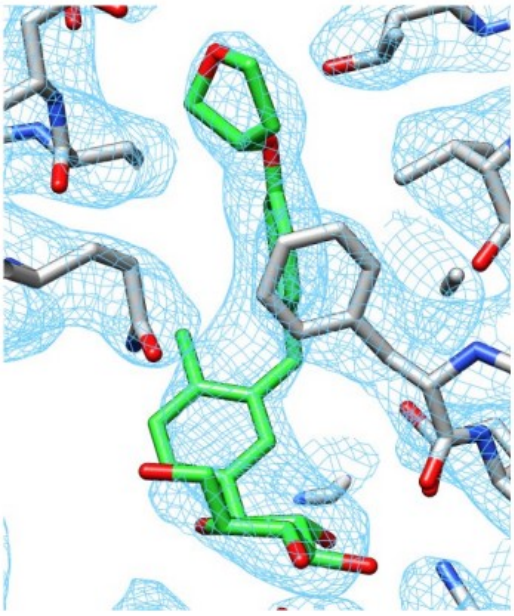


Figure 4: Depiction of empagliflozin at its binding site and nearby amino acid residues. The electron density is shown as a blue mesh and empagliflozin is represented by green sticks (Niu et al., 2021).

1.4.2 Transport mechanism

The occupation of both the sugar binding site and the outer vestibule by SGLT2 inhibitors leads to the lock of SGLT2 in its outward-open conformation which blocks the transport mechanism that is hypothesized to work similarly to vSGLT and LeuT (Niu et al., 2021). This would be a rocking-bundle alternating access mechanism where transported solutes bind to a site in the transporter that can be exposed to the other side of the membrane after conformational changes - in this case to the cytosolic side (Niu et al., 2021; Forrest & Rudnick, 2009).

The first step of the transport of glucose to the cytosolic side is characterized by the binding of sodium to SGLT2 in order to induce the opening of the outer gate of the

protein. Afterwards, the outer gate closes, through which glucose gets "trapped" inside the transporter before the inner gate opens and allows sodium and glucose to exit into the cytoplasm. After unloading sodium and glucose, the transporter finally returns to its outward conformation (figure 5) (Wright, 2020).

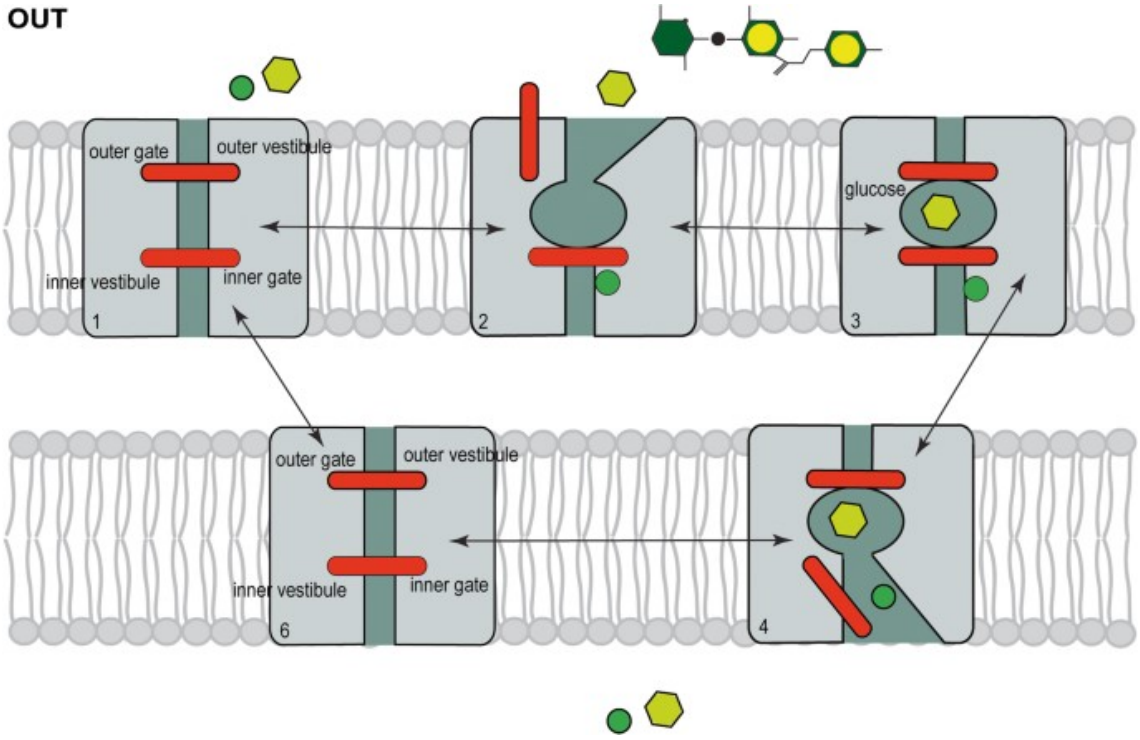


Figure 5: Mechanism of sodium-dependent glucose co-transport. The transporter starts in its outward position (upper left). After binding of sodium to the transporter, the outer gate opens and allows glucose to bind to the sugar binding site (upper middle). The outer gate closes and "traps" glucose inside (upper right). The inner gate opens and allows the exit of both glucose and sodium to the cytoplasm (lower right). Finally, the transporter returns to its starting conformation (lower left) (Wright, 2020).

As a result, SGLT2 transports one molecule glucose together with one sodium ion per transport cycle into the cytoplasm. The presence of sodium is necessary to allow the transport of glucose and it has therefore been proposed that there is an ion binding site at similar position as the Na2 site of LeuT (Niu et al., 2021). The binding of sodium leads to the adoption of the outward-open position of SGLT2 and from this knowledge it can be deduced that the binding of SGLT2 inhibitors also is only possible in the presence of sodium (Wright, 2020). However, the only currently available cryo-EM structure which captured SGLT2 in an open-outward position with empagliflozin bound to it was not able to find a strong density of sodium at the proposed position. This was explained by the authors as a consequence of the low binding affinity of SGLT2 to sodium at 0 mV (Niu et al., 2021).

2 Methods and Materials

2.1 Availability and evaluation of protein structures

In order to evaluate the availability of structures of SGLT2 for the docking process, a thorough search was performed.

The protein data bank (PDB) is a publicly available data bank containing information about the 3D shapes of proteins, nucleic acids, and complex assemblies. In addition to coordinate data in PDB and mmCIF formats, structure factor files and NMR constraint files, PDB entries provide documentation derived data (Berman, 2000)

As the 3D structure of SGLT2 was recently solved for the first time by cryogenic electron microscopy, there was only one structure of the human SGLT2 uploaded to the protein data bank, which was resolved with 2.95 Å and co-crystallized with empagliflozin at its binding site (Niu et al., 2021).

As a consequence of the limited availability of structures, the search for alternative options for the acquirement of more data for docking purposes led to the examination of AlphaFold (Jumper et al., 2021). AlphaFold is an AI system developed for the prediction of the 3D structure of proteins. It calculates the coordinates of all heavy atoms of a protein using the primary amino acid sequence and the aligned sequence of homologues as inputs. The predictions computed by AlphaFold are freely available and cover the complete human proteome as well as the proteome of 47 other organisms (Jumper et al., 2021). Over the course of this thesis, the performance of the structure of SGLT2 predicted by AlphaFold was compared to the cryo-EM structure and its usefulness in docking studies was evaluated.

2.2 Structure alignment and comparison

As one of the aims of this master thesis is to assess the properties of the AlphaFold structure, its differences to the structure obtained from cryo-EM, and its suitability for docking, alignments of the overall structures and binding sites of the two proteins proved to be a useful approach. Likewise to the comparison of ligand poses, which will be explained in chapter 2.3.4.2, similar protein structures or different

conformations of the same protein can be compared to each other by making use of the root-mean-square deviation (RMSD) (Mechelke & Habeck, 2010). Additionally, a visual inspection of the binding site and the contacts and interactions made there was performed.

2.2.1 Maestro

The Protein Structure Alignment panel found in Maestro, which is a graphical user interface used for molecular modelling and part of the Schrödinger Software Suite, allows the alignment of either multiple proteins or the substructures of these proteins. The default setting, which was used for this task, is to align all the amino acid residues of all proteins included in the Maestro Workspace. The additional feature of the Protein Structure Alignment panel to perform an alignment based on the binding site was also used (Schrödinger Release 2021-1: Maestro, 2021; "How do I align independent chains in two or more protein structures? | Schrödinger", 2016).

2.2.2 Molecular Operating Environment (MOE)

There is a number of options for the alignment of proteins found in the alignment function of MOE, which defines itself as an integrated computer-aided molecular design platform for small molecules, peptides and biologics (Molecular Operating Environment (MOE) 2020.09, 2020).

For the purpose of this task – and in order to ensure comparability with the results from Maestro – the settings were chosen to perform an alignment of only the amino acid sequence in the first step and a superposition based on the current alignment for all amino acids (called "structure alignment" in Maestro). In addition to the superposition of all amino acids, the same panel includes the option to superpose the amino acids of the binding site of the proteins based on the current alignment (Molecular Operating Environment (MOE) 2020.09, 2020).

2.3 Molecular docking

Molecular docking is a widely used computational tool for the study of molecular recognition and aims to predict the binding mode and binding affinity resulting from the establishment of a complex of two molecules (Huang & Zou, 2010). Protein-ligand docking represents one of the most important types of docking as it has become a crucial part of drug discovery processes and modern structure-based drug design (Du et al., 2016; Huang & Zou, 2010). Molecular docking, especially its use in virtual screening studies, has been successfully employed for the development and improvement of drugs several times (Sethi et al. 2020; Park et al. 2018; Dadashpour et al., 2015).

The importance of protein-ligand docking has led to the development of a variety of software packages using different algorithms for the placement and scoring of ligands. These methods consist of two steps, the first of which is the search algorithm responsible for searching through different conformations and orientations (poses) of the same ligand. The second step is called the scoring function and is used for the estimation of binding affinities and the ranking of different poses and different ligands (Du et al., 2016).

In theory, the search for protein-ligand binding should include all possible conformations of the protein and the ligand. However, this is not appropriate to be used in drug discovery as the computational expense hinders the application to a large number of compounds, which is often the case in drug discovery settings (Du et al., 2016). For this reason, usually various simplifications are employed in order to reduce computational time. However, the trend for search algorithms is still leaning towards the use of flexible-ligand (semi-rigid) or flexible ligand–flexible protein (induced fit) methods instead of the pure rigid-body algorithms (rigid) (Schrödinger Release 2021-1: Glide, 2021; Du et al., 2016).

The use of rigid receptors but flexible ligands is of particular usefulness in virtual screenings as it has shown the potential to reduce computational expense while still yielding satisfying results (Friesner et al., 2006; Du et al., 2016; Madhavi Sastry et al., 2013). In contrast, the use of an induced fit model in docking, which considers the flexibility of the protein, may be of usefulness when the binding of the ligand to the protein is believed to be dependent on the protein being induced into the correct

binding conformation for a ligand (Sherman et al., 2005; Madhavi Sastry et al., 2013). As both methods have shown their successful application and possess their respective strengths, both have been examined over the course of this project in order to analyse their abilities to assess the activities of SGLT2 inhibitors (Madhavi Sastry et al., 2013).

The docking protocols in this thesis were all provided by the Schrödinger Software Suite and all depictions in this thesis, unless otherwise clarified, were generated using Maestro, which is part of the Software Suite (Schrödinger Release 2021-1: Maestro, 2021).

2.3.1 Ligand selection for docking

2.3.1.1 Ligand selection for virtual screening

Because of the nature of virtual screening workflows and other computational methodologies to provide an *in silico* prediction that is yet to be proven experimentally, there is a need for a validation of the results. This validation can either take place in an *in vitro*, *in vivo* or *in silico* setting in order to prove the correctness of the predictions. *In silico* validation is often performed by screening a set of active compounds and a set of inactive compounds in parallel (Gimeno et al., 2019).

For this reason, it was necessary to obtain a library of compounds with a broad range of molecules to divide them into actives an inactives. ChEMBL is a manually curated database of bioactive molecules with drug-like properties and provides, among other data, the bioactivity data of such molecules. It allows the query of specific target proteins, on which the molecules that are shown as a result of the search have been tested on (Gaulton et al., 2016).

Therefore, ChEMBL proved to be an invaluable resource for retrieving molecules that have been tested on SGLT2. For this purpose, an in-house KNIME workflow (Preisach et al., 2008), which was initially created by members of the Pharmacoinformatics Research Group of the University of Vienna for retrieving and standardizing molecules from ChEMBL and internal datasets for machine learning tasks was adapted to provide molecules with known activity in an appropriate condition for

docking The initial, unchanged workflow has been uploaded to GitHub and is publicly available (Pharmacoinformatics Research Group, 2021).

This workflow provides a simple way to define the desired target protein and choose a threshold of the activity value, which is defined as the pChEMBL value. Above this threshold, the compound gets labelled as active by adding a column which holds the value 1 (active) or 0 (inactive). The output of the workflow includes a file in the SDF format containing the compound structures and their activity values as well as files in the CSV format with calculated descriptors, which were used for the machine learning tasks in the later parts of this thesis (Pharmacoinformatics Research Group, 2021).

After choosing the desired target by uploading an Excel file containing its ChEMBL ID to the workflow, compounds with missing pChEMBL values, activity values that are 0, and values that are not in an IC50 or Ki unit get excluded. Originally, the newest ChEMBL database version that was available for this workflow was set to ChEMBL 29. Therefore a few modifications had to be made in order to allow the use of the newest version of ChEMBL, which was ChEMBL30 (Gaulton et al., 2016).

Afterwards, there is an option for the standardization of the compounds. The settings of the node were changed in order to achieve an appropriate output for docking studies: the stereochemistry was kept as it is and molecules with nonorganic atoms were not removed.

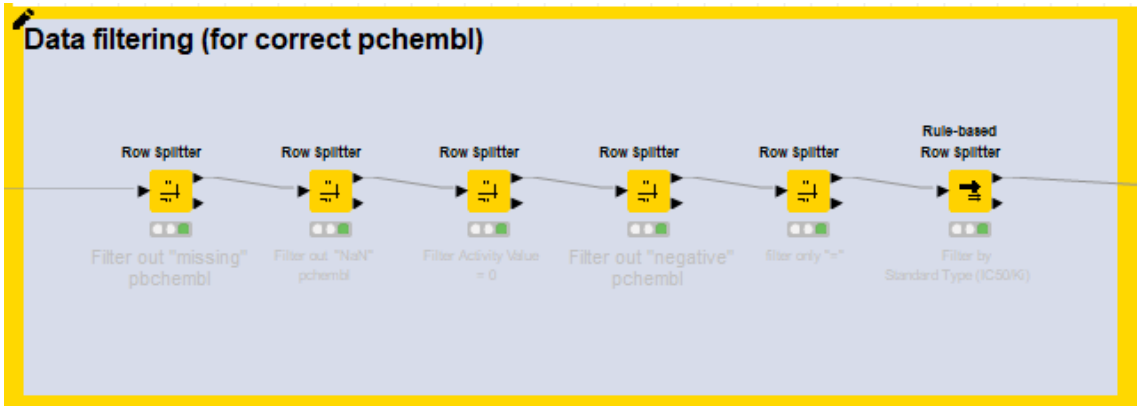


Figure 6: Filtering steps in order to retrieve all compounds with a correct pChEMBL value (Pharmacoinformatics Research Group, 2021).

The next step involves the addition of the activity classification (0 or 1) based on the chosen threshold for the activity values. During the same step, molecules presenting multiple entries because they ran through multiple tests, get merged into one entry

and filtered according to the properties of all the activity values: If all activity values lead to the same activity classification, the compound does not get excluded. If activity values lead to different activity classifications, the compound gets filtered out. Afterwards, the retrieved molecules get saved as a file in the SDF format, which then may be further used to calculate descriptors in the case of using this workflow for machine learning tasks (Pharmacoinformatics Research Group, 2021).

2.3.1.2 Ligand selection for individual docking

Because of the computational expense associated with induced fit docking, which takes the flexibility of the protein during the binding of a ligand into account, it is not possible to screen a significant number of ligands in order to validate the docking protocol (Du et al., 2016). Instead, an alternative route for the validation was taken: A literature search was performed to find a series of 7 congeneric SGLT2 inhibitors with known structure activity relationship, possessing a high difference in activity. In the past, this approach has been successfully applied for revealing binding hypotheses for propafenone type inhibitors of p-glycoprotein (Klepsch, Chiba & Ecker, 2011).

The ligands that were used were analogues of the SGLT2 inhibitor dapagliflozin and were utilized instead of empagliflozin and its analogues because of the plethora of available data on dapagliflozin and congeneric compounds (Braem et al., 2014; Lee et al., 2010; Ng et al., 2017).

Even though the ligands were obtained from three different publications, it had to be made sure to only include activity data from the same assay methodology, which was the intracellular accumulation assay of the SGLT2-selective [^{14}C]-alpha-methyl glucopyranoside (AMG). It was also ensured that the different assays yielded comparable results for dapagliflozin (Braem et al., 2014; Lee et al., 2010; Ng et al., 2017).

2.3.2 Preparation for docking

In the field of molecular docking, it is generally agreed upon that there is a need for proper preparation of the protein crystal structure as well as the ligands and it was shown that the neglection of certain preparation steps may lead to a significant drop

in the performance of virtual screening studies (Madhavi Sastry et al., 2013). Therefore, it was necessary to prepare the protein and the ligands for the docking studies.

2.3.2.1 Ligand preparation

In the case of ligands, the preparation involves the creation of 3D-geometries, the assignment of proper bond orders, and the generation of accessible tautomer and ionization states (Madhavi Sastry et al., 2013).

2.3.2.1.1 Virtual Screening Workflow

The Virtual Screening Workflow, which was used for the step of docking a large library of molecules with known activities to a rigid protein, is part of the Schrödinger Software Suite and provides a ligand preparation option as a part of the workflow. The ligand preparation is run as a LigPrep job and mostly uses the default settings of the LigPrep process (Schrödinger Release 2021-1: LigPrep, 2021).

Some of the options from LigPrep can be changed inside the workflow. However, the only change that was made in this case was the desalting of the ligands, which is a necessary step and is generally performed (Dhanjal et al., 2021; Omer et al., 2022; Schrödinger Release 2021-1: LigPrep, 2021). The possible ionization and tautomerization states at a target pH of 7.0 +/- 2.0 were generated using Epik, which uses the Hammett and Taft approaches for predicting pK_a values. This is necessary because the protonation state of a ligand impacts the conformations that get predicted for the molecule (Shelley et al., 2007; Madhavi Sastry et al., 2013; Schrödinger Release 2021-1: LigPrep, 2021). Additionally, Epik calculates an Epik state penalty which quantifies the energetic costs that are necessary to generate the states of the molecules. After docking, this penalty is combined with the GlideScore and results in a score that is called Docking Score which is used for the final ranking of different compounds (Madhavi Sastry et al., 2013).

At the end of the preparation of the ligands using the standard settings of the Virtual Screening Workflow, only 4 stereoisomers are retained, and one low energy ring conformation is generated (Schrödinger Release 2021-1: LigPrep, 2021).

2.3.2.1.2 Redocking/individual docking

As neither the Ligand Docking panel, which was used for the redocking of empagliflozin and the individual docking with the Extra Precision mode, nor the Induced Fit Docking panel possess the option to prepare the ligands inside the panel itself, it was necessary to pre-process them using the LigPrep panel. The LigPrep panel has the same functions as the above described ligand preparation step of the Virtual Screening Workflow but has some additional setting options. Most of the options were kept in the same way as in the Virtual Screening Workflow: The ligands were desalted and ionization/tautomerization states were generated at a target pH of 7.0 +/- 2.0 by using Epik. However, in contrast to the Workflow, the standard setting for the maximum number of generated stereoisomers is 32 (Schrödinger Release 2021-1 LigPrep, 2021).

2.3.2.2 Protein preparation

Similarly to the above-mentioned case of ligands, the preparation of proteins is a necessary step before the docking studies are started. This involves, among other steps, the addition of hydrogen atoms, the optimization of hydrogen bonds and the removal of atomic clashes (Madhavi Sastry et al., 2013). For this purpose, the Protein Preparation Wizard by Schrödinger was used (Schrödinger Release 2021-1: Prime, 2021).

As an addition to the default settings, the Protein Preparation Wizard allows to fill in missing side chains and loops as well as the capping of termini. These options, while not necessary for the AlphaFold structure, were applicable to the case of the cryo-EM structure of SGLT2 and were used to minimize the potential problems as shown by the View Problems function of the Protein Preparation Wizard panel. Afterwards, the structure was further refined by optimizing H-bonds and performing a restrained minimization (Schrödinger Release 2021-1: Prime, 2021).

In the case of redocking, where the Ligand Docking panel was used to perform the semi-rigid docking studies, it was necessary to define the binding site by utilizing the Receptor Grid Generation panel. The binding site was chosen by manually defining the co-crystallized ligand while keeping the default settings of the panel (Madhavi Sastry et al., 2013, Schrödinger Release 2021-1: Glide, 2021). In contrast,

both the Virtual Screening Workflow and the Induced Fit Docking protocol have an integrated option to choose a grid and therefore the binding site of the protein (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021).

2.3.3 Docking algorithms and workflows utilizing them

The docking algorithms available in the Schrödinger Software Suite which are not based on molecular dynamics simulations can be generally divided into semi-rigid and induced fit docking (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021). For semi-rigid docking, the Glide program is used while the induced fit docking uses both Glide and Prime (Friesner et al., 2004; Sherman et al., 2005).

While semi-rigid methods, where the ligand is considered flexible but the protein is not, are a suitable approach for the screening of large ligand libraries (Du et al., 2016; Schrödinger Release 2021-1: Glide, 2021), induced fit docking takes into account that the binding of different ligands may induce changes in the protein that are not considered in the semi-rigid approach (Du et al., 2016; Madhavi et al., 2013; Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021).

2.3.3.1 Docking and scoring algorithms

2.3.3.1.1 Semi-rigid docking – Standard Precision (SP)

Glide uses a docking method which approximates a search of the complete conformational, orientational, and positional space of the docked ligand. Initial rough positioning, scoring and refinement is followed by a torsionally flexible energy optimization in the field of the receptor on an OPLS grid and a further refining via Monte Carlo sampling (Friesner et al., 2004; Docking and Scoring | Schrödinger, n.d.).

Glide SP performs an exhaustive sampling, is recommended by Schrödinger as a balance between speed and accuracy, and takes about 10 seconds per compound (Docking and Scoring | Schrödinger, n.d.).

The different precision methods use different functional forms for the GlideScore (Docking and Scoring | Schrödinger, n.d.) and even though they use similar terms, they are formulated with different objectives in mind. Glide SP is a “softer” scoring

function which allows the identification of ligands with a reasonable potential to bind, even in cases of the pose having significant imperfections. Therefore, Glide SP is designed to minimize the number of false negatives and is appropriate for many database screening applications (Friesner et al., 2004).

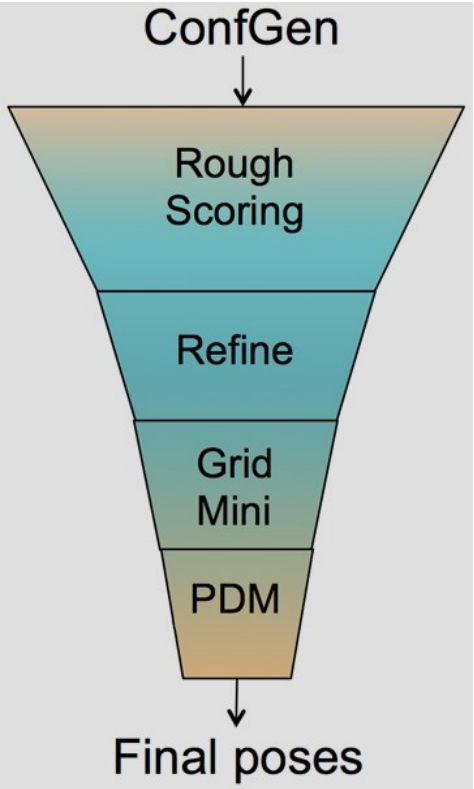


Figure 7: The docking process in Glide starts with the generation of a set of initial ligand conformations. The generated conformations are then roughly positioned, scored, and refined (first two steps of the depicted "docking funnel"). A small number of the best refined poses (about 400) are then minimized on an OPLS force field. Finally, a post docking minimization (PDM) is performed which uses the Monte Carlo sampling method to examine nearby torsional minima (Friesner et al., 2004; "Docking and Scoring | Schrödinger", n.d.).

2.3.3.1.2 Semi-rigid docking – Extra Precision (XP)

XP Glide sampling starts with the same methodology as the aforementioned SP Glide docking algorithm. However, it uses a wider "docking funnel" which provides a greater diversity of docked structures. It is necessary that SP produces at least one structure with a properly docked key fragment (Friesner et al., 2006).

Afterwards XP sampling attempts to build better-scoring poses by assigning various fragments of a molecule as anchors and starting from each anchor. The growing of sidechains from relevant positions of the anchors are initiated and unsuitable sidechains are rejected based on steric clashes, as the anchor fragment is already placed in the binding site and the sidechain building takes place there. Additionally, a rough scoring function is performed to screen the initial side chain conformations. The combination of constant pruning by screening and clustering algorithms leads to the creation of a high resolution sampling (Friesner et al., 2006).

After growing the individual sidechains, candidate structures are selected and minimized and a grid-based water addition is performed (Friesner et al., 2006). Finally, the scoring function for the GlideScore is computed, which is, compared to the SP scoring function, "harder" and applies severe penalties for poses that violate certain physical chemistry principles. The objective of the scoring function of XP is to minimize false positives and is recommended to be used in lead optimization or other docking studies with a small number of ligands (Friesner et al., 2004; Friesner et al., 2006).

2.3.3.1.2 Induced fit docking – Standard

In induced fit docking, not only the ligand, but also the protein is considered flexible. In the Schrödinger Software Suite this is realized by iterative combination of rigid receptor docking and protein structure prediction using the Glide and the Prime software, respectively. The induced fit docking process can be generally described by dividing it into four steps (Sherman et al., 2005):

1. The first step is characterized by docking the ligands into a rigid receptor using a softened energy function to allow more leeway for steric clashes.

2. For each ligand pose generated in the first step, the protein is sampled by sidechain rearrangements and minimization of the ligand-protein complex to allow for minor backbone movement.

3. Thereafter, a second round of ligand docking is performed into the induced fit structures of the previous step while using a hard potential function.

4. Finally, scoring is done by accounting for both the docking energy (GlideScore) and the receptor strain and solvation terms (Prime energy) (Sherman et al., 2005). This results in a scoring term called IFDScore (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021).

During the first step of the Standard Sampling method of the Induced Fit Docking panel a maximum of 20 poses are retained while using a softened energy function. Afterwards, during the third step, which is composed of a second round of docking with a hard energy function, it can be decided if an XP or a SP algorithm is used for the docking (Schrödinger Release 2021-1: Glide, 2021).

2.3.3.1.3 Induced fit docking – Extended Sampling

The Extended Sampling method of the Induced Fit Docking protocol follows the same general steps as the above-mentioned Standard method. However, it performs an initial docking step, which not only involves the use of a softened potential, but also the removal of protein sidechains. The removal is decided upon using properties like solvent accessible surface areas, and up to 80 docking poses are retained after multiple docking runs, some of which use a trimmed receptor while others use an untrimmed one with softened potentials. The results from the docking runs are clustered to obtain representative poses (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021).

During the second docking step with a hard energy function, the Extended Sampling protocol automatically runs the SP algorithm and, unlike for the Standard Sampling method, there exists no option to choose the Extra Precision algorithm (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021).

2.3.3.2 Workflows and panels

The Schrödinger Software Suite provides a number of panels and workflows that use different algorithms that will be further described in the following part of this chapter, which is to provide an overview of these panels and the settings that were

utilized during this thesis. The used panels were the Ligand Docking Panel, the Virtual Screening Workflow, and the Induced Fit Docking panel (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021).

2.3.3.2.1 Ligand Docking panel

The Ligand Docking panel uses the semi-rigid docking methods, where the user can decide between a High Throughput Virtual Screening (HTVS) precision, a Standard Precision (SP), and an Extra Precision mode. For this thesis, the previously described Standard and the Extra Precision modes were used (Schrödinger Release 2021-1: Glide, 2021; Friesner et al., 2006; Friesner et al., 2004).

The settings for the Ligand Docking panel provide a number of options, all of which were kept at their default, except for the number of poses to report which was set to 32 in order to allow an overview over the possible ligand poses. This panel was used for the redocking step, where both the SP and XP mode were used, as well as the individual docking step of this thesis, where the XP mode was used (Schrödinger Release 2021-1: Glide, 2021).

2.3.3.2.2 Virtual Screening Workflow

Similarly to the Ligand Docking panel, the Virtual Screening Workflow provides the option to use semi-rigid docking methods, for which it can be decided if HTVS, SP or XP modes are used. Additionally, different filtering steps may be applied, where each of the filtering modes represents one of the filtering steps (i.e., 10% of the ligands with the best Docking Scores of a HTVS run are kept, of those 10% another 10% are kept after a SP run and of those 10% another 10% are kept after a XP run). The Virtual Screening Workflow possesses an integrated ligand preparation step as well as a binding site selection tool, which are necessary steps for docking (Madhavi Sastry et al., 2013; Schrödinger Release 2021-1: Glide, 2021).

For the purpose of this master thesis, the settings of the Virtual Screening Workflow were mainly kept at their default. However, three changes were made: during the ligand preparation, the desalting of the ligands was performed, as this is a necessary step and usually performed in docking studies (Dhanjal et al., 2005; Omer et al.,

2022; Schrödinger Release 2021-1: Prime, 2021). The second change introduced was to perform the screening by using the SP mode, which is the recommendation for screening purposes because of its property to possess a “softer” and more tolerant scoring function (Friesner et al. 2004; Schrödinger Release 2021-1: Glide, 2021). Finally, the filtering was disabled to retrieve a complete list of the docked ligands in a ranked order after the screening (Schrödinger Release 2021-1: Glide 2021).

2.3.3.2.3 Induced Fit Docking panel

The Induced Fit Docking panel uses the induced fit protocols, where the protein and the ligand are both considered to be flexible, and allows the choice between a Standard Sampling and an Extended Sampling, which changes the number of reported poses (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021; Docking and Scoring | Schrödinger, n.d.).

For the Standard mode of the Induced Fit Docking protocol, the settings were kept at their default, except for the Glide Redocking option, which is part of the panel and was changed to the Extra Precision mode. For the Extended Sampling mode, no changes were made to the default parameters (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021). These settings were used in the following steps of this thesis, whenever the Induced Fit Docking panel was employed.

2.3.4 Redocking and its analysis

2.3.4.1 Redocking

Prior to any large-scaled docking studies it is important to validate the abilities of a docking procedure. One of the methods that have been reported for this purpose is to analyse the ability of the program to recreate the original, native pose of the co-crystallized ligand. The results of a redocking validation can be analysed by comparing the top ranked poses of the docking process to the initial pose of the co-crystallized ligand (Mateev et al., 2022; Cole et al., 2005).

For the redocking studies, using both the cryo-EM and, when possible, the AlphaFold protein structure, all possible docking algorithms were used: The Standard Precision and Extra Precision modes using the Ligand Docking panel, and the Standard and Extended Sampling using the Induced Fit Docking panel (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021).

The results of the Ligand Docking panel may be regarded as representative for the results of the Virtual Screening Workflow as it mostly uses the same default settings and allows the choice between the use of the same precision modes as the ones used during the Virtual Screening Workflow (Schrödinger Release 2021-1: Glide, 2021).

2.3.4.2 Analysis of redocking studies

The first step of a molecular docking study involves the validation of the docking accuracy, virtual screening utility, or scoring accuracy. However, there are no established standards regarding this critical step in the docking process (Jain, 2007).

One of the possible solutions consists in measuring the pose-prediction success rates of the chosen algorithm to recreate the original pose of a co-crystallized ligand (Mateev et al., 2022; Cole et al., 2005). This procedure is called redocking, and a common metric used to measure the distances of the predicted pose and the native ligand pose is the root-mean-square deviation (RMSD) of a redocked pose to the given pose of the co-crystallized ligand (Cole et al., 2005; Bell & Zhang, 2019).

For the calculation of the RMSD of the docked poses, Schrödinger provides a number of options and panels, one of which is the Superposition panel, accessible through the Maestro software. By choosing the "Compute without changing structure" option, the superposition of the structures is done without moving the structures and the RMSD is calculated for the current set of atoms in their existing positions. The chosen method for the superposition was a superposition based on the ligand substructures (Schrödinger Release 2021-1: Maestro, 2021).

The RMSD, based on the heavy atoms of the compared ligand conformations (Schrödinger Release 2021-1: Maestro, 2021), is calculated the following way:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}$$

Figure 8: Equation for the calculation of the root-mean-square deviation, where N is the number of atoms in the ligand and d_i is the Euclidean distance between the i^{th} pair of corresponding atoms (Bell & Zhang, 2019).

However, the exclusive use of the RMSD is not recommended, and an additional careful interpretation of the resulting numerical measures through interaction-based measures and visual inspection is seen as necessary. This is needed because docking programs may present solutions that have a low RMSD but form interactions with the protein different from the experimentally observed ligand (Cole et al., 2005).

The definition of the cut-off for a good result is frequently set at a RMSD of 2 Å and a RMSD of 2-3 Å as partial success (Cole et al., 2005). For example, Mateev et al. have defined values below 1 Å as excellent, from 1-2 Å as good, from 2-3 Å as moderate, and above 3 Å as wrong/incorrect (2022). The choice for these cut-offs is arbitrary but has found common use (Cole et al., 2005).

2.3.5 Virtual Screening and its analysis

In typical drug discovery settings, the main stages include target selection, hit identification, lead optimization, and preclinical and clinical studies. Two of those steps, hit identification and lead optimization, are intertwined with computational modelling, which includes structure based virtual screening (Kontoyianni, 2017). The advantage of virtual screening over *in vitro* high-throughput screenings is the possibility to process thousands of compounds in short time and therefore reduce the number of compounds that are afterwards used for *in vitro* testing. This procedure reduces the costs significantly (Gimeno et al., 2019).

Virtual screening has seen multiple successes over the past years (Sethi et al., 2020) and is based on the idea of docking a library of small compounds into the binding pocket of a protein (Kontoyianni, 2017). A fraction of the ranked compounds is then

moved forward toward hit identification. The underlying principle is the assumption that the virtual screening is able to differentiate between active and inactive compounds and to score all actives at the top of the returned list. As the success rates of screening methods are low and the goal of them is not to find all hits, but only a sufficient number of possible scaffolds for initial discovery efforts, it is necessary to reconfirm the results (Kontoyianni, 2017).

The Virtual Screening Workflow, which was described in the previous chapters (2.3.3.2), was used for the assessment of the ability of Glide to provide a potentially successful virtual screening framework (Schrödinger Release 2021-1: Glide, 2021). This workflow was applied to the protein structure that was obtained by cryo-EM experiments (Niu et al., 2021; Schrödinger Release 2021-1: Glide, 2021).

2.3.5.2 Analysis of virtual screening runs

As previously described, virtual screening workflows consist in computational methodologies, which result in predictions as the output of the screenings. These predictions need to be validated both *in silico* and *in vitro* or *in vivo*, which is often done by screening a library that includes known active molecules as well as known inactives (or decoys). Active compounds are compounds that have been reported to possess a certain activity towards the target protein. As these *in vitro* assays are reporting a range of activity values, the threshold over which the compounds are considered to be active is arbitrary. However, compounds are usually considered to be active when they fall into a micromolar and nanomolar range: the higher the activity threshold for active compounds, the more restrictive the virtual screening. Similarly, the inactive compounds are the ones that have been reported to have a low activity towards the target protein in *in vitro* test. (Gimeno et al., 2019).

Afterwards, a number of statistical measures are calculated in order to assess the performance of the virtual screening and its ability to tendentially rank the active compounds above the inactives (Gimeno et al., 2019; Schrödinger Release 2021-1: Glide, 2021).

The most important statistical characteristics of a virtual screening that have to be inspected in order to judge the performance are called enrichment metrics. Glide enables to calculate these metrics by using the Enrichment Calculator Panel, for

which the output of a screening run, a file containing only the screened active compounds, and the number of decoys/inactive compounds is needed as an input (Schrödinger Release 2021-1: Maestro, 2021; Gimeno et al., 2019).

The enrichment metrics calculated by the Enrichment Calculator Panel are the Boltzmann-enhanced Discrimination Receiver Operator Characteristic (BEDROC) area under the curve, Receiver Operator Characteristic (ROC) area under the curve, Area Under the Accumulation Curve (AUAC), a number of (modified) enrichment factors, Efficiency in distinguishing actives from decoys/inactives (Eff), and the average Fraction of Outranking Decoys (FOD). Additionally, its output includes a plot of the ROC curve and a % Screen Plot, which shows the percentage of actives recovered against the percentage of structures screened (Schrödinger Release 2021-1: Maestro, 2021).

The calculations of some of the various enrichment metrics that were used for the analysis of the virtual screenings during this thesis are performed as described in the following section of this chapter.

As docking produces a continuous output (Docking Score), different thresholds can be applied to the score to produce a discrete (binary) classifier in order to predict the class membership of the screened compounds. In this case, the two classes are the active compounds and the inactive compounds (Schrödinger Release 2021-1: Glide, 2021; Fawcett, 2006).

The ROC curve is depicted by plotting the false positive rate against the true positive rate for various thresholds (Fawcett, 2006) and The ROC area under the curve is typically described as the probability of an active appearing before an inactive (Schrödinger Release 2021-1: Maestro, 2021; Fawcett, 2006).

In virtual screening, the ideal case is the ranking of the active compounds at the top of the screening output. However, realistically this is never the case (Kontoyianni, 2017). Because of the nature of screening methods of being used to move forward a fraction of the top results of a screening run toward hit identification (Kontoyianni, 2017), it is necessary to judge the ability of the used method to enrich a sufficient number of actives at the top of the output (Truchon & Bayly, 2007). The key requirement for success is therefore that it must rank actives very

early in the larger set of compounds, an ability towards which the ROC metric is not sensitive (Truchon & Bayly, 2007).

For this reason, the Enrichment Factor was chosen as an additional metric for the analysis of the results. The Enrichment Factor is a measure of how much the sample is enriched after a filter or a series of filters is applied and is determining how many more actives are found within a defined fraction (for example, the top 10% of the screening results) relative to a random distribution (Truchon & Bayly, 2007; Gimeno et al. 2019).

The Enrichment Factor is defined as:

$$EF = \frac{a/n}{A/N}$$

a is the number of actives found in sample size n and A is the total number of actives found among the total number of ligands N (Schrödinger Release 2021-1: Maestro, 2021). The advantages of this method are its capability to compare the enrichment to a random selection of compounds and not to weigh all compounds equally. One of the disadvantages is that the Enrichment Factor is weighing all compounds within the cut-off equally, which means that it is not able to distinguish between algorithms that have an equal Enrichment Factor but different rankings within the cut-off (Truchon & Bayly, 2007).

A different way to illustrate the early enrichment is provided by the Enrichment Calculator Panel, which is able to display a plot of the percentage of actives recovered against the percentage of structures screened and is therefore a visual representation of the Enrichment Factor while not being bound to certain cut-offs for the weighing of compounds (Schrödinger Release 2021-1: Maestro, 2021).

2.3.6 Individual docking and its analysis

Induced fit docking, which considers not only the docked ligand, but also the protein as flexible, is associated with a significant computational expense and is therefore not suited for the use in virtual screening settings (Du et al., 2016). Similarly, the

Extra Precision mode in Glide is computationally more demanding than the Standard Precision, which was used for the virtual screening tasks of this master's thesis (Friesner et al., 2016).

Consequently, a solution had to be found in order to be able to evaluate these docking modes for their use in scoring and enriching in drug discovery of SGLT2 inhibitors. For this purpose, a series of 7 congeneric SGLT2 inhibitors with significant differences in their activity were docked by using the above-mentioned docking methods.

2.3.6.1 MM-GBSA

As an addition to the scoring functions of the aforementioned docking protocols, which are the Docking Score and the IFD scores, the best complexes of the best performing protocol were further minimized using the MM-GBSA methodology. The resulting MMGBSA dG Bind energies may be used to estimate the binding affinities, and it is claimed that the ranking based on the calculated binding energies can be expected to agree with the ranking based on experimental binding affinity (Genheden & Ryde, 2015; Schrödinger Release 2021-1: Prime, 2021; "Can I relate MM-GBSA energies to binding affinity? | Schrödinger", 2015). For this reason, it was decided to use the MMGBS dG Bind energies for the analysis of the ranking scores.

The space surrounding the ligand, within which the minimization is conducted, was set at 12 Å, the solvent was defined as water, and the force field was changed into the force field used by the IFD protocol, which is OPLS4 (Schrödinger Release, 2021-1: Prime, 2021; Lu et al., 2021; Grillberger 2022).

2.3.6.2 Interaction fingerprint clustering

As the validity of the docking pose of ligands is oftentimes not only dependent on their binding pose orientation but also on the interactions they form with the protein, the Interaction Fingerprints Panel by Schrödinger came in as a handy tool (Cole et al., 2005; Schrödinger Release 2021-1: Maestro, 2021). All the available interaction types of the ligand with the protein within the possible distances were chosen for the calculation of the interaction fingerprint and the settings for the distances

were kept at their default parameters. The interaction fingerprints of the ligands and their poses were then clustered by using the average linkage method (Grillberger, 2022; Schrödinger Release 2021-1: Maestro, 2021).

2.3.6.3 Rank correlation

Even though scoring functions resulting from docking studies are not accurate enough to predict differences in the binding affinity of compounds that have different structures and different predicted binding modes, this may be possible in some cases for structurally similar compounds with the same binding mode. This was the case for the used SGLT2 inhibitors, which was the reason for an attempt at correlating the results of the docking runs with their reported activities. (Gimeno et al., 2019).

In order to correlate the results of the docking runs with the activity of the ligands, it was necessary to convert their activities and their Docking Scores to ranks which allows to calculate the Spearman correlation coefficient for ranked data. The Spearman correlation is used to determine the correlation between two sets of rankings (Myers & Well, 2002).

The Spearman correlation possesses a number of advantages over correlation calculations between two quantitative variables such as the Pearson correlation coefficient, which is a measure of the extent to which two variables are linearly related, while the Spearman correlation is the Pearson correlation applied to ranks. The conversion of scores to ranks is a reasonable choice if it is not assumed that equal differences between the scores necessarily correspond to equal differences in the underlying variable that is measured. Additionally, it is more robust to the effect of outliers and was therefore used in this case (Myers & Well, 2002). The Spearman rank correlation is calculated in the following way for non-tied ranks:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}$$

Figure 9: Calculation of the Spearman rank correlation coefficient where n is the number of pairs of measurement and D_i is the difference of i^{th} pair of the rankings ("Spearman rank correlation coefficient" n.d.; Myers & Well, 2002).

The equation in figure 9 should not be used if there are tied ranks, and there exist modifications to the equation that can adjust for ties (Myers & Well, 2002).

2.4 Machine learning based QSAR modelling

If the three-dimensional structure of a target protein is known, the use of structure-based approaches, some of which have been mentioned during the previous chapters, are recommended for virtual screenings. However, ligand-based approaches may also show effectiveness, especially in cases where structure-activity relationship (SAR) studies have been conducted (Gimeno et al., 2019).

Therefore, the combination of ligand- and structure-based approaches has the potential of being of particular usefulness in identifying compounds which share critical structural characteristics for the presence of activity with structures with known activity, while also taking into account their compatibility with the receptor (Gimeno et al., 2019). For this reason, it was decided to use a ligand-based approach to inspect and perform the prediction of SGLT2 inhibitors as an addition to the discussed structure-based approaches.

Quantitative Structure-Activity Relationship (QSAR) is a popular ligand-based approach that is employed in order to correlate chemical molecules with their biological and pharmaceutical activities based on their chemical structure (Keyvanpour & Shirzad, 2021; Shahlaei, 2013). This is done by calculating mathematical descriptors which are encoding molecular structures and properties in QSAR studies, like topological descriptors, constitutional descriptors, or functional groups (Shahlaei, 2013; Muratov et al., 2020). These descriptors are used to find a reliable relationship between the calculated values and the biological activity for a series of compounds

(Shahlaei, 2013; Danishuddin & Khan, 2016). This would then result in a model which could be used to assess the activity of new chemical entities (Shahlaei, 2013; Danishuddin & Khan, 2016).

Over the years, QSAR modelling has become a tool for virtual screening, used to develop models that are able to screen large databases and find probabilities of the molecules having activity against a protein (Barros et al., 2020).

The use of machine learning and QSAR modelling has been progressively evolving over the last decades and are now closely related fields of work. Machine leaning methods are a subfield of computer science and have emerged from the study of pattern recognition and the theory of computational learning in artificial intelligence. A short definition of machine learning describes it as the study to develop algorithms that are able to learn from their errors and to generate predictions about data by using a sample input to construct a model (Barros et al., 2020).

Machine learning models can be categorized by the desired output of the model. Classification models are characterized by entries that are divided into two or more classes and the produced model is trained by data with known classes. Another type of machine learning model are the regression models, for which the outputs are continuous or discrete. Both of these types are solved by supervised learning (Barros et al., 2020).

There are multiple machine learning algorithms used for QSAR including Random Forest, Support Vector Machine, k-nearest Neighbors, Artificial Neural Networks, naïve Bayes classifiers and logistic regression (Li et al., 2007; Mitchell, 2014; Barros et al., 2020).

For the purposes of this thesis, two publicly available machine learning-based jupyter notebooks provided by the Pharmacoinformatics Research Group were used to build classification models for SGLT2 inhibitors. The notebooks are available on Github (Pharmacoinformatics Research Group, 2021; Pharmacoinformatics Research Group 2022) and will respectively be referred to as the Sandbox and the Retraining notebooks throughout this thesis. The machine learning algorithms that were employed in the codes of these two notebooks are Random Forest, Support Vector Machine, k-Nearest Neighbors, and Logistic Regression. The used algorithms are explained below.

Random Forest

In Random Forest (RF) algorithms, a large number of decision trees are generated and at the end a vote for the most popular case will be held. Decision trees are diagrams that allow the representation of problems involving sequential decisions and highlighting the risks and results during each decision. The decision trees work from top to bottom and choose a variable in each step that best divides the data (Barros et al., 2020).

Support Vector Machine

In Support Vector Machine (SVM) algorithms, the training data is non-linearly mapped to a high-dimension feature space in which a linear decision surface is constructed (Cortes & Vapnik, 1995). In short, a SVM model represents the training data as points in space in a way that leads to the points of different categories separated by a gap as wide as possible. Data with unknown categories will then be mapped in the same space and their category will be predicted depending on the side of the gap they fall into (Barros et al., 2020).

k-Nearest Neighbors

In k-Nearest Neighbors (KNN) algorithms, the output of a classification model is classified by a plurality of votes of its nearest *k* neighbors where *k* is an integer value and positive. So, the object gets assigned to the class which is most common among its neighbors with *k* determining the number of neighbors used for this classification (Barros et al., 2020).

Logistic Regression

Logistic Regression (LR) is a classification algorithm which assumes the decision boundaries to be linear. Each weighted feature vector from the training data is mapped to a value between 0 and 1 through the S-shaped logistic function and this value is interpreted as the probability of an example belonging to one of the classes. The learning algorithm tunes the weights to classify the training data correctly (Gudivada et al., 2016).

2.4.1 Sandbox notebook

The PharminfoVienna Sandbox is a tool that allows the user to gather data and calculate descriptors by using a KNIME workflow, that has been partially described in the ligand selection section of the docking chapter of this thesis (2.3.1.1), and building classification models by making use of a Jupyter notebook for machine learning based QSAR models (Pharmacoinformatics Research Group, 2021).

The output of the KNIME workflow, which requires the setup of a docker and an appropriate Python environment, provides a file in the SDF format with the standardized molecules and their classification (into actives and inactives) depending on their pChEMBL value and if it crosses the set threshold. From this SDF file, the workflow calculates a number of RDKit descriptors and adds CSV format files to the output folder. This includes a training and a test set as well as a file for all compounds containing the descriptors of the molecules (Pharmacoinformatics Research Group, 2021).

The Jupyter Notebook is employed for the model building, using the programming language Python, and accessing the notebook also requires the setup of a docker. The target list in the Jupyter Notebook has to be modified in order to include the desired targets, which was SGLT2 in this case, and the input folder has to include the four CSV files provided by the KNIME workflow (Pharmacoinformatics Research Group, 2021).

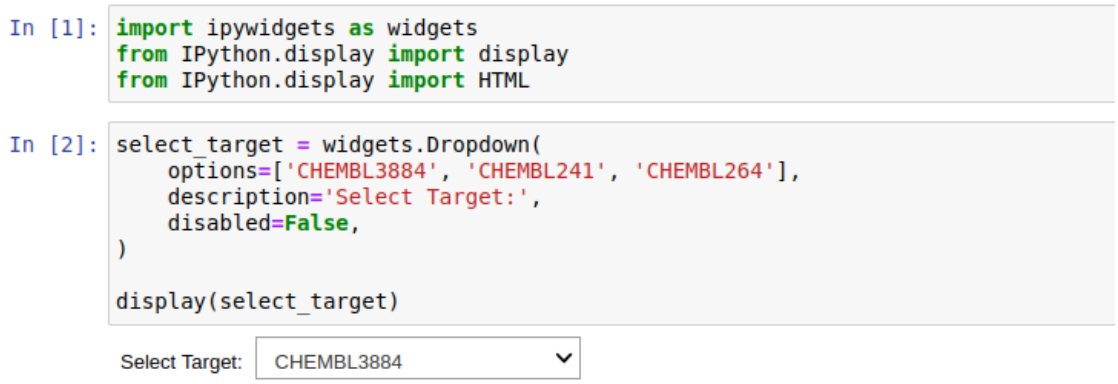


Figure 10: The first step to be done in the Jupyter Notebook requires the selection of the desired target by typing the ChEMBL IDs of the targets into the "options"-brackets. In this case, this was CHEMBL3884, which is the ChEMBL ID of SGLT2 (Pharmacoinformatics Research Group, 2021).

Furthermore, a grid-search for the hyperparameters of the classifiers must be performed, as the default parameters in the notebook are only a trivial example (Pharmacoinformatics Research Group, 2021). The parameters chosen for the grid-search of this thesis were as follows:

For Support Vector Machine:	C: 0.01, 0.1, 1.0, 10.0, 100.0
	Gamma: 0.01, 0.1, 1.0, 10.0, 100.0
For Random Forest:	n_estimators: 10, 25, 50, 75, 100, 200, 300
	max_depth: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20
For k-nearest Neighbors:	n_neighbors: 3, 5, 7, 9, 11, 13, 15, 17, 19

The classifiers used in the notebook are three of the previously discussed classification algorithms: Random Forest, Support Vector Machine, and k-nearest Neighbors. For the validation and evaluation of the models using the test set, a number of statistical metrics are provided as well as a Confusion Matrix, a Precision-Recall Curve, and a ROC curve. The best performing models from each of the classifiers are automatically chosen, but the other models are also included in the output of the code (Pharmacoinformatics Research Group, 2021).

2.4.2 Retraining notebook

Similarly, the Retraining notebook is a tool that provides the user with the option to generate machine learning based models for the classification of compounds by using a Python coded Jupyter Notebook. Like the Sandbox notebook, this notebook includes a section for the standardization of molecules and the calculation of descriptors. However, the standardization of the molecules was carried out by using the KNIME workflow of the PharminfoVienna Sandbox in order to ensure comparability with the output of the Sandbox notebook (Pharmacoinformactics Research Group, 2022).

Afterwards, the SDF file was used for the creation of a training and a test set by modifying the Retraining notebook and a number of RDKit descriptors were calculated. The activity classification was also provided by the KNIME workflow, although the

column name had to be changed in order to be used by the notebook (Pharmacoinformatics Research Group, 2022).

For the generation of the models, all four discussed classifiers were used: Random Forest, Support Vector Machine, k-nearest Neighbors, and Logistic Regression. Analogously to the Sandbox notebook, statistical metrics are provided for the evaluation of each of the models. Despite the parallels shown between the two notebooks, additional modifications had to be made to the script to create comparable outputs between them because of certain differences of the two codes (Pharmacoinformatics Research Group, 2022).

As the Retraining notebook was created for the generation of models for certain transporter proteins, the hyperparameters were already pre-selected. However, SGLT2 was not among the transporter proteins for which the notebook was created, which raised the necessity to modify the notebook for the implementation of a grid-search. The same parameters for the grid-search of the models were chosen as for the grid-search used in the Sandbox notebook (Pharmacoinformatics Research Group, 2022).

2) Support Vector Machine (SVM)

In [40]:
```
from sklearn import svm
from sklearn.model_selection import GridSearchCV
```

In [41]:
```
# defining parameter range
param_grid = {'class_weight': ['balanced'],
              'C': [0.01, 0.1, 1.0, 10.0, 100.0],
              'kernel': ['linear','rbf','poly','sigmoid']}

grid = GridSearchCV(svm.SVC(), param_grid, refit = True, verbose = 0)

# fitting the model for grid search
grid.fit(descriptors, activities)
```

Figure 11: The Retraining notebook provides pre-selected hyperparameters for 6 targets. A grid-search had to be implemented, because SGLT2 is not among the targets for which the hyperparameters are pre-selected. The figure shows the code which realises the grid-search, which was implemented using scikit-learn (Pedregosa et al., 2011)

Unlike the Sandbox notebook, the Retraining notebook does not provide a Confusion Matrix or a ROC curve for the evaluation of each model, which was also added as a modification to the code provided on GitHub (Pharmacoinformatics Research Group group, 2022).

```
In [57]: from sklearn.metrics import confusion_matrix
         from sklearn.metrics import ConfusionMatrixDisplay
         import matplotlib.pyplot as plt
         ClassificationSVM = df_svm_prediction["Classification"]
         PredictionSVM = df_svm_prediction["Prediction"]
         conf = confusion_matrix(ClassificationSVM, PredictionSVM)
         display = ConfusionMatrixDisplay(conf).plot()
         plt.show()
```

Figure 12: The Retraining notebook does not provide Confusion Matrices and ROC curves in order to allow a visual evaluation of the computed models, which is the reason for the addition of the shown modifications. The figure shows the code for the Confusion matrix of the Support Vector Machine model using scikit-learn and Matplotlib (Pedregosa et al., 2011; Hunter, 2007).

2.4.3 Evaluation of the model performances

Retrospective validation of machine learning models and challenging the potential utility of the models for predictions by applying them to data that has not been part of the model building process is recommended to be standard practice. This data is usually called external test set and the validation is carried out through cross-validation. Bender et al. recommend computing an array of statistical metrics for the validation instead of a single user-defined one and to use them as an ensemble to make use of their complementary nature (2022).

Both notebooks provide multiple metrics: Accuracy, Sensitivity, Specificity, Balanced Accuracy, F1 score, ROC AUC, Precision, Matthews Correlation Coefficient (MCC), and Recall were part of the validation sections of the notebooks and their overall results were used for the evaluation of the performances over different activity thresholds and different classifiers (Pharmacoinformatics Research Group 2021; Pharmacoinformatics Research Group 2022).

2.5 Ligands with unknown activity

The aim of this thesis is to assess and provide different methods for the *in silico* prediction of the activity of ligands with unknown *in vitro* and *in vivo* activity.

These methods were then applied to a series of in-house potential inhibitors of SGLT2 with unknown activity that were synthesized and designed by the Pharmaceutical Chemistry subdivision of the University of Vienna (Kirchweger, Rollinger & Kowalska, 2022).

Overall, 89 compounds had been synthesized, some of which displayed activity against the nematode C. elegans. However, C. elegans does not possess the SGLT2 and it was therefore not possible to deduct the ability to bind to SGLT2 from the activity data stemming from tests on C. elegans (Kirchweger, Rollinger & Kowalska, 2022).

As the ligands will potentially be patented, it is not possible to present or describe their molecular structure in this thesis. However, the ligands will be referred to by their codes in order to make the comparison of the results of the different *in silico* methods possible.

3 Results

3.1 Structure alignment

As one of the aims of this thesis is to evaluate the suitability of the SGLT2 structure provided by AlphaFold for docking purposes, the characterization of the binding site and its comparison to the binding site provided by the cryo-EM resolved structure proved to be a necessary step. This was achieved by carrying out a superposition of the two SGLT2 structures, which was performed both by using Schrödinger and MOE (Schrödinger Release 2021-1: Maestro, 2021; Molecular Operating Environment (MOE) 2020.09, 2020). The output was inspected by calculating the RMSD and performing visual inspections of the binding site as explained in chapter 2.2.

3.1.1 Superposition and binding site comparison

The superposition of the two proteins resulted in relatively similar RMSDs in both programs, 1.193 Å in Schrödinger and 1.210 Å in MOE, which correlated with a visual inspection, which allowed to conclude that the backbones of the two structures show a high similarity. However, a superposition based on the binding site resulted in two very different outcomes for the two programs. While MOE showed a RMSD of 1.202 Å for a superposition based on the binding site, the superposition panel in Schrödinger returned a RMSD of 6.891 Å, which is such a significant difference that it prompted a visual inspection of the binding sites regarding possible differences.

The results of the visual inspection led to the conclusion that, while the protein backbone seems to be calculated accurately by AlphaFold, the amino acid orientation at the binding site shows considerable differences which could lead to the arising of issues during the rigid docking processes. For this reason, the interactions and clashes of the co-crystallized empagliflozin with the AlphaFold structure were assessed. This was achieved by superposing the AlphaFold structure with the cryo-EM resolved structure and inspecting the residues within 5 Å of the ligand for clashes and contacts with the ligand. It became obvious that the differences in the orientation of the amino acids led to a high number of clashes with empagliflozin, which was the case for both programs and for both the overall structure superposition and

the superposition based on the biding site. In the following paragraph, an explanation of the observed clashes after performing a superposition based on all amino acid residues is presented as an example for the results of both settings and both programs.

As can be seen in figure 13, the co-crystallized empagliflozin shows a number of clashes and bad contacts with the AlphaFold structure. One of the amino acids involved in the clashes is aspartic acid 454, which clashes with the tetrahydrofuran substructure of empagliflozin. The backbone of phenylalanine 453 also shows clashes with the tetrahydrofuran substructure and, in addition to that, its aromatic sidechain shows bad and "ugly" contacts with the distal benzene ring of empagliflozin. Threonine 87 clashes with the oxygen, which serves as a linker between the distal benzene ring and the tetrahydrofuran, and glycine 83 and leucine 84 also show contacts with the distal benzene. Histidine 80 shows bad contacts with the proximal benzene ring and the carbon linker between the proximal and the distal benzene ring. Finally, serine 287, aspartic acid 75, and glutamic acid 457 show clashes with the glucose moiety.

Figure 13: Empagliflozin in its pose, as elucidated via cryo-EM, inside the binding site of the superposed AlphaFold structure. The picture was generated by using Maestro, where there is an option to show various interactions by using the interactions panel. It can be used to visualize non-covalent bonds, pi interactions, and contacts/clashes. In this case, there are no non-covalent bonds or pi interactions, but a number of "bad" and "ugly" contacts/clashes, which are indicated by dotted lines (orange = bad and red = ugly). Carbon atoms of the ligand are shown as blue, oxygen atoms as red, and chloride as green (Schrödinger Release 2021-1: Maestro, 2021).

3.2 Redocking

3.2.1 Characteristics of the binding site

An overview of the overall- and binding site characteristics of SGLT2 as elucidated with the help of cryogenic electron microscopy (Niu et al., 2021) was briefly presented in the introductory chapter 1.4.1. As the following chapters will include the analysis of the redocking of empagliflozin and the docking of various ligands into the binding site, the interactions of empagliflozin with SGLT2 at its binding site and the pose shown in the cryo-EM structure are introduced here in order to allow a comparison with the docking results.

As can be seen in Figure 14, the ligand shows a number of hydrogen bonds at the sugar binding site of SGLT2, where various amino acids make contact with the hydroxy groups of the ligand. Additionally, the two benzene rings at the hydrophobic

tail of the ligand show pi-pi interactions with two aromatic amino acids. These interactions are usually defined as the attractive interaction between two parallel or face-to-face oriented aromatic systems (Fang, 2013). The amino acids involved are phenylalanine 98, which interacts with the distal benzene ring, and histidine 80, which interacts with the proximal ring.

The binding of SGLT2 inhibitors involve the binding of sodium, but the sodium binding site has not been determined yet (Niu et al., 2021).

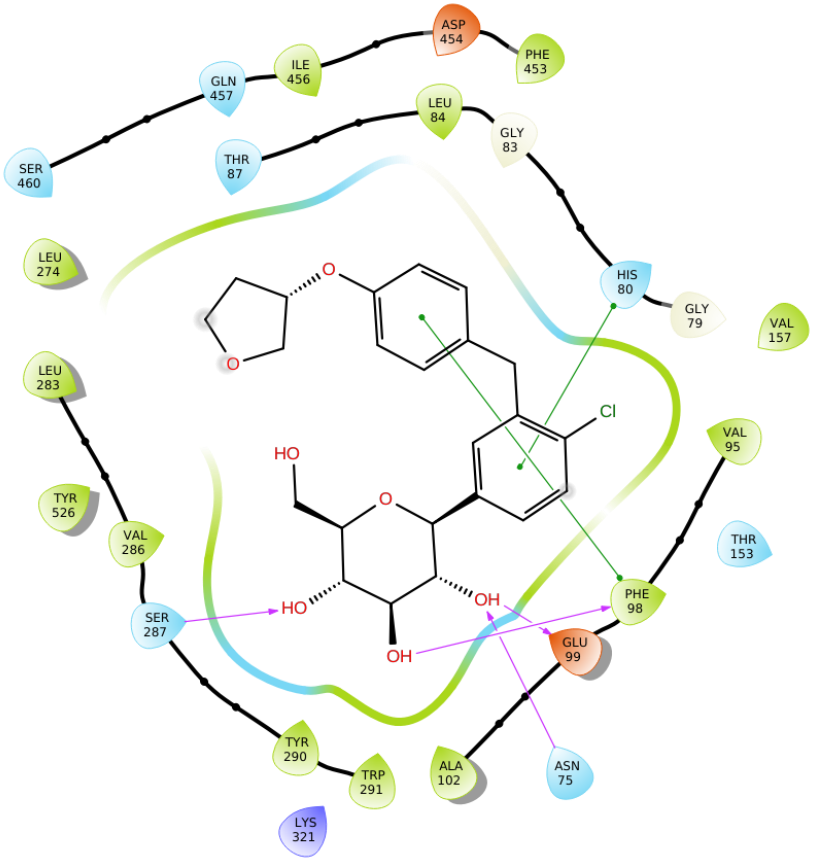


Figure 14: Interactions of empagliflozin with SGLT2. For this illustration, the Ligand Interaction Diagram panel in Maestro was used, which shows the amino acids with polar side chains as light blue, the ones with hydrophobic side chains as green, positively charged side chains as red, negatively charged side chains as blue, and special cases like glycine as white. Furthermore, hydrogen bonds are illustrated as violet arrows and pi-pi interactions as green arrows (Schrödinger Release 2021-1: Maestro, 2021).

3.2.2 Redocking SGLT2

In order to evaluate the ability of the various docking algorithms to dock ligands in a correct and accurate pose into the cryo-EM resolved receptor, the co-crystallized

empagliflozin was redocked into the SGLT2 binding site (Mateev et al., 2022; Cole et al., 2005). The ability to recreate the co-crystallized pose was judged by visual inspection as well as RMSD calculation and inspection of the formed interactions (Cole et al., 2005).

For the analysis of the docking accuracy of each docking algorithm, the best ranked pose was chosen, using the emodel score to rank the poses, as recommended by Schrödinger (Schrödinger Release 2021-1: Glide, 2021).

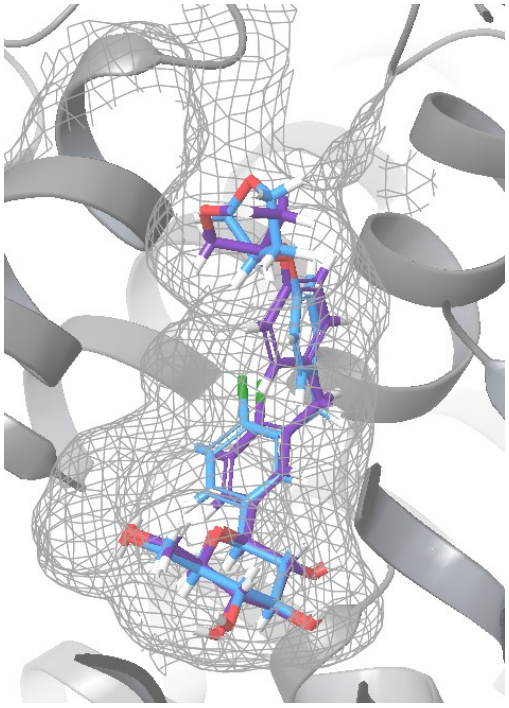


Figure 15: The co-crystallized pose of empagliflozin at the SGLT2 binding site is shown (blue) together with the top ranked pose of the redocked empagliflozin (violet). The surface of the binding site is shown as a grey mesh, while the protein backbone is illustrated as grey ribbons.

This was performed with all four available docking algorithms, where all four of them reached an RMSD that was either in the "excellent" (below 1 Å) or "good" category (1-2 Å) as defined by Mateev et al. (2022)

The results of the redocking studies are summarized in Table1. The Extra Precision mode was able to perform best regarding the RMSD category and the best ranking pose from this mode reached a RMSD of 0.919 Å.

The properties of the docking results will be discussed here by using the results of the Extra Precision mode, which is a semi-rigid docking algorithm, where the ligand is considered flexible, but the protein is not (Schrödinger Release 2021-1: Glide, 2021). Even though the pose calculated by the XP mode for its top ranking entry shows the lowest RMSD compared to the other docking modes, there are still a number of obvious differences regarding the orientation of its tail.

As can be seen in Figure 15, the orientation of the distal benzene ring is slightly twisted in comparison to the co-crystallized pose. This difference can be explained by an evaluation of the differences of the simulated interactions of the docking pro-

cess and the interactions formed by empagliflozin with SGLT2 in their co-crystallized state. Figure 16 shows the interactions formed by the docking process: Unlike the co-crystallized pose, the distal benzene ring does not show any interactions here, which allows the ring to rotate. In contrast, the co-crystallized pose shows an interaction with phenylalanine 98 that can be described as pi-pi-stacking.

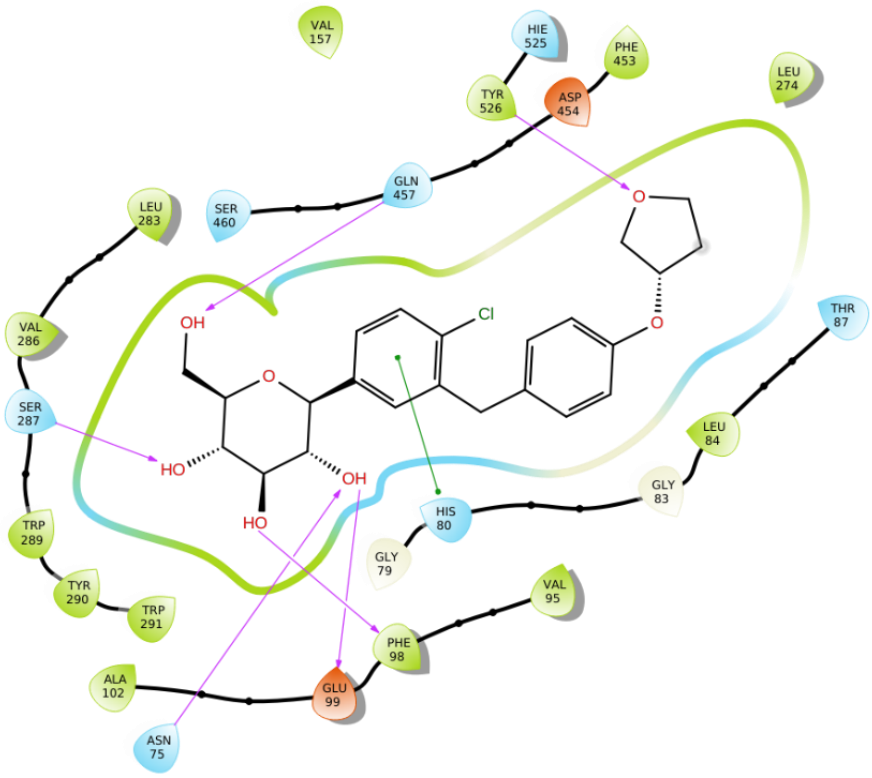


Figure 16: The interactions of the redocked empagliflozin as computed by the Extra Precision docking mode in Glide. The distal benzene ring does not form a pi-pi stacking, which is the reason for its deviation from the co-crystallized pose. Furthermore, the tetrahydrofuran shows an interaction which is not formed in the co-crystallized structure and leads to another difference in the ligand's orientation.

Additionally, the tetrahydrofuran located at the distal end of empagliflozin shows a slight deviation from the co-crystallized pose. This is explained by its interaction with tyrosine 526, which is an interaction computed by the docking algorithm, but is not apparent in the co-crystallized pose.

Similarly, the Standard Precision, the IFD-Standard, and the IFD-Extended Sampling modes formed the same interactions at the tetrahydrofuran ring of empagliflozin as formed by the Extra Precision mode and showed a deviation from the co-crystallized ligand at this position.

In contrast, the divergence of the ligand at the distal ring that was shown by the Extra Precision docking was not shared by the other docking modes, which were able to simulate the pi-pi-stacking of the distal ring with phenylalanine 98. In conclusion, all of the docking protocols for the cryo-EM resolved SGLT2 structure met the predefined criteria (Mateev et al., 2022) for being considered appropriate for further analyses of their ability to dock and score SGLT2 inhibitors. The best performance regarding the achievement of a low RMSD was shown by the Extra Precision mode, and the methods considering the protein to be rigid performed better in general. However, this has to be seen in the context of analysing the properties of the formed interactions, where the Extra Precision mode showed a deviation from the co-crystallized ligand, which was not shown by the other docking algorithms. Of the Induced Fit Docking algorithms, the Extended Sampling method performed worse than the Standard Method, which may be explained by the use of the Standard Precision algorithm during the Extended Sampling protocol of the Induced Fit Docking (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021).

Docking Algorithm	RMSD (Å)
Standard Precision	1.135
Extra Precision	0.919
IFD – Standard	1.2098
IFD – Extended	1.2328

Table 1: The redocking results in the cryo-EM structure of SGLT2. All algorithms performed well and within the predefined bounds, with the Extra Precision mode achieving the lowest RMSD for its top ranked pose.

3.2.3 Redocking AlphaFold

Because of the clashes that the AlphaFold structure displayed with the co-crystallized empagliflozin during the superpositioning tasks, which were explained in chapter 3.1, the AlphaFold structure was deemed inappropriate for docking pro-

cesses where the protein is considered to be rigid. However, the possibility of docking ligands into proteins, for which the protein flexibility is taken into account, made it possible to analyse the ability of the AlphaFold structure to serve as a target for studies using the Induced Fit Docking protocol (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021).

However, the results were not encouraging, and the recreated poses did not resemble the pose that was shown by the co-crystallized empagliflozin. For the Standard mode of the Induced Fit Docking protocol, the top ranked pose according to the emodel score possessed a RMSD of 9.142 Å to the co-crystallized pose.

These results from the Standard mode were reflected by the Extended Sampling mode, where the recreated pose was also not placed into an accurate orientation and conformation, and the RMSD achieved by this protocol was 7.803 Å (figure 17).

For these reasons, it was decided not to use the AlphaFold structure of SGLT2 for any of the following tasks and to regard the cryo-EM structure as more appropriate for the purposes of this thesis.

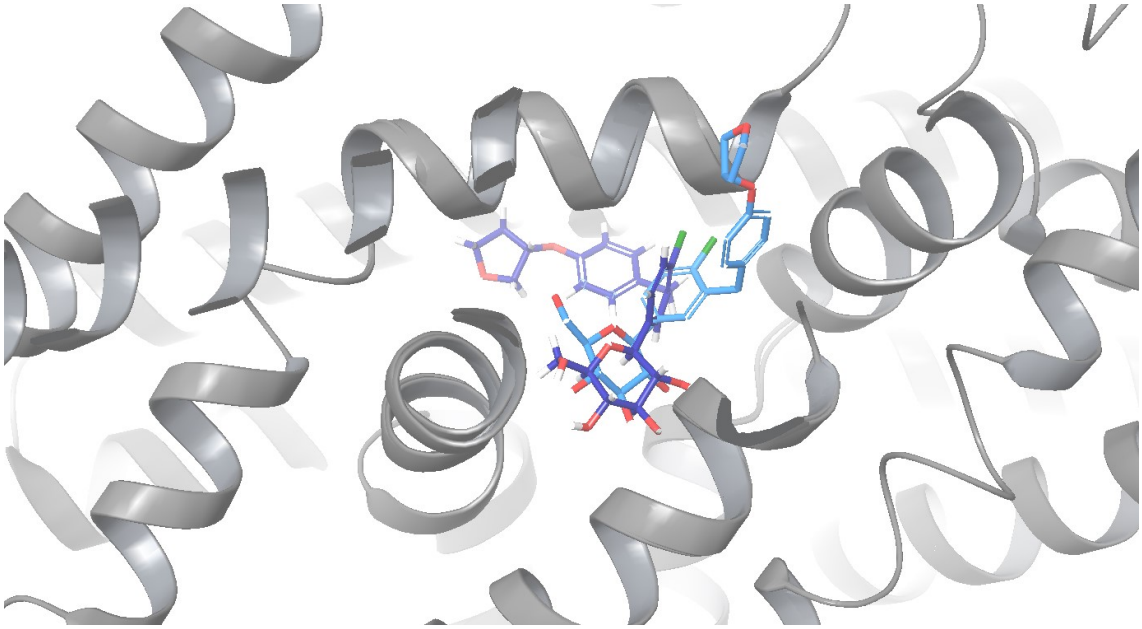


Figure 17: Top ranked pose from the redocking study using the AlphaFold structure and the Extended Sampling IFD protocol. The original pose from the cryo-EM structure is shown in light blue, while the redocked pose is shown in violet.

3.3 Virtual screening

As mentioned in chapter 3.1 and 3.2, the AlphaFold structure proved to be unsuitable for the purposes of this thesis because of the clashes it showed with the co-crystallized ligand when superimposed with the cryo-EM structure, and the inability of the program to redock the ligand into an accurate pose.

For this reason, the virtual screening was performed by using the cryo-EM structure and the Standard Precision mode, as recommended (Friesner et al., 2004).

3.3.1 Ligands

The number of the ligands that were retrieved from the ChEMBL data base was 1229 and the activity they showed ranged from a pChEMBL of 4 to slightly below 10.

The ChEMBL data base provides a standardised value to convey the potency of the tested compounds, which is called pChEMBL and is calculated as the negative log 10 molar of the IC50, XC50, EC50, AC50, Ki, Kd, or potency (Allaway et al., 2018; Gaulton et al., 2016).

The activities and the number of ligands that belonged to certain pChEMBL ranges are summarized in Figure 18 in the form of a histogram plot. For the calculation of this plot, the ligands that had been tested multiple times, across multiple assays, were excluded as their number proved to be insignificant.

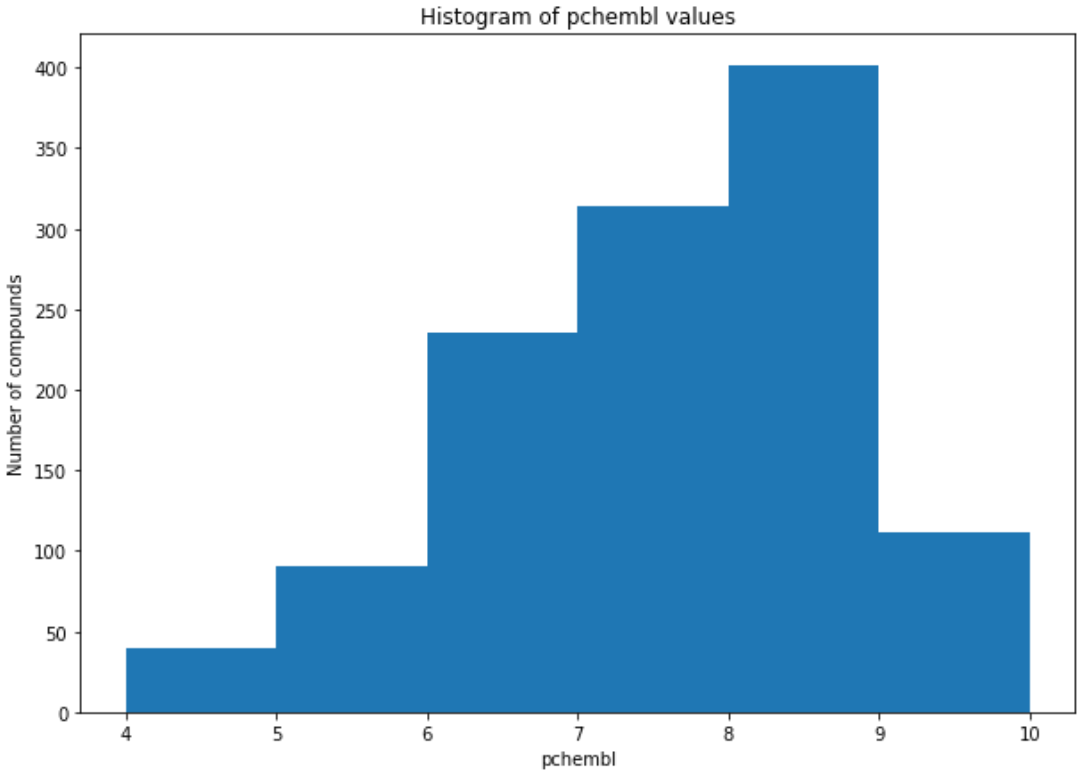


Figure 18: A histogram plot of the retrieved ligands for docking purposes and their activities. The x-axis shows the pChEMBL ranges and the y-axis depicts the number of compounds belonging to each of the pChEMBL ranges. The plot was created by using Matplotlib, which is a visualization tool for the programming language Python (Hunter, 2007).

3.3.2 Enrichment calculations

The choice of a threshold for the classification of ligands with known activity values into actives and inactives is arbitrary, but the threshold above which compounds are considered to be actives is usually set in the micromolar or nanomolar range (Gimeno et al., 2019). As the pChEMBL value is defined as the negative log 10 molar of the different activity values, a micromolar value of 1 (= 0.000001) would correspond to a pChEMBL value of 6 and a nanomolar value of 1 (= 0.000000001) would correspond to a pChEMBL value of 9.

An analysis of the SGLT2 inhibitors currently approved for therapeutical use reveals that the majority of compounds have an IC50 value ranging from 1 to 10 nM, with empagliflozin having an IC50 of 3.1 nM (Grempler et al., 2011), dapagliflozin one of 1.0 nM (Braem et al., 2014), canagliflozin one of 2.2 nM (Nomura et al., 2010), and sotagliflozin one of 1.8 nM (Lapuerta et al., 2015). These activities correspond to

pChEMBL values which are located between 8 and 9, which can be used for the choice of an appropriate activity threshold for the discovery of new potential inhibitors and as a basis for the calculation of enrichment metrics.

In addition to the general agreement of using a micromolar to nanomolar range for the activity threshold (Gimeno et al., 2019), other factors influence the choice of an appropriate threshold. For the assessment of virtual screening protocols, it is important to have a reasonable quantitative relationship of the actives and inactives. Because of the property of virtual screening to be a method that aims to retrieve a significant larger fraction of true positives from a database than a random compound selection, it is necessary to use a database with a large number of inactives (Kirchmair et al., 2008). Therefore, it came in handy to select a pChEMBL threshold that is restrictive enough to classify the actives and inactives into an appropriate quantitative relationship and is located in the range of the activity values of established inhibitors (Gimeno et al., 2019; Kirchmair et al., 2008).

For above-mentioned reasons, the calculation of the enrichment was conducted twice, with the threshold set to 8 and 9 for each calculation, respectively. These thresholds are in the micromolar or nanomolar range and were restrictive enough to ensure an appropriate balance between the inactives and actives.

The results of both thresholds were comparable to each other and are presented in table 2, with the early enrichment of a threshold of pChEMBL 8, which is reflected by the Enrichment Factor, and the Receiver Operator Characteristics (ROC) area under the curve being slightly better. The 10% Enrichment Factor for a threshold of pChEMBL 8 was 1.5, which means that the fraction of retrieved actives in the first ten percent of the virtual screening results was 1.5 times higher than the fraction of actives in the whole database (Schrödinger Release 2021-1: Maestro, 2021). The result of the ROC area under the curve was 0.72 for a threshold of 8 and 0.68 for a threshold of 9. The ROC curves and the % Screen Plots of the pChEMBL threshold of 8 are presented in the figures 19 and 20.

The sum of the ligands that were included in the output of the Virtual Screening Workflow numbered 1147. Out of the 1229 ligands retrieved by the KNIME workflow from the ChEMBL data base, the remaining 82 were rejected during the docking process and therefore not included in the output of the Virtual Screening Workflow.

Threshold (pChEMBL)	8	9
Enrichment Factor (10%)	1.5	1.3
Enrichment Factor (20%)	1.5	1.6
ROC area under the curve	0.72	0.68

Table 2: Results of the enrichment calculation.

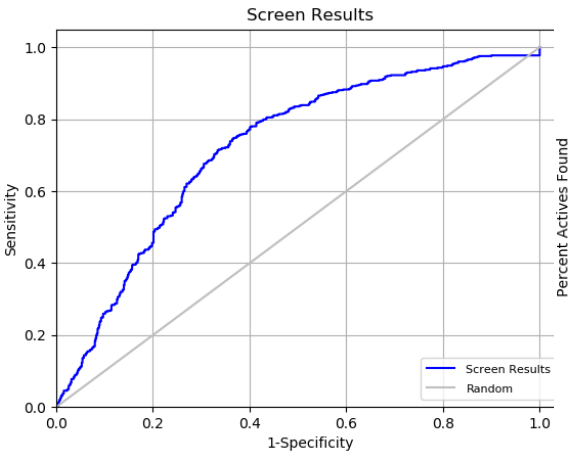


Figure 19: ROC curve of the virtual screening results, using a threshold of pChEMBL 8 for the classification into inactives and actives. The ROC curve shows the true positive rate (y-axis; Sensitivity) plotted against the false positive rate (x-axis; 1-Specificity) for different Docking Score thresholds (Schrödinger Release 2021-1: Maestro, 2021; Fawcett, 2006).

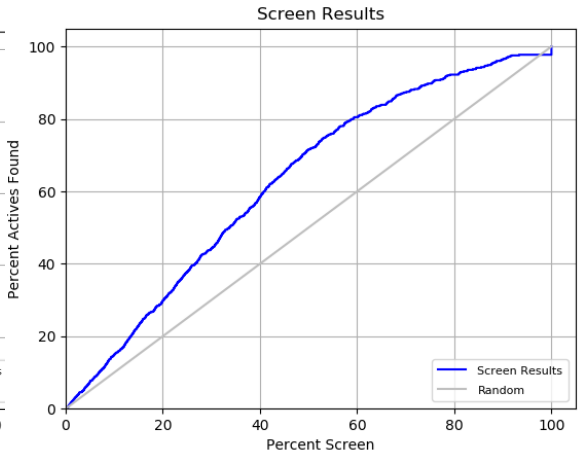


Figure 20: % Screen Plot, which plots the percentage of structures screened (x-axis; Percent Screen) against the percentage of actives recovered up to this point (y-axis; Percent Actives Found) (Schrödinger Release 2021-1: Maestro, 2021)

3.3.2 Docking Score based classification

In addition to the already discussed enrichment calculations of the virtual screening result provided by the Enrichment Calculator of the Schrödinger software, it was decided to create a classification model based on the Docking Score provided by the virtual screening output for a better visualization of the results and a basis for later appliance of the virtual screening workflow to untested data.

For this purpose, a kernel density estimation (KDE) of the theoretical distributions of the actives and the inactives was calculated in order to visualize the differences

between the scoring of the active molecules and the inactive ones, which were classified by using the slightly better performing threshold of pChEMBL 8. A KDE plot is a method for the visualization of distributions and is analogous to a histogram ("seaborn.kdeplot — seaborn 0.11.2 documentation", n.d.). Figure 21 shows the theoretical distribution of the molecules.

The choice of a Docking Score threshold for a predicted classification of ligands into actives and inactives is a context-depending process and has to take into account practical considerations of the investigator (Hubbard & Bayarri, 2003; Triballeau et al., 2005). Nevertheless, using the kernel density estimation allows the calculation of a reasonable threshold by finding the intersection point of the two distribution functions. This results in the choice of a classification threshold that is the equivalent of a point on the upper left part of a ROC curve for various thresholds as shown in figure 19 (Kirchmair et al., 2008; Triballeau et al., 2005).

The choice of a point on the upper left part of the ROC curve is a liberal strategy that prefers sensitivity over specificity and possesses a number of advantages over choosing a point on the lower left corner of the ROC curve. Choosing a point on the lower left corner would result in a more conservative approach that would allow to push the majority of inactives aside. In contrast, choosing a point on the upper left corner allows to take into account the uncertainty of the model while fewer actives would be lost (Triballeau et al., 2005).

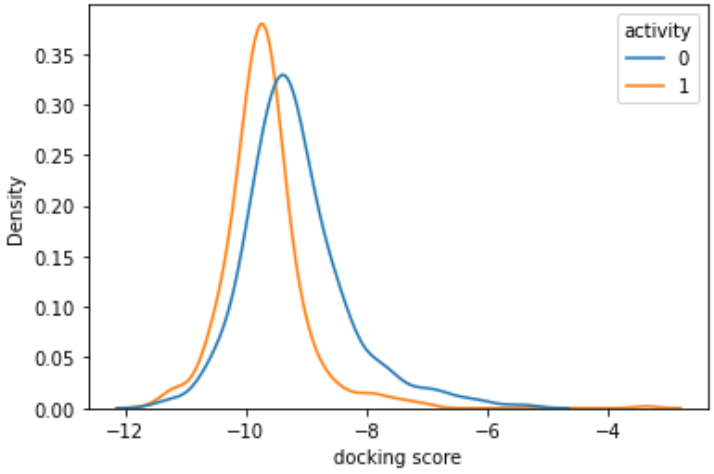


Figure 21: Kernel density estimation used for the visualization of the theoretical distribution of active and inactive ligands based on their Docking Score. Inactive ligands (0) are shown in blue, while active ligands (1) are shown in orange. This figure was plotted by using Seaborn, which is a Python data visualization library based on Matplotlib (Waskom, 2021; Hunter, 2007).

The intersection of the two KDE plots happens to be at a Docking Score of -9.4042, which was used to perform a classification of the obtained data. Every ligand that obtained a lower Docking Score during the virtual screening was therefore classified as an active, and every ligand that obtained a higher Docking Score was classified as inactive, which resulted in 482 ligands that were predicted to be active and 665 ligands that were predicted to be inactive.

The classification based on the Docking Score resulted in a True Positive Rate of 0.773 (Sensitivity) and a True Negative Rate (Specificity) of 0.578, and the calculation of the ROC curve, based on the Docking Score, led to an area under the curve (AUC) of 0.68.

The obtained confusion matrix and the corresponding ROC curve for this classification are displayed in figures 21 and 22. Furthermore, a number of metrics for the evaluation of classification based models were calculated (table 3) in order to allow comparability to the machine learning based classification models, which will be presented in chapter 3.5.

Metrics	Results
Accuracy	0.6661
Sensitivity	0.7733
Specificity	0.5785
Balanced Accuracy	0.6759
F1 Score	0.6757
AUC	0.6759
Precision	0.6000
Matthews correlation coefficient (MCC)	0.3545
Recall	0.7733

Table 3: Relevant metrics for the evaluation of classification based models. The majority of these metrics were calculated by using scikit-learn, which is a Python module integrating machine learning algorithms and providing model evaluation tools, which were used for this thesis (Pedregosa et al., 2011),while the sensitivity and specificity were calculated manually.

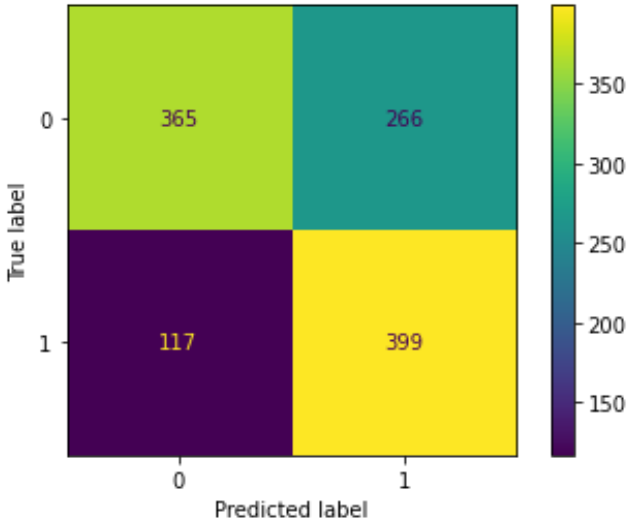


Figure 22: Confusion matrix for a Docking Score based classification model of SGLT2 inhibitors. In this figure, the active class corresponds to the label "1" and the inactive class corresponds to the label "0". The classification resulted in 399 ligands to be correctly classified as active and 266 ligands to be incorrectly classified as active. At the same time 365 ligands were correctly classified as inactive while 117 ligands were incorrectly classified as inactive. This confusion matrix was computed by using scikit-learn (Pedregosa et al., 2011).

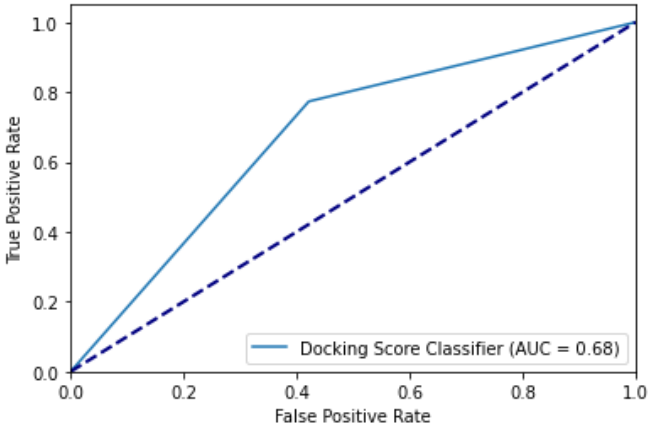


Figure 23: ROC curve of the Docking Score based classification for a Docking Score threshold of -9.402. The False Positive Rate of the classification (0.422) was plotted against the True Positive Rate (0.773), and the curve results in an area under the curve of 0.68. The dotted line represents the performance of a set of actives and inactives with randomly distributed scores (Triballeau et al., 2005; Kirchmair et al., 2008). The ROC curve was computed by using scikit-learn (Pedregosa et al., 2011)

3.3.4 Applying the Virtual Screening Workflow

The results of the evaluation of the Virtual Screening Workflow led to the conclusion that the workflow provides a sufficient differentiation between actives and inactives to apply it to new, untested ligands in order to provide a basis for further *in vitro*

testing. For this reason, the virtual screening workflow was applied to a series of potential inhibitors of SGLT2 with unknown activity that were synthesized by the pharmaceutical chemistry subdivision of the University of Vienna (Kirchweger, Rollinger & Kowalska, 2022).

Because of a significant portion of the ligands possessing a reversed configuration at the C1 position of their sugar moiety that is known to cause the diminishing of activity (Bhattacharya et al., 2020), only ligands with an appropriate configuration were considered for the docking study in order to minimize the risk of incorrectly ranked ligands and to reduce costs for further *in vitro* tests.

After this filtering step, only 42 of the 89 ligands were left for an application of the Virtual Screening Workflow and the described docking based classification. As the virtual screening is usually applied for finding a few promising compounds as a fraction from a bigger database (Gimeno et al., 2019), the ten best performing compounds and their docking scores are presented in table 4.

However, the results of the virtual screening are not encouraging. None of the compounds exceed the set threshold of a Docking Score of -9.4042 from the classification model of the previous chapter (3.3.2), and when the model is applied to the compounds and their accompanying Docking Scores, all compounds are predicted to be inactive.

Compound	Docking Score
GJB1224	-9.338
GJB1244	-8.209
GJB1182	-8.106
GJB1141	-8.043
GJB1096	-8.029
GJB407	-7.811
GJB1126	-7.697
GJB392	-7.606
GJB916	-7.583
GJB539	-7.530

Table 4: Results of the Virtual Screening Workflow for the compounds with unknown activity (Kirchweger, Rollinger & Kowalska, 2022). The left column shows the compound codes to make a comparison with the results from individual docking and the machine learning tasks possible and the right column presents the respective Docking Scores. None of the synthesized compounds exceeds the set threshold (-9.4042) for the prediction as an active.

3.4 Individual docking

As mentioned in chapter 2.3.6, the Extra Precision docking mode and the two Induced Fit Docking protocols provided by the Schrödinger Software Suite are not recommended for the purpose of virtual screening, because of their high computational demand (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021; Friesner et al., 2004; Friesner et al., 2006; Du et al., 2016).

3.4.1 Ligands

For this reason, a series of congeneric SGLT2 inhibitors with significant differences in their activity were docked by using the three above-mentioned docking methods, and the ability of the programs to rank them correctly was assessed.

Even though, the ligands were obtained from three different publications, only activity data from comparable assay methodologies was used (Braem et al., 2014; Lee et al., 2010; Ng et al., 2017).

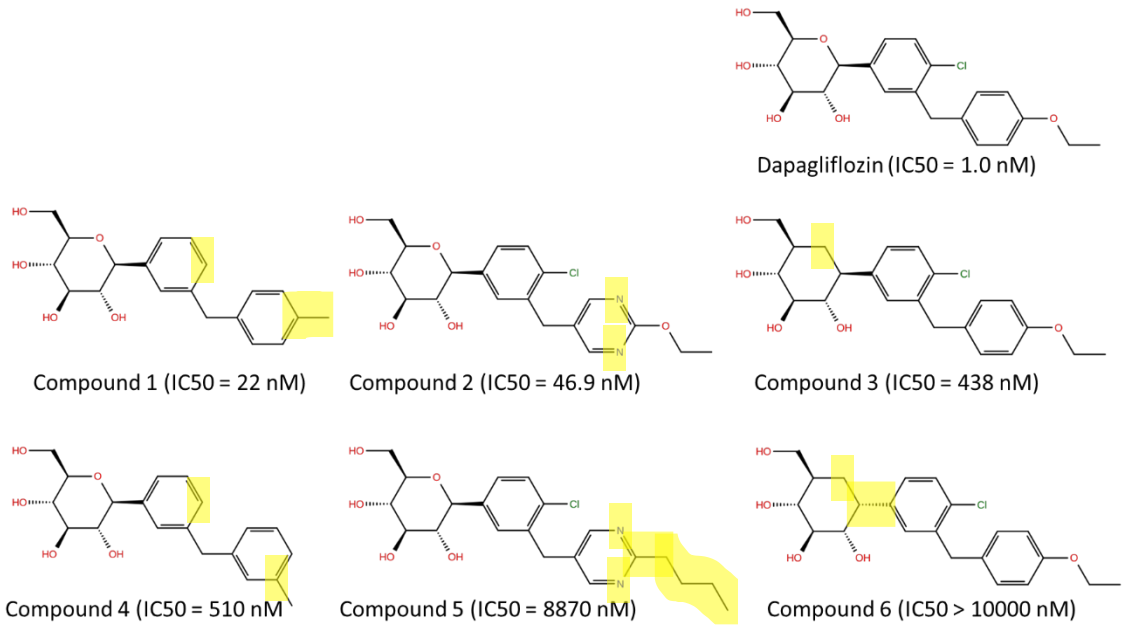


Figure 24: The SGLT2 inhibitor dapagliflozin and 6 of its analogues, which possess small differences in their structure, but significant differences regarding their activity. Their differences are highlighted in yellow. The activity data is taken from three separate sources, with dapagliflozin and compounds 1 and 4 taken from Braem et al. (2014), compounds 2 and 5 taken from Lee et al. (2010) and compounds 3 and 6 from Ng et al. (2017).

Compound 1 differs from dapagliflozin regarding its para position of the proximal benzene, where no chloride is present, and its para position of distal benzene, where a methyl group replaces the ethoxy group of dapagliflozin. Compounds 1 and 4 differ only regarding the position of the methyl substitute, which is placed at the ortho position in compound 4 and leads to a significant activity drop (Braem et al., 2014).

Compound 2 possesses a pyrimidine ring instead of the distal benzene of dapagliflozin, and its difference to compound 5 is characterized by the para position of the pyrimidine ring, which substitutes the ethoxy group with a hydrophobic alkyl chain. The added butyl chain leads to an almost two hundred fold difference in activity (Lee et al. 2010).

Compound 3 shows a four hundred fold activity difference to dapagliflozin, which is caused by the absence of the oxygen at the sugar moiety, making it a cyclohexane. Compound 6 possesses a reversed configuration at position C1 of the cyclohexane, rendering the compound essentially inactive with an IC50 of more than 10000 (Ng et al., 2017).

The presented differences in activity caused by relatively subtle differences in the molecular structure of the ligands make them an appropriate tool for the assessment of the docking protocols. For this reason, the performances of the above-mentioned docking algorithms were judged by ranking the ligands according to their obtained scores and calculating the Spearman rank correlation of the predicted ranks with the ranking created by the *in vitro* testing.

3.4.2 Docking results

In addition to the scores provided by the output of the docking runs, the resulting best ranked poses according to the emodel score of the docked ligands were assessed by using the MMGBSA dG Bind, which provides an estimation of the binding

energies for the results of the MM-GBSA based refinement (Schrödinger Release 2021-1: Prime, 2021). The results of the docking runs are presented in table 5.

Scores	Dapagliflozin	Compound 1	Compound 2	Compound 3	Compound 4	Compound 5	Compound 6	Correlation
Activity IC50 (nM)	1.0	22	46.9	438	510	8870	> 10000	/
IFD Standard Docking Score	-14.68	-12.65	-12.89	-11.46	-12.62	-11.73	-12.25	0.7143
IFD Extended Docking Score	-10.93	-11.455	-11.414	-11.143	-10.742	-10.598	-10.917	0.6786
IFDscore Standard	-1194.73	-1191.59	-1195.69	-1190.35	-1190.08	-1193.61	-1189.19	0.6429
XP Docking Score	-12.462	-9.902	-9.798	-11.758	-11.164	-9.848	-10.409	0.25
MMGBSA dG bind	-99.18	-85.62	-94.11	-98.25	-98.05	-109.8	-69.32	0.1429
IFDscore Extended	-23788.84	-23699.57	-23879.87	-23774.51	-23736.62	-23865.92	-23749.48	0.0357

Table 5: Results of the individual docking studies. The ligand poses of each ligand from the output of the docking protocols were ranked by their emodel scores and the best scoring poses were then ranked by the displayed scoring functions. As seen in the furthest right column (highlighted in green), the highest Spearman rank correlation coefficient was obtained by the Docking Score of the IFD Standard Sampling protocol with a coefficient of 0.7143.

As seen in table 5, the best performing scoring function was the Docking Score provided by the Standard Sampling of the Induced Fit Docking protocol, which was therefore used for further docking studies of the novel synthesized ligands (Kirchweger, Rollinger & Kowalska, 2022).

3.4.3 Applying the IFD – Standard Sampling

In parallel to the application of the Virtual Screening Workflow as presented in chapter 3.3.4, 42 of the 89 synthesized compounds (Kirchweger, Rollinger & Kowalska,

2022) were tested by using the Induced Fit Docking – Standard Sampling protocol, which was the best performing docking approach from the individual docking step.

The ten best performing ligands can be seen in table 6, which is shown below.

Compound	Docking Score
Dapagliflozin (comparison)	-14.684
GJB1244	-12.850
GJB1141	-12.816
GJB1230 (hydrolysed)	-12.230
GJB1093	-11.999
GJB407	-11.866
GJB1098	-11.763
GJB1224	-11.710
GJB1097	-11.624
GJB268	-11.274
GJB1182	-11.241

Table 6: Results of the Virtual Screening Workflow for the newly synthesized compounds (Kirchweger, Rollinger & Kowalska, 2022). The left column shows the compound codes to make a comparison with the results from individual docking and the machine learning tasks possible and the right column presents the respective Docking Scores. The compounds highlighted in green are also present in the ten best performing compounds of the Virtual Screening Workflow output.

As the Docking Scores from the output of the IFD – Standard Sampling protocol are based on the Extra Precision scoring function, which uses a different algorithm for the scoring of the ligands than the Standard Precision (Friesner et al., 2004), the results can not be used for the Docking Score based classification model presented in chapter 3.3.2. For this reason, dapagliflozin, which is a known active molecule with an activity of IC50 (nM) = 1.0 and was used as one of the ligands for the assessment of the IFD protocol as discussed in chapters 3.4.1 and 3.4.2, is shown here as a tool for the comparison of the Docking Scores achieved by the synthesized compounds (Kirchweger, Rollinger & Kowalska, 2022).

The compounds GJB1244, GJB1141, GJB407, GJB1224, and GJB1182 also appear in the top ranks of the Virtual Screening Output and are therefore the most promising candidates. However, it has to be considered that none of them were predicted to be active when applying the classification model presented in chapter 3.3.2.

3.5 Machine learning

In addition to the described structure-based methods, a ligand based analysis of the SGLT2 inhibitors was conducted. The molecules retrieved and standardised by the KNIME workflow, which is part of the PharminfoVienna Sandbox, were used for the creation of various QSAR based machine learning models for the classification of molecules. For this purpose, the Sandbox and Retraining notebooks, which were described in chapter 2.4, were utilized (Pharmacoinformatics Research Group, 2021).

For the classification of the activity of the ligands from ChEMBL, the thresholds of 7 and 8 were used and the performances of the thresholds were evaluated by using the balanced accuracy metric for each classifier of each notebook. This was done because the balanced accuracy is sensitive to class imbalances and can be used to deal with imbalanced datasets (Bender et al., 2022; Brodersen et al., 2010).

For perfectly balanced datasets, the differences between the accuracy and the balanced accuracy should not be substantial, while they could differ for imbalanced datasets with balanced accuracy being lower in the case of applying a biased classifier to an imbalanced data set (Brodersen et al., 2010). A threshold of pChEMBL 7 resulted in balanced accuracies that were generally lower than the accuracies, while a pChEMBL resulted in balanced accuracies that were virtually equal to the accuracies. This was true for all of the models across both notebooks. This assessment of the balanced accuracy metric led to the conclusion that a threshold of 8 is more appropriate for this task, which is also the threshold that was chosen for the docking score based classification as presented in chapter 3.3.2, allowing a better comparability

Both notebooks provide a multitude of metrics for the evaluation of the performances of the models and, as it is recommended to use an array of metrics for validation purposes, table 7, which is presented below, includes all of the metrics calculated for the models. As can be seen in the table, the Random Forest classifier performed the best for both notebooks, with the Sandbox notebook achieving better results than the Retraining notebook.

	LR(1)	SVM(1)	RF(1)	KNN(1)	SVM(2)	RF(2)	KNN(2)
Accuracy	0.722	0.743	0.840	0.772	0.85	0.86	0.82
Sensitivity	0.692	0.760	0.817	0.798	0.87	0.87	0.88
Specificity	0.744	0.729	0.857	0.752	0.83	0.86	0.78
Balanced accuracy	0.718	0.744	0.837	0.775	0.85	0.86	0.83
F1 score	0.686	0.721	0.817	0.755	0.84	0.86	0.82
ROC AUC	0.718	0.744	0.837	0.775	0.85	0.86	0.83
Precision score	0.679	0.687	0.817	0.716	0.82	0.84	0.78
MCC	0.436	0.485	0.674	0.546	0.7	0.72	0.65
Recall	0.692	0.760	0.817	0.798	0.87	0.87	0.88

Table 7: Results of the machine learning models when applied to the test sets, shown as the metrics that were achieved. The best performing models are highlighted in green, which is the Random Forest classifier for both notebooks. Classifiers annotated with (1) are part of the Retraining notebook and classifiers annotated with (2) belong to the Sandbox notebook. Abbreviations: LR = Logistic Regression, SVM = Support Vector Machine, RF = Random Forest, KNN = k-nearest Neighbors, ROC AUC = Receiver Operator Characteristics area under the curve, MCC = Matthews Correlation Coefficient.

In addition to the presented metrics in table 7, the performances of the best performing models are visualized in figure 25 where the ROC-curves and the confusion matrices are displayed.

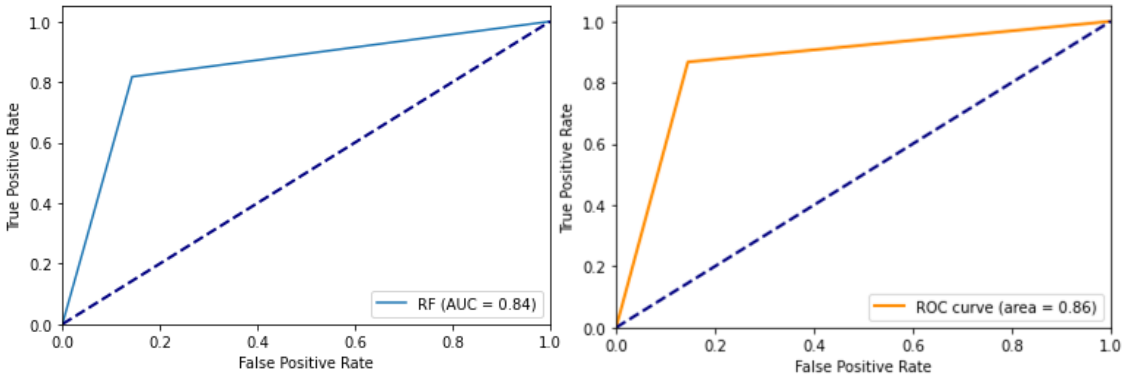


Figure 25: ROC curves and the corresponding AUC of the best performing models from the Retraining and the Sandbox notebook, respectively. The ROC curve of the Retraining notebook is displayed in blue on the left side, while the ROC curve of the Sandbox is shown in orange on the right side. The ROC curves were computed using scikit-learn (Pedregosa et al., 2011)

As can be seen in figure 26, the Random Forest model from the Retraining notebook classified the test set of the retrieved SGLT2 data the following way: 85 ligands were correctly predicted to be active (true positives), 19 ligands were incorrectly predicted to be active (false positives), 114 were correctly predicted to be inactive (true

negatives), and 19 were incorrectly predicted to be inactive (false negatives). Similarly, the Sandbox notebook resulted in 98 true positives, 18 false positives, 106 true negatives, and 15 false negatives.

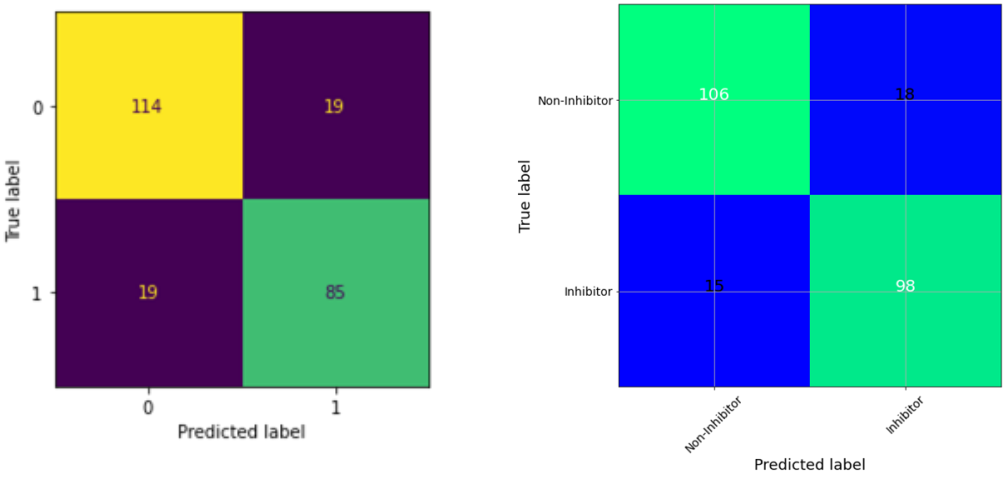


Figure 26: Confusion matrices of the best performing models from the Retraining and the Sandbox notebook, respectively. The confusion matrix of the Retraining notebook is displayed on the left side, while the confusion matrix of the Sandbox notebook is shown on the right side.

3.5.1 Applying the machine learning models

As the Random Forest models of the two notebooks showed the best performance during their application to the test sets, it was decided to apply them to the 42 previously described molecules with unknown activity (Kirchweger, Rollinger & Kowalska, 2022).

For the preparation of the ligands, the KNIME workflow was utilized, which was also used for the standardization of the molecules during the model generation step. The implemented RDKit Standardiser, which removes the stereochemistry and filters out molecules with nonorganic atoms, was used for the standardization of the ligands. As some of the molecules showed no differences except for their stereochemistry, redundant structures were filtered out using the GroupBy node (Pharmacoinformatics Research Group, 2021).

For the application of the Random Forest model created by the Sandbox notebook, the preselected descriptors offered by KNIME workflow were chosen and calculated using the RDKit Descriptor Calculation node. In contrast, for the application of the model created by the Retraining notebook, the descriptor calculation was conducted

inside the notebook (Pharmacoinformatics Research Group, 2021; Pharmacoinformatics Research Group 2022).

For both models, all of the synthesized molecules were predicted to be inactive, which supports the results found by the docking based classification presented in chapter 3.3.2, where none of the ligands achieved a Docking Score high enough to be considered active.

4 Conclusion

The aim of this master's thesis was to provide an *in-silico*-based method, mainly driven by the use of molecular docking, and QSAR modelling as an additional tool, for the analysis of the sodium/glucose co-transporter 2 (SGLT2) and the prediction of the activity of its inhibitors.

In addition to the aforementioned aims of this thesis, one of the objectives was to investigate SGLT2 structures adequate for docking based approaches and the assessment of the utility of the structure computed by AlphaFold. It was found that, while the SGLT2 structure elucidated through cryo-EM has the potential to serve as a powerful tool for drug discovery purposes, the structure provided by AlphaFold is not able to recreate the desired poses and is therefore not of utility (Niu et al., 2021; Jumper et al., 2021).

For the assessment of the activity of potential SGLT2 inhibitors that have not yet been tested *in vitro*, multiple classification models were created and the ability of various docking algorithms and scoring functions to correctly rank molecules according to their activity was analysed.

Over the course of this study, a docking based classification model and a number of QSAR machine learning models were computed. For the docking based classification model the Virtual Screening Workflow (Schrödinger Release 2021-1: Glide, 2021) was utilized, and a Docking Score threshold of -9.4042 revealed to be the most appropriate for a predicted classification into actives and inactives. The basis for this classification was a pChEMBL threshold of 8.

This approach displayed a number of limitations, one of which was the high number of false positives obtained from the classification. This could be explained by the nature of the Standard Precision mode, which was used for this classification model and was designed to minimize the number of false negatives (Friesner et al., 2004). This necessitates to further validate and inspect any positive result stemming from this approach.

From the machine learning models that were analysed, it was concluded that the models utilizing the Random Forest classifier performed the best.

The docking algorithms that were assessed were the IFD – Standard Sampling, IFD – Extended Sampling, Extra Precision, and the Standard Precision modes, which are part of the Schrödinger Software Suite (Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, 2021). Their performance was judged by the Spearman rank correlation coefficient that they were able to achieve, and which is used for correlating two rankings (Myers & Well, 2002). In this case, the rankings of scores computed by the docking algorithms were correlated with the ranking of compounds by their activity. It was found that the IFD – Standard Sampling using the Docking Score for the ranking was the best performing method.

This approach was limited by the small number of ligands (7), which was used for the docking and the subsequent correlation calculations. The small number of ligands used is caused by the nature of the docking protocols, which is their high computational expense, and by the time as a limiting factor in the docking processes (Du et al., 2016; Friesner et al., 2006). For further analyses of the capabilities of the IFD – Standard Sampling method, it would be desirable to validate the results through docking a larger number of compounds with known activities towards SGLT2.

Furthermore, the methods that were developed were applied to a series of compounds with unknown activity with the objective to support the development of new SGLT2 inhibitors. Out of a dataset of 89 compounds with unknown activities (Kirchweger, Rollinger & Kowalska, 2021), 42 molecules were determined to be appropriate (Bhattacharya et al., 2020) for the prediction by the created classification models. However, none were predicted to be active.

As the IFD protocol and the Virtual Screening Workflows were found to be appropriate tools for the ranking of molecules according to their activity, the top ranked

results were analysed and the overlapping molecules of the two methods were determined. This suggested that the compounds GJB1244, GJB1141, GJB407, GJB1224, and GJB1182 were the ones with the greatest potential to be active.

To conclude, despite the discussed limitations, the docking based virtual screening and classification, along with the individual docking of a small number of compounds using the induced fit protocols proved to be useful tools for the assessment of the activities of potential SGLT2 inhibitors and may play a role in further *in silico* approaches for drug discovery purposes of SGLT2 inhibitors. Additionally, a combination with ligand based approaches, such as machine learning based QSAR models, may serve as a handy tool to further validate the results of the structure based approaches.

5 References

Allaway, R. J., La Rosa, S., Guinney, J., & Gosline, S. J. C. (2018). Probing the chemical–biological relationship space with the Drug Target Explorer. *Journal of Cheminformatics*, *10*(1). https://doi.org/10.1186/s13321-018-0297-4

American Diabetes Association. (2017). 2. Classification and Diagnosis of Diabetes: *Standards of Medical Care in Diabetes—2018*. *Diabetes Care*, *41*(Supplement_1), S13–S27. https://doi.org/10.2337/dc18-s002

Barros, R. P. C., Sousa, N. F., Scotti, L., & Scotti, M. T. (2020). Use of Machine Learning and Classical QSAR Methods in Computational Ecotoxicology. *Methods in Pharmacology and Toxicology*, 151–175. https://doi.org/10.1007/978-1-0716-0150-1_7

Bell, E. W., & Zhang, Y. (2019). DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. *Journal of Cheminformatics*, *11*(1). https://doi.org/10.1186/s13321-019-0362-7

Bender, A., Schneider, N., Segler, M., Patrick Walters, W., Engkvist, O., & Rodrigues, T. (2022). Evaluation guidelines for machine learning tools in the chemical sciences. *Nature Reviews Chemistry*, *6*(6), 428–442. https://doi.org/10.1038/s41570-022-00391-9

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Bhattacharya, S., Rathore, A., Parwani, D., Mallick, C., Asati, V., Agarwal, S., Rajoriya, V., Das, R., & Kashaw, S. K. (2020). An exhaustive perspective on structural insights of SGLT2 inhibitors: A novel class of antidiabetic agent. *European Journal of Medicinal Chemistry*, *204*, 112523. https://doi.org/10.1016/j.ejmech.2020.112523

Braem, A., Deshpande, P. P., Ellsworth, B. A., & Washburn, W. N. (2014). Discovery and Development of Selective Renal Sodium-Dependent Glucose Cotransporter 2 (SGLT2) Dapagliflozin for the Treatment of Type 2 Diabetes. *Topics in Medicinal Chemistry*, 73–94. https://doi.org/10.1007/7355_2014_41

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. *2010 20th International Conference on Pattern Recognition*. https://doi.org/10.1109/icpr.2010.764

Cai, W., Jiang, L., Xie, Y., Liu, Y., Liu, W., & Zhao, G. (2015). Design of SGLT2 Inhibitors for the Treatment of Type 2 Diabetes: A History Driven by Biology to Chemistry. *Medicinal Chemistry*, *11*(4), 317–328. https://doi.org/10.2174/1573406411666150105105529

Can I relate MM-GBSA energies to binding affinity? | Schrödinger. (2015, November 10). Schrödinger. Retrieved August 10, 2022, from https://www.schrodinger.com/kb/1647?original_search=mmgbsa

Chrysant, S. (2017). Promising cardiovascular and blood pressure effects of the SGLT2 inhibitors: a new class of antidiabetic drugs. *Drugs of Today*, *53*(3), 191. https://doi.org/10.1358/dot.2017.53.3.2555985

Clar, C., Gill, J. A., Court, R., & Waugh, N. (2012). Systematic review of SGLT2 receptor inhibitors in dual or triple therapy in type 2 diabetes. *BMJ Open*, *2*(5), e001007. https://doi.org/10.1136/bmjopen-2012-001007

Coady, M. J., el Tarazi, A., Santer, R., Bissonnette, P., Sasseville, L. J., Calado, J., Lussier, Y., Dumayne, C., Bichet, D. G., & Lapointe, J. Y. (2016). MAP17 Is a Necessary Activator of Renal Na+/Glucose Cotransporter SGLT2. *Journal of the American Society of Nephrology*, *28*(1), 85–93. https://doi.org/10.1681/asn.2015111282

Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D., & Taylor, R. (2005). Comparing protein-ligand docking programs is difficult. *Proteins: Structure, Function, and Bioinformatics*, *60*(3), 325–332. https://doi.org/10.1002/prot.20497

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/a:1022627411411

Dadashpour, S., Tuylu Kucukkilinc, T., Unsal Tan, O., Ozadali, K., Irannejad, H., & Emami, S. (2015). Design, Synthesis and In Vitro Study of 5,6-Diaryl-1,2,4-triazine-3-ylthioacetate Derivatives as COX-2 and β-Amyloid Aggregation Inhibitors. *Archiv Der Pharmazie*, *348*(3), 179–187. https://doi.org/10.1002/ardp.201400400

Danishuddin, & Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Today*, *21*(8), 1291–1302. https://doi.org/10.1016/j.drudis.2016.06.013

Dhanjal, J. K., Kumar, V., Garg, S., Subramani, C., Agarwal, S., Wang, J., Zhang, H., Kaul, A., Kalra, R. S., Kaul, S. C., Vrati, S., Sundar, D., & Wadhwa, R. (2021). Molecular mechanism of anti-SARS-CoV2 activity of Ashwagandha-derived withanolides. *International Journal of Biological Macromolecules*, *184*, 297–312. https://doi.org/10.1016/j.ijbiomac.2021.06.015

Docking and Scoring | Schrödinger. (n.d.). Schrödinger. Retrieved August 10, 2022, from https://www.schrodinger.com/science-articles/docking-and-scoring

Du, X., Li, Y., Xia, Y. L., Ai, S. M., Liang, J., Sang, P., Ji, X. L., & Liu, S. Q. (2016). Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *International Journal of Molecular Sciences*, *17*(2), 144. https://doi.org/10.3390/ijms17020144

Ehrenkranz, J. R. L., Lewis, N. G., Ronald Kahn, C., & Roth, J. (2005). Phlorizin: a review. *Diabetes/Metabolism Research and Reviews*, *21*(1), 31–38. https://doi.org/10.1002/dmrr.532

Fang, Y. F. (2013, May). *HYDROGEN BOND, PI-PI STACKING, AND VAN DER WAALS INTERACTIONS INVESTIGATED WITH DENSITY FUNCTIONAL THEORY* (Thesis). https://digitallibrary.tulane.edu/islandora/object/tulane%3A25609/datastream/PDF/view

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Forrest, L. R., & Rudnick, G. (2009). The Rocking Bundle: A Mechanism for Ion-Coupled Solute Flux by Symmetrical Transporters. *Physiology*, *24*(6), 377–386. https://doi.org/10.1152/physiol.00030.2009

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., & Shenkin, P. S. (2004). Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, *47*(7), 1739–1749. https://doi.org/10.1021/jm0306430

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C., & Mainz, D. T. (2006). Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein–Ligand Complexes. *Journal of Medicinal Chemistry*, *49*(21), 6177–6196. https://doi.org/10.1021/jm051256o

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2016). The ChEMBL database in 2017. *Nucleic Acids Research*, *45*(D1), D945–D954. https://doi.org/10.1093/nar/gkw1074

Gellman, S. H. (1997). Introduction: Molecular Recognition. *Chemical Reviews*, *97*(5), 1231–1232. https://doi.org/10.1021/cr970328j

Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, *10*(5), 449–461. https://doi.org/10.1517/17460441.2015.1032936

Gimeno, A., Ojeda-Montes, M., Tomás-Hernández, S., Cereto-Massagué, A., Beltrán-Debón, R., Mulero, M., Pujadas, G., & Garcia-Vallvé, S. (2019). The Light and Dark Sides of Virtual Screening: What Is There to Know? *International Journal of Molecular Sciences*, *20*(6), 1375. https://doi.org/10.3390/ijms20061375

Grempler, R., Thomas, L., Eckhardt, M., Himmelsbach, F., Sauer, A., Sharp, D. E., Bakker, R. A., Mark, M., Klein, T., & Eickelmann, P. (2011). Empagliflozin, a novel selective sodium glucose cotransporter-2 (SGLT-2) inhibitor: characterisation and comparison with other SGLT-2 inhibitors. *Diabetes, Obesity and Metabolism*, *14*(1), 83–90. https://doi.org/10.1111/j.1463-1326.2011.01517.x

Grillberger, K. (2022). Structural modelling of Molecular Initiating Event interactions: Docking neonicotinoids into human nAChR (Thesis).

Gudivada, V., Irfan, M., Fathi, E., & Rao, D. (2016). Cognitive Analytics. *Handbook of Statistics*, 169–205. https://doi.org/10.1016/bs.host.2016.07.010

How do I align independent chains in two or more protein structures? | Schrödinger. (2016, December 2). Schrödinger. Retrieved August 10, 2022, from

https://www.schrodinger.com/kb/1426?original_search=protein%20structure%20alignment

Huang, S. Y., & Zou, X. (2010). Advances and Challenges in Protein-Ligand Docking. *International Journal of Molecular Sciences*, *11*(8), 3016–3034. https://doi.org/10.3390/ijms11083016

Hubbard, R., & Bayarri, M. J. (2003). Confusion Over Measures of Evidence (*p*'s) Versus Errors (*α*'s) in Classical Statistical Testing. *The American Statistician*, *57*(3), 171–178. https://doi.org/10.1198/0003130031856

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.).

Jain, A. N. (2007). Bias, reporting, and sharing: computational evaluations of docking methods. *Journal of Computer-Aided Molecular Design*, *22*(3–4), 201–212. https://doi.org/10.1007/s10822-007-9151-x

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Keyvanpour, M. R., & Shirzad, M. B. (2021). An Analysis of QSAR Research Based on Machine Learning Concepts. *Current Drug Discovery Technologies*, *18*(1), 17–30. https://doi.org/10.2174/1570163817666200316104404

Kirchmair, J., Markt, P., Distinto, S., Wolber, G., & Langer, T. (2008). Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes? *Journal of Computer-Aided Molecular Design*, *22*(3–4), 213–228. https://doi.org/10.1007/s10822-007-9163-6

Kirchweger, B., Rollinger, J. M., & Kowalska, A. (2022). *Effects of C-glycosyls on Nile red staining of Caenorhabditis elegans* [Unpublished manuscript]. Subdivision of Pharmaceutical Chemistry, Department of Pharmaceutical Sciences, University of Vienna.

Klepsch, F., Chiba, P., & Ecker, G. F. (2011). Exhaustive Sampling of Docking Poses Reveals Binding Hypotheses for Propafenone Type Inhibitors of P-Glycoprotein. *PLoS Computational Biology*, *7*(5), e1002036. https://doi.org/10.1371/journal.pcbi.1002036

Kontoyianni, M. (2017). Docking and Virtual Screening in Drug Discovery. *Methods in Molecular Biology*, 255–266. https://doi.org/10.1007/978-1-4939-7201-2_18

Lapuerta, P., Zambrowicz, B., Strumph, P., & Sands, A. (2015). Development of sotagliflozin, a dual sodium-dependent glucose transporter 1/2 inhibitor. *Diabetes and Vascular Disease Research*, *12*(2), 101–110. https://doi.org/10.1177/1479164114563304

Lee, J., Kim, J. Y., Choi, J., Lee, S. H., Kim, J., & Lee, J. (2010). Pyrimidinylmethylphenyl glucoside as novel C-aryl glucoside SGLT2 inhibitors. *Bioorganic & Medicinal Chemistry Letters*, *20*(23), 7046–7049. https://doi.org/10.1016/j.bmcl.2010.09.103

Li, Y., Pan, D., Liu, J., Kern, P. S., Gerberick, G. F., Hopfinger, A. J., & Tseng, Y. J. (2007). Categorical QSAR Models for Skin Sensitization based upon Local Lymph Node Assay Classification Measures Part 2: 4D-Fingerprint Three-State and Two-2-State Logistic Regression Models. *Toxicological Sciences*, *99*(2), 532–544. https://doi.org/10.1093/toxsci/kfm185

Lu, C., Wu, C., Ghoreishi, D., Chen, W., Wang, L., Damm, W., Ross, G. A., Dahlgren, M. K., Russell, E., von Bargen, C. D., Abel, R., Friesner, R. A., & Harder, E. D. (2021). OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *Journal of Chemical Theory and Computation*, *17*(7), 4291–4300. https://doi.org/10.1021/acs.jctc.1c00302

Macalino, S. J. Y., Gosu, V., Hong, S., & Choi, S. (2015). Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research*, *38*(9), 1686–1701. https://doi.org/10.1007/s12272-015-0640-5

Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, *27*(3), 221–234. https://doi.org/10.1007/s10822-013-9644-8

Mateev, E. M., Valkova, I. V., Angelov, B. A., Georgieva, M. G., & Zlatkov, A. Z. (2022). VALIDATION THROUGH RE-DOCKING, CROSS-DOCKING AND LIGAND ENRICHMENT IN VARIOUS WELL-RESOLUTED MAO-B RECEPTORS. *INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES AND RESEARCH*, *13*(3), 1099–1107. https://doi.org/10.13040/IJPSR.0975-8232.13(3).1099-07

Mechelke, M., & Habeck, M. (2010). Robust probabilistic superposition and comparison of protein structures. *BMC Bioinformatics*, *11*(1). https://doi.org/10.1186/1471-2105-11-363

Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. *WIREs Computational Molecular Science*, *4*(5), 468–481. https://doi.org/10.1002/wcms.1183

Molecular Operating Environment (MOE), 2020.09, Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2020.

Morando, M. A., Saladino, G., D'Amelio, N., Pucheta-Martinez, E., Lovera, S., Lelli, M., López-Méndez, B., Marenchino, M., Campos-Olivas, R., & Gervasio, F. L. (2016). Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the c-Src Kinase. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep24439

Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., Isayev, O., Curtalolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D. A., Agrafiotis, D., Cherkasov, A., & Tropsha, A. (2020). QSAR without borders. *Chemical Society Reviews*, *49*(11), 3525–3564. https://doi.org/10.1039/d0cs00098a

Myers, J. L., & Well, A. D. (2002). *Research Design & Statistical Analysis* (2nd ed.). Routledge.

Ng, W. L., Li, H. C., Lau, K. M., Chan, A. K. N., Lau, C. B. S., & Shing, T. K. M. (2017). Concise and Stereodivergent Synthesis of Carbasugars Reveals Unexpected

Structure-Activity Relationship (SAR) of SGLT2 Inhibition. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-05895-9

Niu, Y., Liu, R., Guan, C., Zhang, Y., Chen, Z., Hoerer, S., Nar, H., & Chen, L. (2021). Structural basis of inhibition of the human SGLT2–MAP17 glucose transporter. *Nature*, *601*(7892), 280–284. https://doi.org/10.1038/s41586-021-04212-9

Nomura, S., Sakamaki, S., Hongu, M., Kawanishi, E., Koga, Y., Sakamoto, T., Yamamoto, Y., Ueta, K., Kimata, H., Nakayama, K., & Tsuda-Tsukimoto, M. (2010). Discovery of Canagliflozin, a Novel *C*-Glucoside with Thiophene Ring, as Sodium-Dependent Glucose Cotransporter 2 Inhibitor for the Treatment of Type 2 Diabetes Mellitus. *Journal of Medicinal Chemistry*, *53*(17), 6355–6360. https://doi.org/10.1021/jm100332n

Omer, S. E., Ibrahim, T. M., Krar, O. A., Ali, A. M., Makki, A. A., Ibraheem, W., & Alzain, A. A. (2022). Drug repurposing for SARS-CoV-2 main protease: Molecular docking and molecular dynamics investigations. *Biochemistry and Biophysics Reports*, *29*, 101225. https://doi.org/10.1016/j.bbrep.2022.101225

Park, H., Jung, H. Y., Mah, S., & Hong, S. (2018). Systematic Computational Design and Identification of Low Picomolar Inhibitors of Aurora Kinase A. *Journal of Chemical Information and Modeling*, *58*(3), 700–709. https://doi.org/10.1021/acs.jcim.7b00671

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, *9*(3), 203–214. https://doi.org/10.1038/nrd3078

Pedregosa, F. P., Varoquaux, G. V., Gramfort, A. G., Michel, V. M., Thirion, B. T., Grisel, O. G., Blondel, M. B., Prettenhofer, P. P., Weiss, R. W., Dubourg, V. D., Vanderplas, J. V., Passos, A. P., Cournapeau, D. C., Brucher, M. B., Perrot, M. P., & Duchesnay, E. D. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Pharmacoinformatics Research Group. (2021, November 29). *GitHub - PharminfoVienna/sandbox: The PharminfoVienna Sandbox is a tool combining data gathering and model building of classification models.* GitHub. Retrieved August 12, 2022, from https://github.com/PharminfoVienna/sandbox

Pharmacoinformatics Research Group. (2022). *GitHub - PharminfoVienna/Retraining_Notebook: Using Jupyter Notebooks for Re-training Machine Learning Models*. GitHub. Retrieved August 12, 2022, from https://github.com/PharminfoVienna/Retraining_Notebook

Phlorizin | CAS:60–81-1. (n.d.). Manufacturer ChemFaces. Retrieved August 10, 2022, from https://www.chemfaces.com/natural/Phlorizin-CFN97038.html

Preisach, C., Burkhardt, H., Schmidt-Thieme, L., & Decker, R. (2008). *Data Analysis, Machine Learning and Applications*. Springer Publishing.

Scheepers, A., Joost, H., & Schurmann, A. (2004). The glucose transporter families SGLT and GLUT: molecular basis of normal and aberrant function. *Journal of Parenteral and Enteral Nutrition*, *28*(5), 364–371. https://doi.org/10.1177/0148607104028005364

Schrödinger Release 2021-1: Glide, Schrödinger, LLC, New York, NY, 2021.

Schrödinger Release 2021-1: Induced Fit Docking protocol; Glide, Schrödinger, LLC, New York, NY, 2021; *Prime*, Schrödinger, LLC, New York, NY, 2021.

Schrödinger Release 2021-1: LigPrep, Schrödinger, LLC, New York, NY, 2021.

Schrödinger Release 2021-1: Maestro, Schrödinger, LLC, New York, NY, 2021.

Schrödinger Release 2021-1: Prime, Schrödinger, LLC, New York, NY, 2021.

seaborn.kdeplot — seaborn 0.11.2 documentation. (n.d.). Seaborn. Retrieved August 10, 2022, from https://seaborn.pydata.org/generated/seaborn.kdeplot.html

Sethi, A., Joshi, K., Sasikala, K., & Alvala, M. (2020). Molecular Docking in Modern Drug Discovery: Principles and Recent Applications. *Drug Discovery and Development - New Advances*. https://doi.org/10.5772/intechopen.85991

Shahlaei, M. (2013). Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study. *Chemical Reviews*, *113*(10), 8093–8103. https://doi.org/10.1021/cr3004339

Shelley, J. C., Cholleti, A., Frye, L. L., Greenwood, J. R., Timlin, M. R., & Uchimaya, M. (2007). Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *Journal of Computer-Aided Molecular Design*, *21*(12), 681–691. https://doi.org/10.1007/s10822-007-9133-z

Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., & Farid, R. (2005). Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *Journal of Medicinal Chemistry*, *49*(2), 534–553. https://doi.org/10.1021/jm050540c

Shubrook, J., Baradar-Bokaie, B., & Adkins, S. (2015). Empagliflozin in the treatment of type 2 diabetes: evidence to date. *Drug Design, Development and Therapy*, 5793. https://doi.org/10.2147/dddt.s69926

Spearman rank correlation coefficient. (n.d.). Atozee. Retrieved August 10, 2022, from http://atozee.co.uk/S151/spearman1.html

Triballeau, N., Acher, F., Brabet, I., Pin, J. P., & Bertrand, H. O. (2005). Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *Journal of Medicinal Chemistry*, *48*(7), 2534–2547. https://doi.org/10.1021/jm049092j

Truchon, J. F., & Bayly, C. I. (2007). Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *Journal of Chemical Information and Modeling*, *47*(2), 488–508. https://doi.org/10.1021/ci600426e

Tsimihodimos, V., Filippas-Ntekouan, S., & Elisaf, M. (2018). SGLT1 inhibition: Pros and cons. *European Journal of Pharmacology*, *838*, 153–156. https://doi.org/10.1016/j.ejphar.2018.09.019

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Whaley, J., Tirmenstein, Reilly, Poucher, Saye, Parikh, & List. (2012). Targeting the kidney and glucose excretion with dapagliflozin: preclinical and clinical evidence for SGLT2 inhibition as a new option for treatment of type 2 diabetes mellitus. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, 135. https://doi.org/10.2147/dmso.s22503

Wright, E. M. (2020). SGLT2 and cancer. *Pflügers Archiv - European Journal of Physiology*, *472*(9), 1407–1414. https://doi.org/10.1007/s00424-020-02448-4

Wright, E. M., Loo, D. D. F., & Hirayama, B. A. (2011). Biology of Human Sodium Glucose Transporters. *Physiological Reviews*, *91*(2), 733–794. https://doi.org/10.1152/physrev.00055.2009

6 Appendix

6.1 Complete redocking results

In addition to the results shown in chapter 3.2, the pose of the co-crystallized Empagliflozin and the resulting pose of the redocking using the Standard Precision mode are presented in figures 27 and 28. Furthermore, the interactions as computed by the IFD – Standard Sampling, IFD – Extended Sampling and the Standard Precision modes are presented in the figures 29-31.

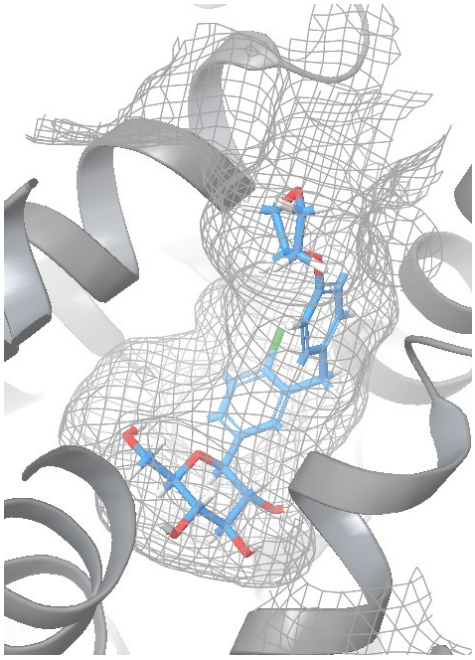


Figure 27: Empagliflozin at its binding site. The co-crystallized pose as elucidated by cryo-EM is shown in blue and binding site is represented by a grey mesh.

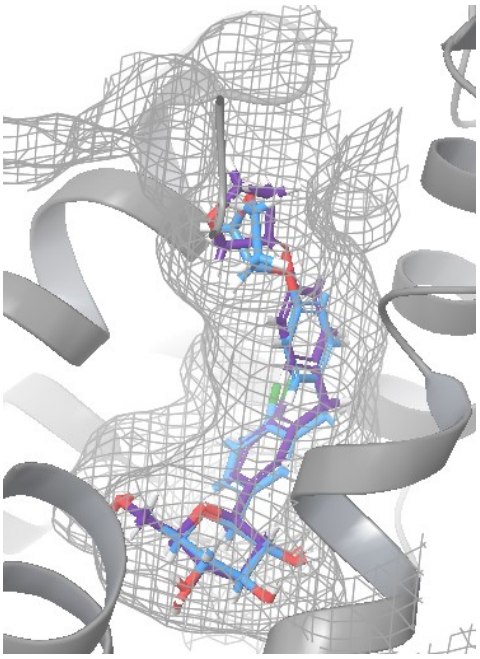


Figure 28: Redocked pose of Empagliflozin using the Standard Precision mode (purple) alongside the co-crystallized pose (blue)

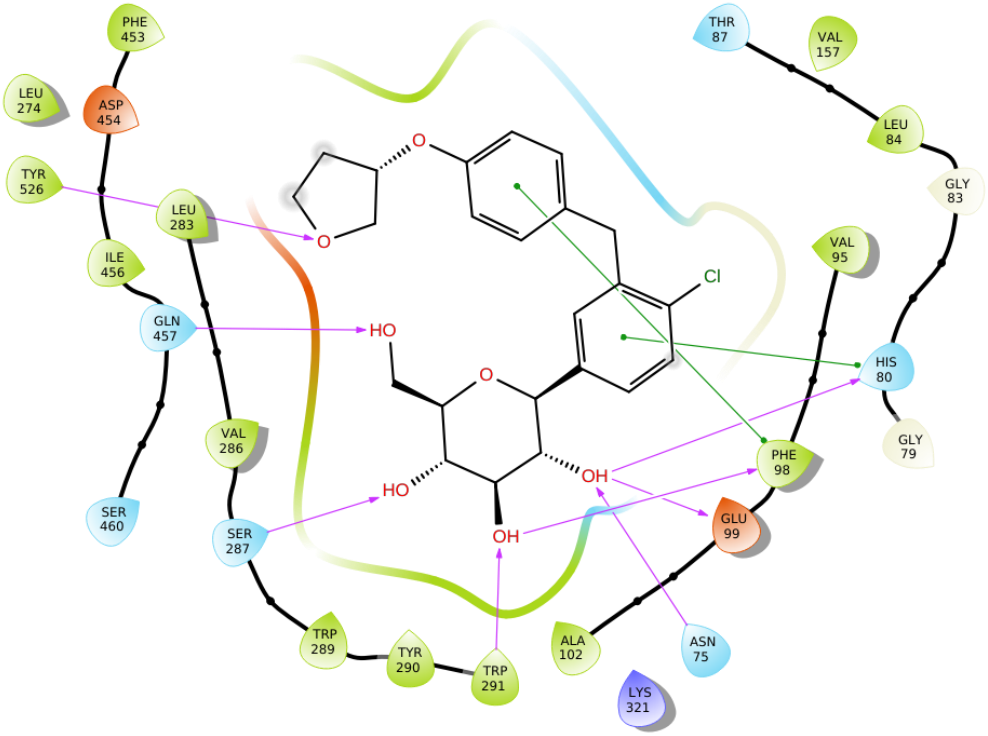


Figure 29: Interactions of the redocked empagliflozin with SGLT2 as computed by the Standard Precision mode.

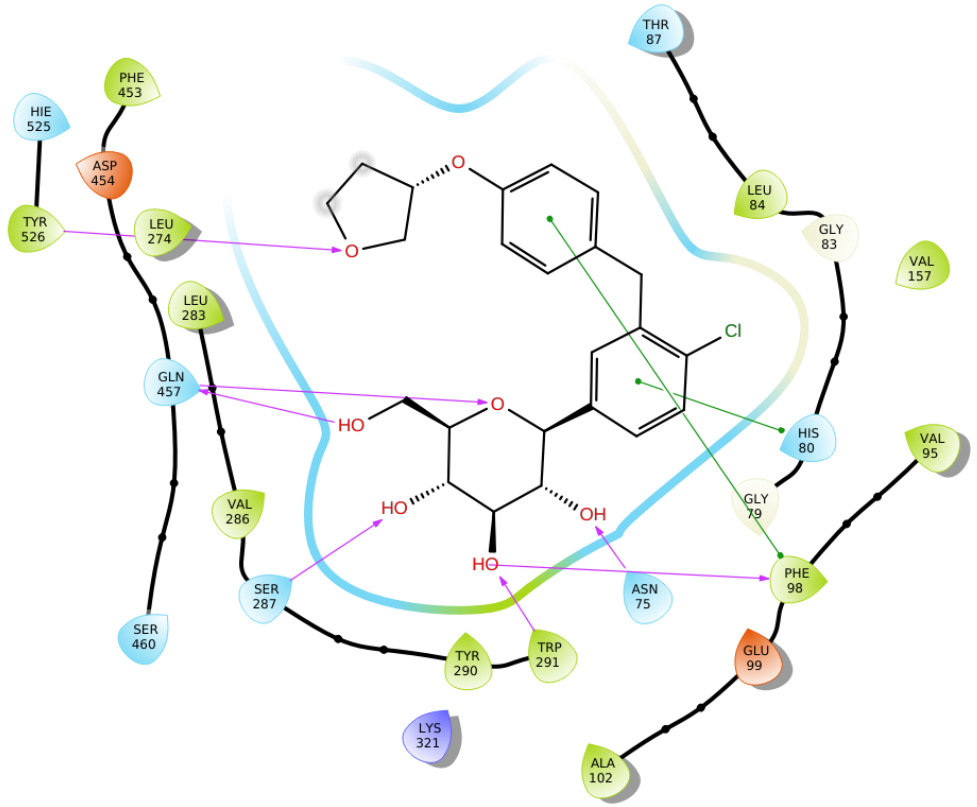


Figure 31: Interactions of the redocked empagliflozin with SGLT2 as computed by the IFD – Standard Sampling mode.

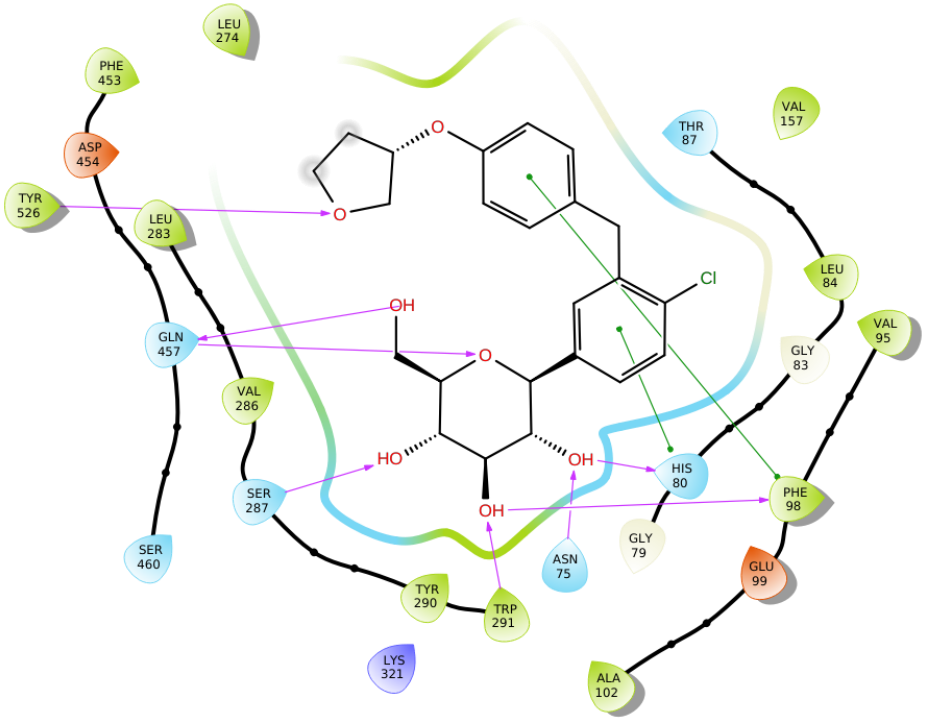


Figure 30: Interactions of the redocked empagliflozin with SGLT2 as computed by the IFD – Extended Sampling mode.

6.2 Complete docking results

The complete list of the docked structures with unknown activity (Kirchweger, Rollinger & Kowalska, 2022) and their Docking Scores from the IFD run and the VSW run are shown in the table below.

Induced Fit Docking - Standard	
Compound	Docking Score
GJB1244	-12.85
GJB1141	-12.816
GJB1230_hydrolized	-12.23
GJB1093	-11.999
GJB407	-11.866
GJB1098	-11.763
GJB1224	-11.71
GJB1097	-11.624
GJB268	-11.274
GJB1182	-11.241
GJB1179	-11.2
GJB1099	-11.13
GJB1096	-11.114
GJB1126	-11.083
GJB393	-11.032
GJB1167	-11.023
GJB539	-10.984
GJB406	-10.97
GJB499	-10.935
GJB392	-10.89
GJB1129	-10.811
GJB916	-10.667
GJB437	-10.644
GJB493	-10.598
GJB1178	-10.587
GJB1172	-10.49
GJB1177-1	-10.44
GJB1023	-10.321
GJB1248	-10.245
GJB431	-10.129
GJB1127	-10.037
GJB426	-9.588
GJB1101	-9.542
GJB1231	-9.526
GJB1230	-9.521
GJB1231_hydrolized	-9.446
GJB1173	-9.236
GJB1117	-9.079
GJB1128	-8.875
GJB1100	-8.582
GJB1225	-8.478
GJB1201	-7.785

Table 8: Results of the Induced Fit Docking – Standard Sampling applied to ligands with unknown activity

Virtual Screening Workflow	
Compound	Docking Score
GJB1224	-9.338
GJB1244	-8.209
GJB1182	-8.106
GJB1141	-8.043
GJB1096	-8.029
GJB407	-7.811
GJB1126	-7.697
GJB392	-7.606
GJB916	-7.583
GJB539	-7.53
GJB1098	-7.527
GJB1128	-7.476
GJB431	-7.461
GJB1099	-7.403
GJB1117	-7.357
GJB1097	-7.326
GJB1231	-7.32
GJB1093	-7.313
GJB1101	-7.311
GJB493	-7.294
GJB499	-7.248
GJB1100	-7.248
GJB406	-7.233
GJB1201	-7.221
GJB1248	-7.174
GJB1129	-7.12
GJB1127	-7.011
GJB1023	-6.948
GJB1230_hydrolized	-6.931
GJB1177-1	-6.923
GJB1179	-6.911
GJB1230	-6.893
GJB1178	-6.824
GJB1173	-6.813
GJB268	-6.782
GJB426	-6.759
GJB1225	-6.755
GJB1231_hydrolized	-6.718
GJB1172	-6.566
GJB1167	-6.557
GJB393	-6.548
GJB437	-6.032

Table 9: Results of the Virtual Screening Workflow using the Standard Precision mode applied to ligands with unknown activity

6.3 Complete machine learning results

In the figures below, the performances of the machine learning models are presented. The confusion matrices and ROC curves at pChEMBL thresholds of 7 and 8 were computed using scikit-learn (Pedregosa et al., 2011).

pChEMBL Threshold 7	Retraining LR	Retraining SVM	Retraining RF	Retraining KNN	Challenge SVM	Challenge RF	Challenge KNN
Accuracy	0.792	0.775	0.805	0.801	0.84	0.8	0.83
Sensitivity	0.936	0.803	0.873	0.919	0.85	0.81	0.88
Specificity	0.397	0.698	0.619	0.476	0.82	0.75	0.69
Balanced accuracy	0.667	0.751	0.746	0.698	0.84	0.78	0.78
F1 score	0.869	0.84	0.868	0.871	0.89	0.86	0.89
AUC	0.667	0.751	0.746	0.698	0.84	0.78	0.78
Precision	0.81	0.88	0.863	0.828	0.93	0.9	0.89
MCC	0.41	0.472	0.497	0.449	0.63	0.52	0.56
Recall	0.936	0.803	0.873	0.919	0.85	0.81	0.88

pChEMBL Threshold 8	Retraining LR	Retraining SVM	Retraining RF	Retraining KNN	Challenge SVM	Challenge RF	Challenge KNN
Accuracy	0.722	0.743	0.84	0.772	0.85	0.86	0.82
Sensitivity	0.692	0.76	0.817	0.798	0.867	0.867	0.876
Specificity	0.744	0.729	0.857	0.752	0.83	0.855	0.775
Balanced accuracy	0.718	0.744	0.837	0.775	0.85	0.86	0.83
F1 score	0.686	0.721	0.817	0.755	0.84	0.86	0.82
AUC	0.718	0.744	0.837	0.775	0.85	0.86	0.83
Precision	0.679	0.687	0.817	0.716	0.82	0.84	0.78
MCC	0.436	0.485	0.674	0.546	0.7	0.72	0.65
Recall	0.692	0.76	0.817	0.798	0.87	0.87	0.88

Figure 32: Performances of the generated machine learning models using a threshold of 7 and 8. Various relevant metrics for the assessment of classification models are shown, which were calculated after applying the generated models to a test set.

6.3.1. Sandbox notebook

Support Vector Machine – Threshold 7

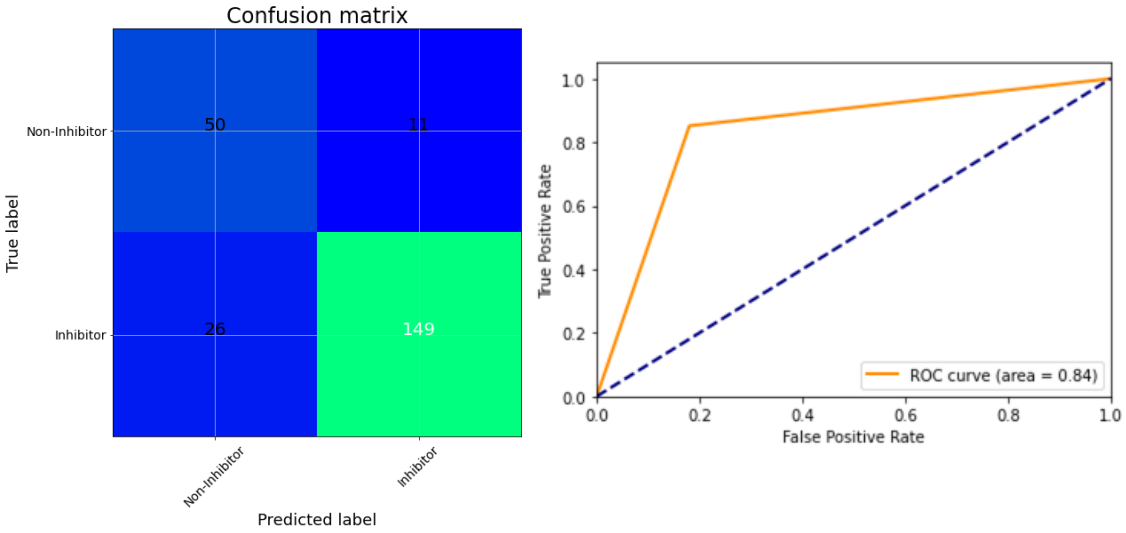


Figure 33: Confusion matrix and ROC curve of a SVM model, computed with a pChEMBL threshold of 7, after application to a test set.

Support Vector Machine – Threshold 8

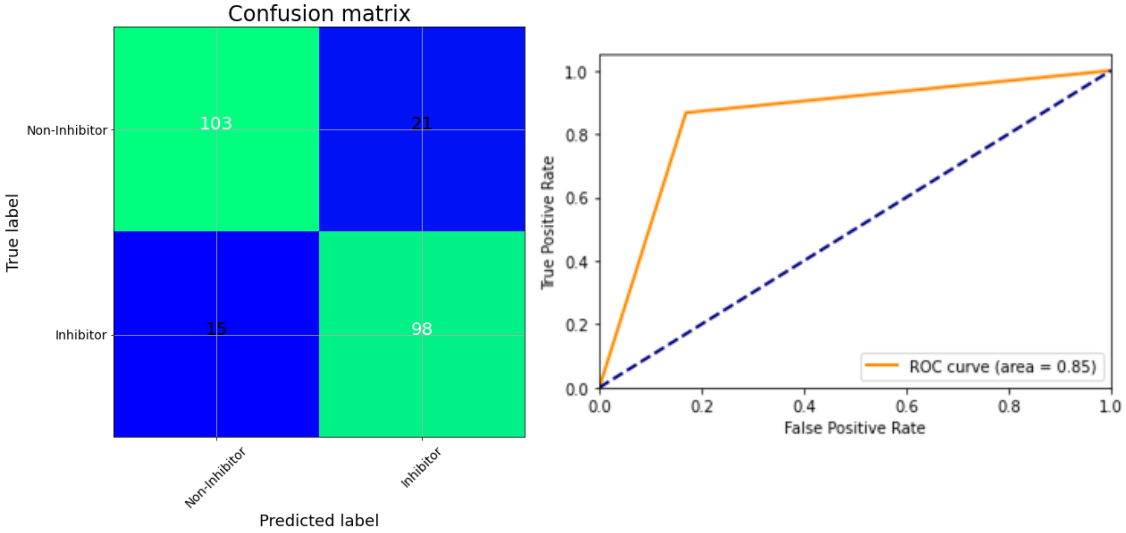


Figure 34: Confusion matrix and ROC curve of a SVM model, computed with a pChEMBL threshold of 8, after application to a test set.

Random Forest – Threshold 7

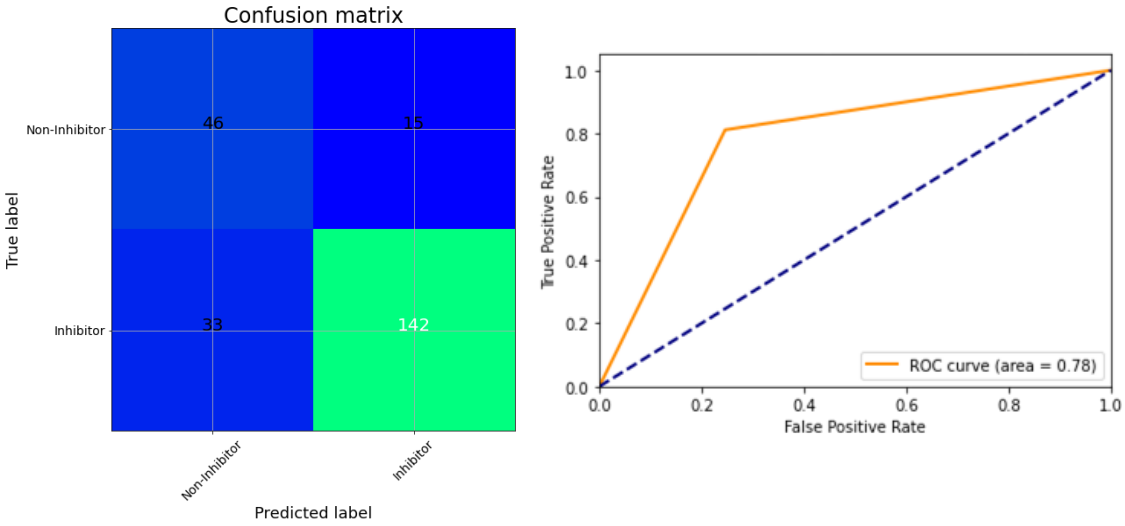


Figure 35: Confusion matrix and ROC curve of a RF model, computed with a pChEMBL threshold of 7, after application to a test set.

Random Forest – Threshold 8

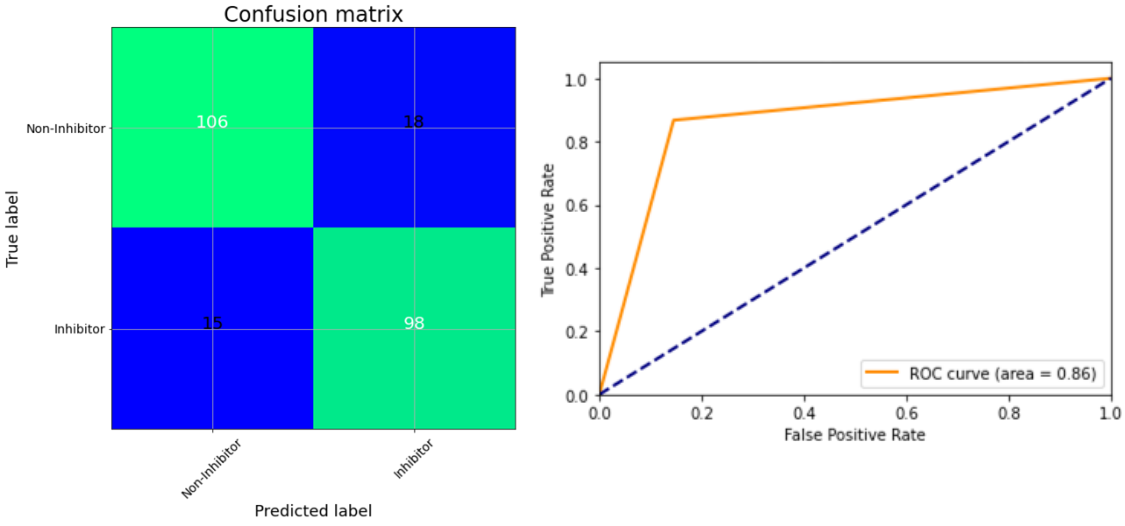


Figure 36: Confusion matrix and ROC curve of a RF model, computed with a pChEMBL threshold of 8, after application to a test set.

K-nearest Neighbors – Threshold 7

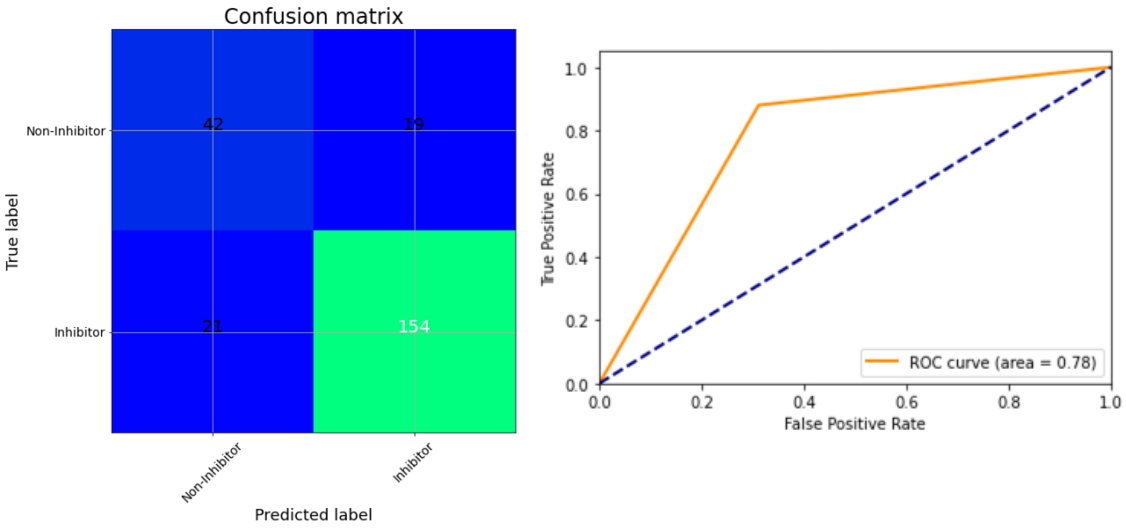


Figure 37: Confusion matrix and ROC curve of a KNN model, computed with a pChEMBL threshold of 7, after application to a test set.

K-nearest Neighbors – Threshold 8

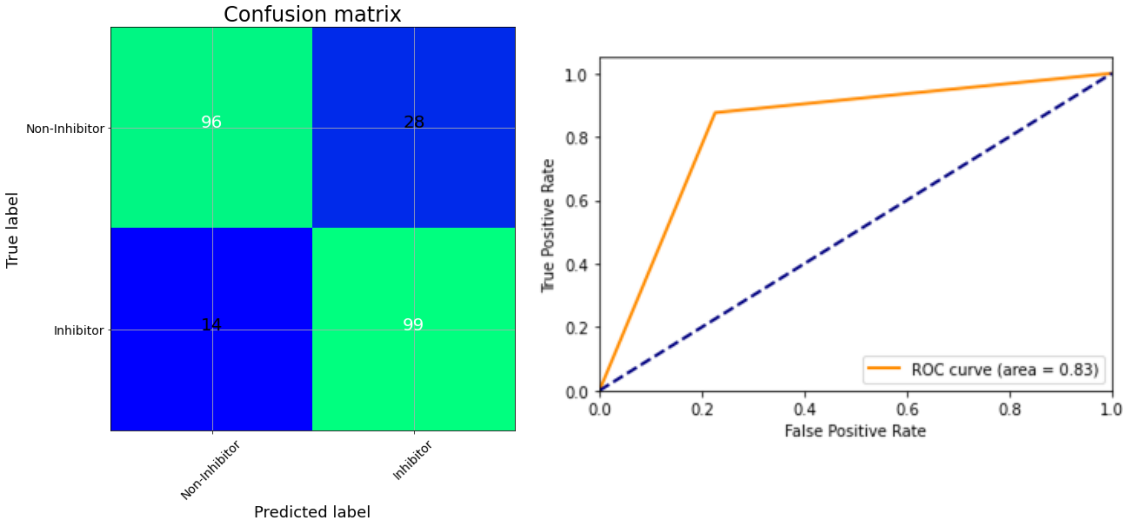


Figure 38: Confusion matrix and ROC curve of a KNN model, computed with a pChEMBL threshold of 8, after application to a test set.

6.3.2 Retraining notebook

Logistic Regression – Threshold 7

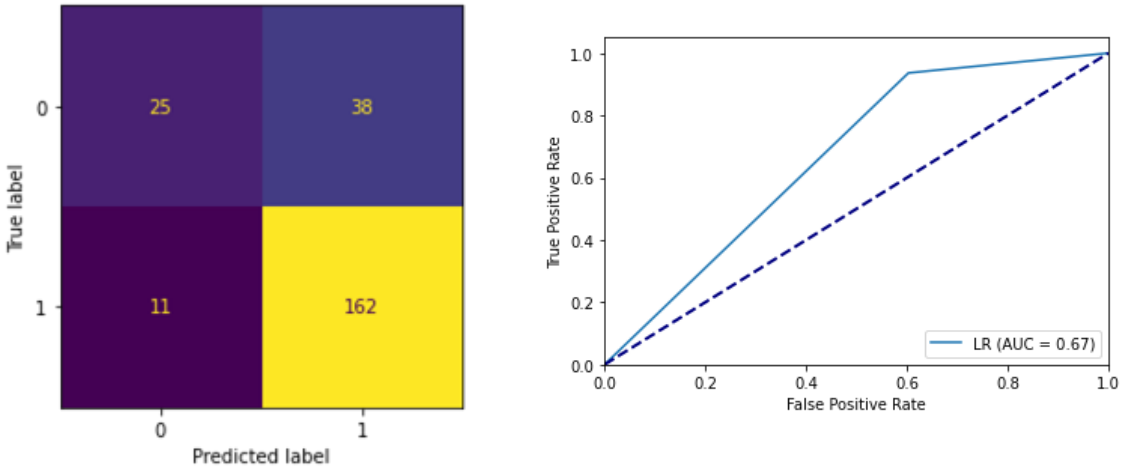


Figure 39: Confusion matrix and ROC curve of a LR model, computed with a pChEMBL threshold of 7, after application to a test set.

Logistic Regression – Threshold 8

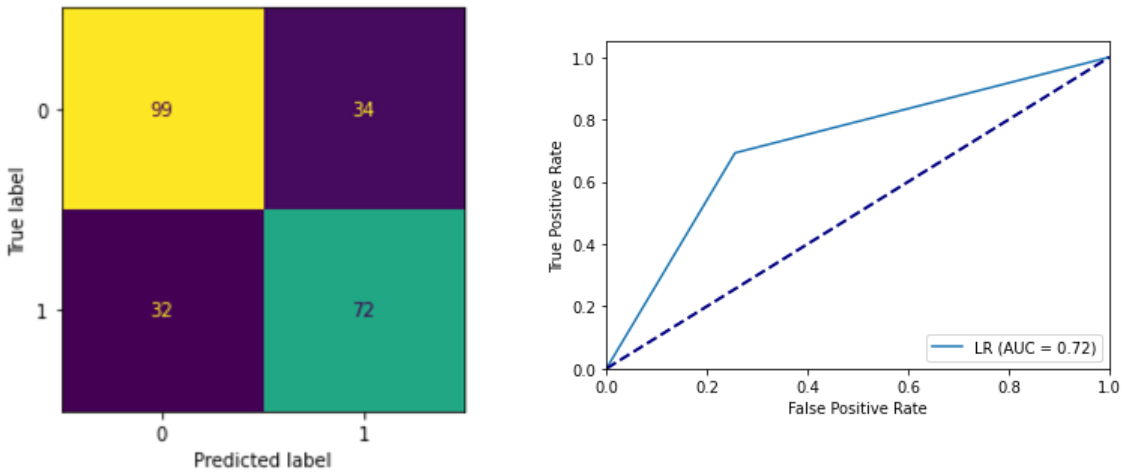


Figure 40: Confusion matrix and ROC curve of a LR model, computed with a pChEMBL threshold of 8, after application to a test set.

Support Vector Machine -Threshold 7

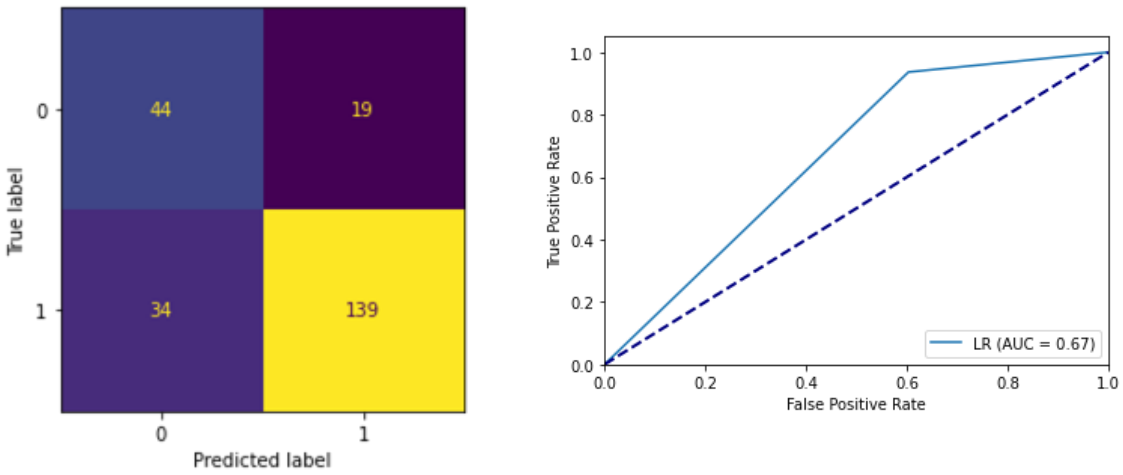


Figure 41: Confusion matrix and ROC curve of a SVM model, computed with a pChEMBL threshold of 7, after application to a test set.

Support Vector Machine – Threshold 8

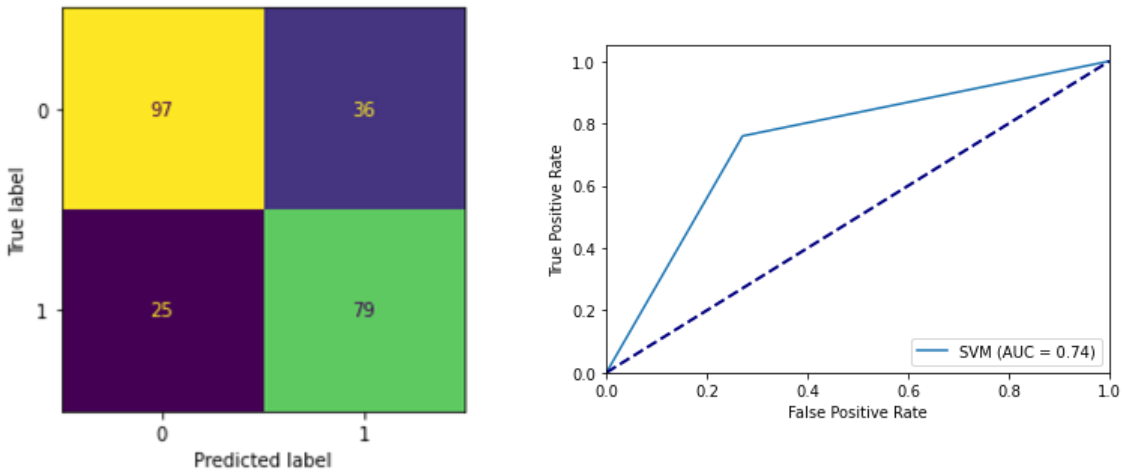


Figure 42: Confusion matrix and ROC curve of a SVM model, computed with a pChEMBL threshold of 8, after application to a test set.

Random Forest – Threshold 7

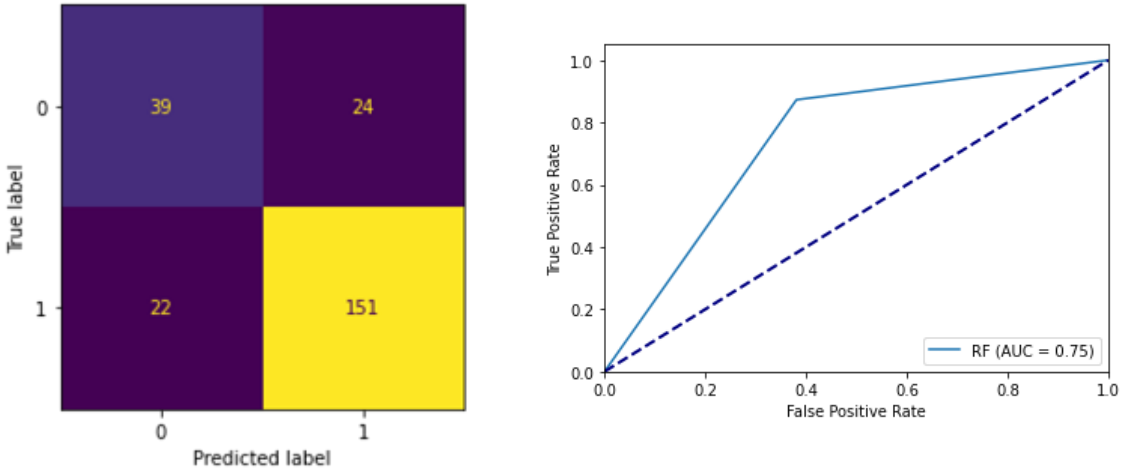


Figure 43: Confusion matrix and ROC curve of a RF model, computed with a pChEMBL threshold of 7, after application to a test set.

Random Forest – Threshold 8

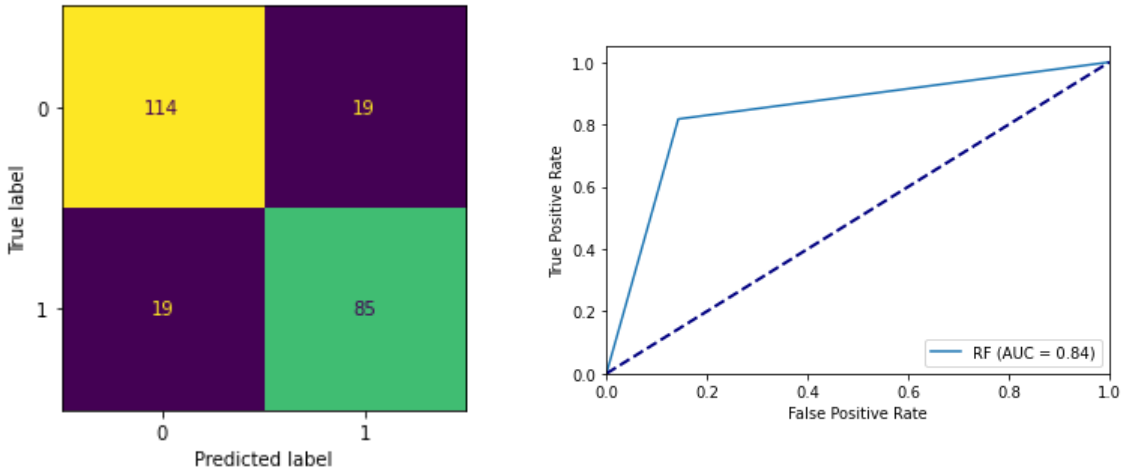


Figure 44: Confusion matrix and ROC curve of a RF model, computed with a pChEMBL threshold of 8, after application to a test set.

K-nearest Neighbors – Threshold 7

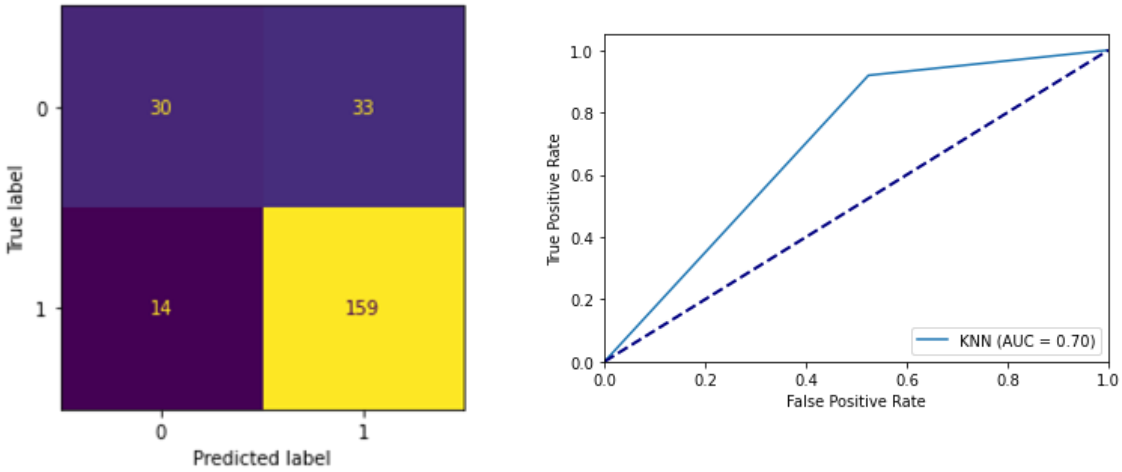


Figure 45: Confusion matrix and ROC curve of a KNN model, computed with a pChEMBL threshold of 8, after application to a test set.

K-nearest Neighbors – Threshold 8

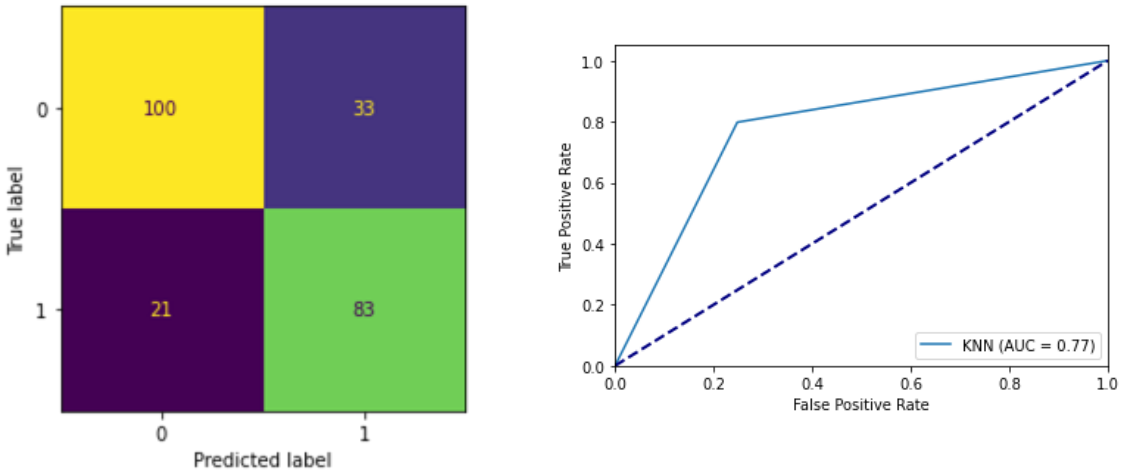


Figure 46: Confusion matrix and ROC curve of a KNN model, computed with a pChEMBL threshold of 8, after application to a test set.

6.4 Abstract

The aim of this master's thesis is to provide an *in-silico*-based method, mainly driven by the use of molecular docking, and QSAR modelling as an additional tool, for the analysis of the sodium/glucose co-transporter 2 (SGLT2) and the prediction of the activity of its potential inhibitors, which have been evolving to an important contribution to the treatment of diabetes mellitus.

To attain this objective, a classification model based on the Docking Scores obtained from a docking based virtual screening was created. Furthermore, the ability of various docking programs and their scoring functions to create compound rankings correlating to the ranking by activity was assessed. Finally, as an additional tool for the evaluation of results attained by the structure based approaches, a number of machine learning based QSAR models for SGLT2 inhibitors were generated and their performances were compared.

The methods developed for the analysis of the activity of potential inhibitors were subsequently applied to a number of compounds with unknown activity in order to predict their ability to inhibit SGLT2.

6.5 Zusammenfassung

Das Ziel dieser Masterarbeit war die Entwicklung einer *in-silico*-Methode zur Analyse des Natrium/Glucose-Co-Transporter 2 (SGLT2) und Vorraussage der Aktivität von potenziellen SGLT2-Inhibitoren, die sich zu einem wichtigen Teil der Behandlung von Diabetes Mellitus entwickelt haben. Dieser Prozess wurde hauptsächlich durch die Verwendung von Molecular Docking sowie QSAR Modelling als ein zusätzliches Werkzeug angetrieben.

Um dieses Ziel zu erreichen wurde ein Klassifikationsmodell basierend auf den Docking Scores eines Virtual Screenings entwickelt. Zusätzlich wurde die Fähigkeit der verschiedenen Docking-Programme und deren Scoring-Funktionen analysiert, eine Rangfolge für Moleküle zu erschaffen, die mit der Rangfolge korreliert, die von den Aktivitäten der Moleküle vorgegeben wurde. Als ein zusätzliches Werkzeug für die Evaluierung der Ergebnisse der Strukturbasierten Methoden wurden abschließend einige Machine-Learning-basierte QSAR-Modelle für SGLT2-Inhibitoren entwickelt und verglichen.

Die entwickelten Methoden für die Analyse der Aktivität von potenziellen Inhibitoren wurden anschließend auf einige Verbindungen mit unbekannten Aktivitäten angewendet, um deren Fähigkeit, SGLT2 zu hemmen, vorauszusagen.