Fachhochschul-Masterstudiengang
DATA SCIENCE UND ENGINEERING
4232 Hagenberg, Österreich

Automatic identification of important interactions and interaction-frequency-based scoring in protein-ligand complexes

Masterarbeit

zur Erlangung des akademischen Grades
Master of Science in Engineering

Eingereicht von

Micha Johannes Birklbauer, BSc.

Betreuerinnen: Univ. Prof. Dr. Daniela Schuster, Institut für Pharmazie, Paracelsus Medizinische Privatuniversität, Salzburg
Mag. Veronika Temml, PhD, Institut für Pharmazie, Paracelsus Medizinische Privatuniversität, Salzburg
Begutachter: FH-Prof. MMag. Dr. Gerald Lirk

August 2021

Declaration

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information have been used, they have been indicated as such and properly acknowledged. I further declare that this or similar work has not been submitted for credit elsewhere. This printed copy is identical to the submitted electronic version.

St. Martin bei Traun, August 31, 2021

Micha Johannes Birklbauer

Acknowledgements

First and foremost I want to thank my supervisors Mag. Veronika Temml, PhD, and Univ. Prof. Dr. Daniela Schuster, who not only came up with the idea and topic of this thesis but provided continuous support and insight throughout the research process. I am deeply thankful for their supervision and very much enjoyed working together with them.

Secondly, I would like to thank my university supervisor FH-Prof. MMag. Dr. Gerald Lirk for his teachings, his supervision and his input on this thesis.

Last but not least I want to thank my family and friends for their continuous emotional support.

Abstract

Molecular docking is an important tool in virtual screening for the discovery and design of new active agents for drug usage. The docking process is influenced by how well molecules fit in the binding site and which interactions occur between the protein and the ligand. Detection of these interactions can be automated with tools like the Protein-Ligand Interaction Profiler (PLIP) by PharmAI. However, identification and assessment of the importance of the different interactions in a protein-ligand complex is still a manual task that requires additional experimental data or domain knowledge about the target. The goals of this thesis are twofold: Firstly, to automatically identify those interactions that have a significant influence on ligand binding, and secondly, to develop a novel scoring function which is able to discriminate active molecules from inactive ones if possible. The underlying data basis were selected targets of the Directory of Useful Decoys: Enhanced (DUD-E) and available structures from the Protein Data Bank (PDB). Specifically 11 targets were analysed: 11-Beta-Hydroxysteroid Dehydrogenase 1 (HSD11B1), Acetylcholinesterase (ACHE), Coagulation Factor XA (FXA), Cyclooxygenase 1 and 2 (COX1/COX2), Dipeptidyl Peptidase IV (DPP4), Monoamine Oxidase B (MAOB), P38 Mitogen-Activated Protein Kinase 14 (MAPK14), Phosphodiesterase 5 (PDE5A), Protein-Tyrosine Phosphatase 1B (PTP1B) and Soluble Epoxide Hydrolase (SEH). PLIP is used to extract interactions present in a protein-ligand complex and the respective interaction's frequency is measured across all target structures. Cofactors were excluded from the analysis and hydrophobic interactions were only counted once per residue. Additionally, when analysing docking poses only the pose that had the most interactions contributed to the calculation. Furthermore, four different scoring functions that are based on the differences in frequencies between active and inactive compounds were established and their performance was assessed on an independent test partition containing unseen ligands. The results show that interactions which are known from literature to be important for ligand binding are found for all targets except ACHE, in many cases among the top ranked interactions in terms of frequency. This behaviour implies a relationship between interaction frequency and the interaction's significance in ligand binding. Interaction-frequency-based scoring was tested in five targets and performed above baseline accuracy in four of the five targets. In all targets scoring led to an enrichment of active compounds and false positive rates fluctuated between 0 and 33%. Interaction frequency analysis and interaction-frequency-based scoring could therefore be used as supporting tools in virtual screening to further enhance results.

Kurzfassung

Molecular Docking ist ein wichtiges Werkzeug im Entdeckungs- und Entwicklungsprozess neuer Medikamente. Die Passgenauigkeit des Liganden in der Bindetasche und die Interaktionen, die er mit dem Protein eingeht, sind maßgebliche Faktoren, die das Docking beeinflussen. Das Auffinden und Charakterisieren eben jener Interaktionen kann mit Programmen wie dem Protein-Ligand Interaction Profiler (PLIP) von PharmAI automatisiert werden. Das Identifizieren und Bestimmen, wie wichtig die einzelnen Interaktionen für die Bindung sind, ist jedoch immer noch ein manuelles Unterfangen, das ausschlaggebende experimentelle Daten oder Domänen Know-how voraussetzt. Das Ziel dieser Arbeit teilte sich in zwei Aspekte: (i) Die automatische Identifikation von bindungswichtigen Interaktionen und (ii) die Entwicklung einer neuen Bewertungsfunktion basierend auf den Frequenzdaten der einzelnen Interaktionen, um aktive von inaktiven Molekülen unterscheiden zu können. Das Fundament bildeten Daten aus dem Directory of Useful Decoys: Enhanced (DUD-E) und Protein-Strukturen aus der Protein Data Bank. Die folgenden 11 Proteine wurden analysiert: 11-Beta-Hydroxysteroid Dehydrogenase 1 (HSD11B1), Acetylcholinesterase (ACHE), Coagulation Factor XA (FXA), Cyclooxygenase 1 und 2 (COX1/COX2), Dipeptidyl Peptidase IV (DPP4), Monoamine Oxidase B (MAOB), P38 Mitogen-Activated Protein Kinase 14 (MAPK14), Phosphodiesterase 5 (PDE5A), Protein-Tyrosine Phosphatase 1B (PTP1B) und Soluble Epoxide Hydrolase (SEH). Die Software PLIP wurde genutzt, um die einzelnen Interaktionen aus den Protein-Strukturen zu extrahieren, und die Frequenz jeder Interaktion wurde anschließend aus allen Strukturen eines Proteins ermittelt. Cofaktoren wurden aus dem Analyseprozess ausgeschlossen und hydrophobe Interaktionen wurden nur einmal pro Residue gezählt. In der Analyse von Docking-Daten wurde pro Ligand nur jene Pose miteinbezogen, die die größte Anzahl an Interaktionen aufwies. Des Weiteren wurden vier Bewertungsfunktionen entwickelt, die auf den Unterschieden zwischen den Frequenzen in aktiven und inaktiven Molekülen basierten. Zur Evaluierung der Bewertungsfunktionen wurde ein eigenständiger Testdatensatz herangezogen, der ausschließlich aus für die Bewertungsfunktionen unbekannten Liganden bestand. Die Ergebnisse zeigen, dass literaturbekannte, bindungswichtige Interaktionen in fast allen Protein-Auswertungen vorkommen, die einzige Ausnahme bildet ACHE. In vielen Fällen finden sich diese Interaktionen sogar unter den Interaktionen mit den höchsten Frequenzen und ein Zusammenhang zwischen Frequenz und Bindungssignifikanz liegt daher nahe. Die Bewertungsfunktionen wurden an fünf der Proteine getestet und in vier Fällen toppte die Performance die Baseline Accuracy. In allen fünf Proteinen kam es zu einem Enrichment der aktiven Moleküle und die Falsch-Positiv-Rate fluktuierte zwischen 0 bis 33%. Die Analyse von Protein-Ligand-Interaktionen und deren Frequenz sowie darauf basierende Bewertungsfunktionen könnten daher in Zukunft den Entwicklungsprozess von Medikamenten unterstützen und die Ergebnisse vorhandener Werkzeuge wie Docking verbessern.

Table of Contents

List of Abbreviations

ACC – accuracy
ACHE – acetylcholinesterase
AI – artificial intelligence
ATP – adenosine triphosphate
AUC – area under the receiver operating characteristic curve
cAMP – cyclic adenosine monophosphate
CATS – chemically advanced template search
cGMP – cyclic guanine monophosphate
COX1 – cyclooxygenase 1
COX2 – cyclooxygenase 2
DOGS – design of genuine structures
DPP4 – dipeptidyl peptidase IV
DUD-E – Directory of Useful Decoys: Enhanced
EC number – Enzyme Commission number (in the PDB also Enzyme Classification Number)
EF – enrichment factor
FPR – false positive rate
FXA – coagulation factor Xa
GIP – glucose-dependent insulinotropic peptide
GLP-1 – glucagon-like peptide 1
HSD11B1 - 11β-hydroxysteroid dehydrogenase type 1
HTS – high-throughput screening
LBVS – ligand-based virtual screening
MAO – monoamine oxidase
MAOA – monoamine oxidase A
MAOB – monoamine oxidase B
MAPK – mitogen-activated protein kinase
MAPK14 – p38 mitogen-activated protein kinase 14
ML – machine learning
NMR – nuclear magnetic resonance
NSAID – nonsteroid anti-inflammatory drug
PDB – Protein Data Bank
PDE – phosphodiesterase
PDE5/PDE5A – phosphodiesterase 5
PIA – Protein Interaction Analyzer
PLIP – Protein-Ligand Interaction Profiler
PSO – particle swarm optimization
PTP1B – protein-tyrosine phosphatase 1B
REF – relative enrichment factor
ROC – receiver operating characteristic
SBVS – structure-based virtual screening

SEH – soluble epoxide hydrolase
VS – virtual screening
Ya/YA – yield of actives

List of Amino Acid Abbreviations

ALA (A) – alanine
ARG (R) – arginine
ASN (N) – asparagine
ASP (D) – aspartic acid (Aspartate)
CYS (C) – cysteine
GLN (Q) – glutamine
GLU (E) – glutamic acid (Glutamate)
GLY (G) – glycine
HIS (H) – histidine
ILE (I) – isoleucine
LEU (L) – leucine
LYS (K) – lysine
MET (M) – methionine
PHE (F) – phenylalanine
PRO (P) – proline
SER (S) – serine
THR (T) – threonine
TRP (W) – tryptophan
TYR (Y) – tyrosine
VAL (V) – valine

3-letter and 1-letter amino acid codes according to the IUPAC-IUB Joint Commission on Biochemical Nomenclature (IUPAC-IUB Joint Commission on Biochemical Nomenclature, 1984).

1. Introduction

The discovery of new drugs – or generally compounds that can be used for new drugs – is a both expensive and time consuming process. From the identification of the first promising molecules (leads) over to clinical trials and until market release it is estimated that this process takes about 14 years and costs 800 million US dollars (Lavecchia & Di Giovanni, 2013). Efforts to shorten this process and to minimize costs is an obvious and logical conclusion. For finding new leads two different approaches have been established in the past: High-throughput screening (HTS) and virtual screening (VS).

In HTS a large number of molecules are tested for their impact on a specific target, usually a protein, by addressing whether the molecule biochemically reacts with the target or not. This is done on so called HTS assays and compounds showing positive results (hits) are then used for further research. It is important that these positive hits are further analysed and re-confirmed as actually being positive because in case of a false positive a lot of time and money could potentially be wasted. Contrary, false negative results could mean that a valuable drug candidate will not be further considered, although this tends to be only an issue if there are no positives found at all. It should be emphasized however that the goal of HTS is not to find all potential candidates in a library collection of compounds but enough to have as set for initial discovery efforts. HTS is a time-consuming process, requires specific infrastructure and has low success rates of below 5%, yet it still has been the method of choice for the last 20 years (Kontoyianni, 2017).

One the other hand, while HTS is an experimental, *in vitro* approach the contrary is the case for VS which is a theoretical, *in silico* approach. In VS a digital library of chemically diverse compounds is screened for leads. Since VS is a computational method it is faster, more cost-efficient and less resource intensive than its counterpart HTS (Tang & Marshall, 2011). VS can be further divided into two sub-categories, namely ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS). In LBVS strategies the structure-activity data from a set of known, active ligands is used to identify possible targets for experimental evaluation. Among the methods used in LBVS are similarity and substructure searching, quantitative structure-activity relationships and 3D shape matching. On the other hand, SBVS makes use of the 3D structure of the biological target. Consequentially the structure of the target has to be either known beforehand or analysed via X-ray crystallography, NMR spectroscopy or computationally via homology modelling. Then a set of ligands is fitted into the binding site of the target (docking) and a score – usually based on the predicted binding affinity – is used to rank the ligands and determine if they are active or not (Lavecchia & Di Giovanni, 2013).

SBVS can be further sub-divided by distinguishing approaches that use rigid docking versus approaches that use flexible docking (McInnes, 2007). In rigid-body docking the ligand is searched in a six-dimensional translational or rotational space to fit in the binding pocket of the target protein. The complex is then evaluated in terms of shape of the fitted ligand and the binding site, as well as the electrostatic, van der Waals and Coulombic interactions that may

take place. The docking score is usually then the sum of these terms. The accuracy for rigid-body docking approaches tends to be much greater for bound complexes than in unbound complexes. Even though the structural differences between bound and unbound molecules are small they affect the docking accuracy noticeably (Pagadala, Syed & Tuszynski, 2017). Biological targets are in fact not rigid in nature but adjust dynamically. However, experimentally resolved structures of targets with bound or unbound ligands are isolated snapshots and do not reflect the flexibility that happens in nature (Kontoyianni, 2017). Consequentially new docking approaches were developed that allowed ligand or receptor flexibility. In the simple case the softening of van der Vaals potentials – also called "soft docking" – can allow small overlaps between the receptor and the ligand in the binding pocket. The downside is that this may lead to an increase in false positives as more diverse structures are allowed to bind (Lavecchia & Di Giovanni, 2013). The more sophisticated and most common approach in standard virtual docking studies is having a flexible ligand while fitting into a rigid receptor. Generally four different strategies are in use when docking a flexible ligand: The application of Monte Carlo methods; combinatorial search; ligand buildup algorithms, where ligands are built directly in the binding site of the protein; and site-mapping and fragment assembly which extends the ligand buildup approach by connecting molecular fragments to mapped functional groups in the binding site (Pagadala, Syed & Tuszynski, 2017). In nature, however, the receptor is also flexible and the binding site is altered according to the orientation of the ligand by movement of the side chains (Pagadala, Syed & Tuszynski, 2017). Therefore flexible receptor docking also became of scientific interest and first programs as well as theoretical approaches that are in constant development exist (Lavecchia & Di Giovanni, 2013). One of these approaches is ensemble docking where a ligand is docked into multiple conformations of the same protein (McInnes, 2007). The structures for ensemble docking are usually taken from the Protein Data Bank (PDB), if available, or from molecular dynamics simulations or from normal mode analyses (Berman et al., 2000; Kontoyianni, 2017).

In all SBVS approaches docking is followed up by scoring of the ligands. Even though predicting one or more potential binding poses is possible most of the times with available docking methods, identifying the correct binding pose and ranking the ligands are still challenging tasks which are tackled by scoring functions. Firstly, scoring functions aim to identify the energetically preferred pose out of a set of bound poses that were generated by the docking algorithm for a single ligand. Secondly, the scoring function is used to rank different docked ligands in order to discriminate between active and inactive compounds. Scoring functions are a major research topic in the docking community with many problems still to be overcome and procedures to be refined. To name an example, one of the difficulties of scoring functions stems from the fact that a lot of factors, like molecular interactions, are not easy to parameterize. Generally existing scoring functions can be divided into three broad groups: Force field-based scoring functions, knowledge-based scoring functions and empirical scoring functions. Additionally some scoring functions also combine these approaches (Lavecchia & Di Giovanni, 2013).

In force field-based scoring functions the binding free energy is estimated by the sum of the independent molecular mechanic force field potentials like Coulomb, van der Waals and hydrogen bonds. Furthermore, solvation and entropy contributions are also considered in some cases. On the other hand in empirical scoring functions interaction terms like hydrogen bonds and hydrophobic contacts are estimated by fitting the scoring function to the experimental binding affinity data of a training dataset of protein-ligand complexes. Subsequently the binding free energy for the docked ligands is calculated as the weighted sum of these terms. Thirdly, knowledge-based scoring functions are exclusively derived by statistically analysing the atom-pair frequencies of known 3D structures from protein-ligand complexes (Lavecchia & Di Giovanni, 2013).

Since this work makes use of an SBVS approach, current developments and state of the art in the field are described in the following section.

1.1 Current developments and state of the art

Today a variety of both open-source and commercial docking software exists. To name a few examples DOCK (Venkatachalam et al., 2003), AutoDock (Österberg et al., 2002), GOLD (Jones et al., 1997), LigandFit (Venkatachalam et al., 2003), Surflex (Jain, 2003) and Glide (Friesner et al., 2004) should be mentioned. Even though all programs share the same goal of accurately predicting the correct binding pose, they apply different strategies to do so. DOCK, for example, is driven by a shape-based algorithm while GOLD applies genetic algorithms. Glide makes use of systematic search techniques and LigandFit predicts docking poses via Monte Carlo simulation. Almost all of the currently available flexible docking software treats the receptor as rigid, GOLD being the only exception (Pagadala, Syed & Tuszynski, 2017). In 2016 Wang et al. evaluated the performance of ten different – both academic and commercial – docking programs on a dataset of 2002 protein-ligand complexes. They differentiated between sampling power, which was denoted as the accuracy of predicting the correct binding pose, and scoring power, which was defined as how accurately binding affinity is estimated. Finally, they concluded that GOLD and LeDock (Zhao & Caflisch, 2013) had the best sampling power with GOLD showing an accuracy of 59.8% for the top scored poses and LeDock yielding 80.8% accuracy for the best poses. AutoDock Vina (Trott & Olson, 2009) achieved the best scoring power for both the top scored and best poses (Wang et al., 2016).

Newer approaches include for instance the application of particle swarm optimization (PSO) algorithms as demonstrated in PSOVina. PSOVina extended AutoDock Vina's Broyden-Fletcher-Goldfarb-Shannon local search and achieved an execution time reduction of 51-60% compared to traditional AutoDock Vina (Ng et al., 2015). Furthermore, machine learning (ML) and artificial intelligence (AI) have become driving forces in computational biology with AI programs like AlphaFold by DeepMind even making headlines in the mainstream media (Senior, 2020; Müller-Jung, 2020). Especially in VS machine learning can be utilized in many different ways and at various different stages of the drug discovery process. Pham and Jain demonstrated in 2008 how a scoring function – specifically that of the docking software Surflex

– can be tuned by optimizing its parameters via multiple instance learning (Pham & Jain, 2008). Moreover, machine learning has become increasingly popular in LBVS for its ability to accurately quantify structure-activity relationships. Both regression and classification methods like Linear Regression, Nearest Neighbour, Naïve Bayesian classification, Support Vector Machines, Artificial Neural Networks and Decision Trees have been successfully applied. The goal of all these ML models is to learn from training data to discriminate between active and inactive molecules in order to find new molecules that interact with the target of interest. ML algorithms in VS are prone to the same risks as any ML approach and their performance is largely dependent on the quality of the underlying training data and how well they can deal with unbalanced datasets because inactive compounds are usually several factors more frequent than active compounds (Lavecchia & Di Giovanni, 2013).

One of the newest methods involving ML in computer-assisted drug discovery is the *de novo* design of active compounds based on natural template products which was recently demonstrated by the Institute of Pharmaceutical Sciences of ETH Zürich. The goal of this procedure is to discover molecules that mimic the function of the natural product but are easier to synthesize. In their approach they apply the so called DOGS (design of genuine structures) algorithm that constructs new molecules by combining molecular building blocks in accordance to a defined list of *in silico* chemical transformations. The process is optimized by a fitness function that is denoted as the pairwise molecular graph similarity between the generated molecule and the template compound. The similarity is measured in the CATS (chemically advanced template search) distance metric where a lower value symbolizes better similarity, therefore the fitness function is minimized. Exercising this strategy utilizing Marinopyrrole A as a template they were able to design a novel Cyclooxygenase-1 inhibitor (Friedrich et al., 2021).

Cycling back it should be noted here that this thesis builds on top of SBVS and protein-ligand docking by further analysing the produced results. The goals of this work are laid out in the following.

1.2 Goals

There were two objectives defined for this thesis:

- Firstly, the automatic identification of interactions that are important for binding in protein-ligand complexes derived either from experimental structures or from dockings.
- Secondly, the design of a novel scoring function which is based on the frequency of interactions found in docked protein-ligand complexes that is able to discriminate between active and inactive molecules.

Moreover, these goals are not independent from each other but strongly intertwined – frequency based scoring makes little sense with interactions that are not contributing to the binding between protein and ligand. Therefore the second goal can be viewed as an extension of the first.

1.3 Motivation

Improving the predictability of active compounds in VS is of utmost importance as new potential molecules for use in medical applications can be more efficiently detected, reducing the cost in both time and money of the drug discovery process. Developing new supporting tools that can enhance the predictions of molecular docking and improve false positive rates means less *in vitro* experiments and therefore can save lots of resources. Software like the Protein-Ligand Interaction Profiler (PLIP) by PharmaAI (Salentin et al., 2015) – which is also used as the basis of this thesis – can reliably detect the non-covalent interactions in protein-ligand complexes, however, automatically assessing the importance of these interactions as well as using the interaction frequency to predict active compounds seems to be a novel approach where little to no research was found that explores this direction. Combining molecular docking (or generally other VS approaches) with the information gained about interactions happening between the protein and the ligand could significantly improve results, especially in cases where structural data is available but the relationships between protein and ligand are not yet fully understood.

1.4 Interaction types

The current version of PLIP is able to characterize eight (originally seven on release) different protein-ligand interactions: Hydrogen bonds, water bridges, salt bridges, halogen bonds, hydrophobic interactions, pi-stacking, pi-cation interactions and metal complexation (Salentin et al., 2015). Because of their significance in this work they shall be shortly described here.

1.4.1 Hydrogen bonds

Hydrogen bonds play an important role in ligand binding and enzyme catalysis. Their bonding properties strongly influence the specificity of binding, transportation, absorption, distribution, metabolization and excretion of the respective molecules and therefore have to be considered in every drug design process. Furthermore, because hydrogen bonds are ubiquitous and flexible they are considered to be the most important physical interactions in biomolecules in aqueous solution (Williams & Ladbury, 2003).

A hydrogen bond is defined as an attractive interaction between a hydrogen atom – either from a molecule or a molecular fragment – that is attached to an atom that is more electronegative than H and another atom (or group of atoms) in the same or a different molecule. Hydrogen bonds are denoted as X–H ⋯ Y–Z where the three dots represent the bond, H is the hydrogen atom and X the more electronegative atom. X–H is called the hydrogen bond donor and Y (or Y–Z) the hydrogen bond acceptor, where Y is either a single atom or anion or in case of Y–Z a molecule or a fragment of a molecule where Y is bonded to Z. Atoms X and H form a covalent bond that is polarized and the strength of the hydrogen bond between H and Y is dependent on the electronegativity of X, higher electronegativity of X leads to a stronger hydrogen bond. The angle between X–H ⋯ Y is usually around 180° and the closer the angle is to 180°, the stronger is the hydrogen bond (Arunan et al., 2011).

The typical binding free energy in hydrogen bonds ranges from -2 kJ/mol (amide – amide in protein core) to -46 kJ/mol (squaramides ⋯ F^- in CH_3CN) (Biedermann & Schneider, 2016).

The role of hydrogen bonds in drugs has been thoroughly studied in the past, in fact Lipinski's rule of five states that a majority of orally administered drugs tend to form more than five but less than ten hydrogen bonds. However, naturally many exceptions exist (Lipinski, 2004).

1.4.2 Water bridges

Although "water bridge" is not a universally defined chemical term, it is used by Salentin et al. to denote water-bridged hydrogen bonds. If an atom in a protein forms a hydrogen bond with a water molecule and that same water molecule forms a hydrogen bond with an appropriate atom in the ligand, this interaction is categorized as a water bridge (Salentin et al., 2015).

1.4.3 Salt bridges

Together with hydrogen bonds salt bridges form the structural basis for molecular complexes (Biedermann & Schneider, 2016). Salt bridges are defined as ion pairs between two side chains of a protein. However, the term salt bridge is often also used to denote ion pairs in general – as in the case of protein-ligand binding where the paired ions are located at protein and ligand respectively. An ion pair is defined as a cation and anion that are located close enough in space that their electrostatic attraction is larger than the thermal energy available to separate them. Ion pairing is therefore classified as an electrostatic interaction (Anslyn & Dougherty, 2006). Typical binding free energy for salt bridges and ion pairs ranges from near 0 kJ/mol in ionic groups at protein surface to -20 kJ/mol for ionic groups in the protein core (Biedermann & Schneider, 2016).

1.4.4 Halogen bonds

Halogen bonds are defined as attractive interactions between an electrophilic region associated with a halogen atom in one molecule and a nucleophilic region in another or the same molecule. Halogen bonds are denoted similarly to hydrogen bonds as R–X ⋯ Y. R–X is in this case the halogen bond donor where X is any halogen atom with an electrophilic region and R is a group of atoms covalently bound to X. On the other hand Y is the halogen bond acceptor and is typically a molecule with at least one nucleophilic region. Halogen bond strength increases with decreasing electronegativity of X as well as increasing electron-withdrawing ability of R (Desiraju et al., 2013). Typical binding free energies of -1 kJ/mol to -19 kJ/mol have been observed for halogen bonds (Biedermann & Schneider, 2016).

1.4.5 Hydrophobic interactions

The tendency of hydrocarbons and lipophilic hydrocarbon-like groups in solutes to form intermolecular or intramolecular aggregates in an aqueous medium is called hydrophobic interaction. The name originates from the hydrophobic effect that describes the repulsion between water and hydrocarbons (Muller, 1994). The aggregation of molecular structures is explained by the reduction of solvent-accessible surface area. Hydrophobic interactions are weaker interactions than hydrogen bonds, salt bridges and halogen bonds with binding free energies around -1 kJ/mol to -3 kJ/mol per CH_2 (Biedermann & Schneider, 2016).

1.4.6 Pi-stacking

Pi-stacking or pi-pi interactions describe interactions between neighbouring aromatic rings. The pi electron density on most aromatic rings creates a quadrupole moment with partial negative charge above both aromatic faces and a partial positive charge around the periphery. This leads to attraction between the aromatic rings and to one of several possible alignments (stacking). However, the term pi-stacking (and pi-pi interaction) has been deemed not appropriate anymore by parts of the scientific community as this interaction seems to be not necessarily unique to aromatic molecules. Furthermore it is also questioned whether pi-stacking is actually based on the attraction between pi cloud electron density or not – which is another aspect why this term may be misleading (Martinez & Iversion, 2012). Yet this discussion goes beyond the scope of this thesis and it should be noted that this thesis largely follows the terms also used by PLIP and interactions between aromatic rings will hereinafter be named pi-stacking.

1.4.7 Pi-cation interactions

Pi-cation interactions or also called cation-pi interactions are non-covalent interactions between cations and the faces of pi systems. As described in 1.4.6 the face of a pi system forms a quadrupole moment with negative charge while the edges are positively charged. It therefore comes naturally that the cation is attracted to the face of the pi system and can electrostatically bind there. Pi-cation interactions are comparable in strength to salt bridges and in some cases even to hydrogen bonds (Anslyn & Dougherty, 2006).

1.4.8 Metal complexation

Metal complexation – or coordination complexation or just complexation – refers to a molecular structure where a central atom that is often a metal ion is bound to surrounding small molecules or ions (Hartshorn et al., 2015). Metal complexation primarily appears in proteins that have to bind metal ions to function (metalloproteins) (Andreini et al., 2006).

All these interaction types will be reappearing throughout this thesis and proper understanding of the underlying binding mechanisms can give additional insight where data or results may be ambiguous.

The end of this this chapter will be concluded by a short overview of the structure of the thesis and some general remarks.

1.5 Thesis overview

This thesis is divided into five major chapters:

- The **Introduction** gives an overview of the topic, current developments, state of the art, goals, motivation and essentials and will be concluded with this section.
- **Methods** will discuss the data, especially the 11 used targets, the Protein-Ligand Interaction Profiler which serves as a basis for follow up approaches, as well as the custom built workflows, scoring functions and the metrics that were used to evaluate them.
- **Results** summarizes the outcomes of the applied methods.
- The **Discussion** will mention noticeable aspects of the results as well as faced challenges and an outlook for the future.
- The thesis is finalized with a **Conclusion** that highlights the most important parts of the work.

Furthermore, it should be mentioned here that all the data, code and results are publicly available on GitHub via this repository: https://github.com/michabirklbauer/protein_docking

2. Methods

This chapter will cover the data and methods used in this thesis and especially give insight on the selected target proteins and which ligands were used for docking and subsequent scoring. The analysis of interactions as well as the scoring approaches will be discussed after establishing the data. Furthermore the chapter will be concluded with a description of the applied quality metrics to evaluate the performance of scoring.

2.1 Data

In total 11 targets were chosen for subsequent analysis: 11β-hydroxysteroid dehydrogenase type 1 (HSD11B1), acetylcholinesterase (ACHE), coagulation factor Xa (FXA), cyclooxygenase 1 (COX1) and cyclooxygenase 2 (COX2), dipeptidyl peptidase IV (DPP4), monoamine oxidase B (MAOB), p38 mitogen-activated protein kinase 14 (MAPK14), phosphodiesterase 5 (PDE5/PDE5A), protein-tyrosine phosphatase 1B (PTP1B) and soluble epoxide hydrolase (SEH). The selection of these targets was based on personal interest (research interest of the Institute of Pharmacy of the Paracelsus Medical University Salzburg) and availability of ligands in the Directory of Useful Decoys: Enhanced (DUD-E) (Mysinger et al., 2012). All in all 868 molecular structures have been manually selected from the PDB, downloaded and analysed. Inclusion criteria for these structures were:

- Belonging to a certain species (mostly Homo sapiens).
- Having a co-crystallized ligand.
- Not being mutated, chimeric or part of a fusion protein.

Further insights on the specific targets will be given in the respective subchapters.

2.1.1 Targets: 11β-hydroxysteroid dehydrogenase type 1

Basic information:

- EC number: 1.1.1.146
- Encoding gene name: HSD11B1
- Encoding gene location: 1q32 – q41
- Organism: Homo sapiens
- Number of residues: 292
- Molecular weight: 32400.665
- Cellular location: Endoplasmic reticulum membrane

Data taken from DrugBank (DrugBank - P28845, 2021; Wishart et al., 2018).

HSD11B1 is a microsomal enzyme belonging to the short-chain dehydrogenase/reductase family and catalyses the NADPH-dependent conversation of 11-ketosteroid cortisone to the glucocorticoid hormone cortisol in humans. Glucocorticoid hormones play essential roles in various physiological processes, among them lipid and bone metabolism, maturation and differentiation of cells as well as in inflammatory response and stress modulation. Therefore HSD11B1 is highly expressed in the respective glucocorticoid target tissues like the liver tissue,

adipose tissue and skeletal muscle tissue. Furthermore elevated levels of HSD11B1 dependent glucocorticoids have been associated with several different diseases, for example insulin and leptin resistance, visceral obesity, dyslipidemia, type 2 diabetes and cardiovascular complications. HSD11B1 is an attractive target for inhibition to manipulate glucorticoid levels and treat the corresponding diseases (Classen-Houben et al., 2009; Thomas & Potter, 2011).

The ligand binding site of HSD11B1 is a predominantly hydrophobic pocket that is open at both ends so that ligands that are too long to fit into the binding site can extend out of it. Contacts with the following residues are known from experimental structures: ILE121, THR122, ASN123, THR124, SER125, LEU126, SER170, LEU171, ALA172, VAL175, TYR177, PRO178, MET179, VAL180, TYR183, GLY216, LEU217, THR220, THR222, ALA223, ALA226, VAL227, VAL231 and MET233 (Thomas & Potter, 2011).

Cofactor(s): Human HSD11B1 is co-crystallized with NADP(H) in the cofactor binding site (Thomas & Potter, 2011). The cofactors are labelled with 3-letter codes NAP and NDP in the PDB respectively.

Analysed structures: The PDB was queried for Enzyme Classification Number = 1.1.1.146 AND Scientific Name of Source Organism = Homo sapiens. In total 28 structures were manually selected from the resulting hits for further analysis. The complete list of structures can be found in the GitHub repository in the respective data folder for HSD11B1 and in the appendix.

2.1.2 Targets: Acetylcholinesterase

Basic information:

- EC number: 3.1.1.7
- Encoding gene name: ACHE
- Encoding gene location: 7q22
- Organism: Homo sapiens
- Number of residues: 614
- Molecular weight: 67795.525
- Cellular location: Cell junction

Data taken from DrugBank (DrugBank - P22303, 2021).

The principle biological role of ACHE is the termination of impulse transmission at cholinergic synapses by hydrolysing the neurotransmitter acetylcholine into choline and acetate (Dvir et al., 2010; Tripathi & Srivastava, 2008). ACHE is critically important for the regulation of neurotransmissions at synapses in all areas of the nervous system and consequentially the inactivation of large amounts of ACHE leads to the death of any organism with a nervous system. Irreversible ACHE inhibitors have been utilized in the past as insecticides and in chemical warfare. On the other hand reversible inhibitors of ACHE such as donepezil, galantamine, rivastigmine and huperzine A have been used to treat neurodegenerative disorders

that exhibit defects in cholinergic neurotransmission such as Alzheimer's disease (Cheung et al., 2012; Tripathi & Srivastava, 2008).

The active site of ACHE consists of three major domains and one peripheral domain. Firstly, an esteratic locus containing the catalytic machinery of the enzyme, namely SER200, HIS440 and GLU327. Secondly, the anionic subsite that is ≥ 4.7 Å away from the esteratic SER and is defined by TRP84, PHE330 and PHE331. The anionic subsite is the binding location for the quaternary ammonium pole of acetylcholine and is responsible for the orientation of entering substrates by aligning the charged part (of the substrate). This is mainly carried out by TRP84. Thirdly, there is a hydrophobic region near the esteratic and anionic subsite that is important for binding. The fourth and final domain is called the peripheral anionic site and is > 20 Å away from the three major domains. It can bind cationic ligands such as gallamine, *d*-tubo-curarine and decamethonium and binding in this site frequently leads to a conformation change of the active center (Quinn, 1987; Tripathi & Srivastava, 2008).

Cofactor(s): None.

Analysed structures: The PDB query for ACHE was Enzyme Classification Number = 3.1.1.7 AND Scientific Name of Source Organism = Homo sapiens. In total 53 structures were selected out of the resulting hits for further analysis. For docking and scoring the PDB entry 4EY7 was used (Cheung at al., 2012). The complete list of used structures can be found in the respective folder for ACHE in the GitHub repository or in the appendix.

2.1.3 Targets: Coagulation factor Xa

Basic information (for coagulation factor X):

- EC number: 3.4.21.6
- Encoding gene name: F10
- Encoding gene location: 13q34
- Organism: Homo sapiens
- Number of residues: 488
- Molecular weight: 54731.255
- Cellular location: Secreted

Data taken from DrugBank (DrugBank - P00742, 2021).

Coagulation factor Xa denotes the activated form of coagulation factor X which is an important enzyme in the cascade of blood coagulation. Coagulation factor X is activated by coagulation factor VIIIa which is also the activated product of a chain of interactions with different other coagulation factors. FXA activates prothrombin to thrombin, which subsequently catalyses the conversion of fibrinogen to fibrin, which is the basis for all blood clots. Logically FXA has become a compelling target for treating coagulation disorders like pulmonary embolism or deep vein thrombosis. However, since FXA belongs to the trypsine-like serine protease family which is involved in numerous physiological functions in the body, the discovery and design of FXA

inhibitors pose a challenge. Inhibitors have to specifically and selectively bind to FXA to avoid toxicity and adverse side effects (Rai et al., 2001). Examples for FXA inhibitors are fondaparinux and otamixaban (Kohrt et al., 2007).

FXA has an active site catalytic triad compromised of amino acids SER195, HIS57 and ASP102. The binding site of FXA is divided into five regions S1, S1', S2, S3 and S4. Key residues are located in the S1 pocket, namely ASP189, ALA190 and GLN192 which likely influence inhibitor binding and selectivity (Rai et al., 2001).

Cofactor(s): None.

Analysed structures: The corresponding PDB query for FXA was Enzyme Classification Number = 3.4.21.6 AND Scientific Name of Source Organism = Homo sapiens. Out of the resulting hits 129 entries were considered for further analysis. The complete list of structures can be found in the FXA data folder in the GitHub repository or in the appendix section.

2.1.4 Targets: Cyclooxygenase 1 & Cyclooxygenase 2

Basic information:

	Cyclooxygenase 1	**Cyclooxygenase 2**
• EC number:	1.14.99.1	1.14.99.1
• Encoding gene name:	PTGS1	PTGS2
• Encoding gene location:	9q32-q33.3	1q25.2-q25.3
• Organism:	Homo sapiens	Homo sapiens
• Number of residues:	599	604
• Molecular weight:	68685.82	68995.625
• Cellular location:	Microsome membrane	Microsome membrane

Data taken from DrugBank (DrugBank - P23219, 2021; DrugBank - P35354, 2021).

The two cyclooxygenases (often also named prostaglandin H_2 synthases) are the two enzymes that catalyse the first two steps in the biosynthesis of prostaglandins from arachidonic acid in the human body. COX1 is constitutive, meaning it is present in nearly all cell types at a constant level, while COX2 activity is induced, meaning normally absent in cells but when induced by external stimuli the protein levels increase and decrease in a matter of hours. COX1 is involved in the production of prostaglandins for stomach and intestine to maintain the integrity of the mucosal epithelium as well as in the production of prostaglandins that preserve normal renal function in compromised kidneys. Inhibition of COX1 leads to gastric damage, haemorrhage and ulceration. On the other hand COX2 is induced by pro-inflammatory cytokines and growth factors and consequently the inductively produced prostaglandins are involved in both inflammation and control of cell growth. Additionally COX2 is also constitutively present in the brain and the spinal cord where it may be involved in nerve transmissions for pain and fever. Furthermore, prostaglandins synthesised by COX2 also have shown to be important in ovulation and the birth process. Because COX2 is inherently known for its role in inflammation,

COX1 and COX2 are sometimes also labelled as physiological and pathological, respectively. However, categorization into constitutive and induced is more encouraged. Both cyclooxygenase isoforms can be inhibited by aspirin and other nonsteroid anti-inflammatory drugs (NSAIDs). Aspirin inhibits the catalytic reaction by irreversibly binding to the active site of the enzymes while other NSAIDs such as ibuprofen and indomethacin compete with the substrate arachidonic acid for the binding site and inhibit it either reversibly or irreversibly. Despite both isoforms being able to be inhibited by NSAIDs, selective inhibition of COX2 is preferred to reduce inflammation without removing any protective prostaglandins in the stomach and kidney produced by COX1 (Vane, Bakhle & Botting, 1998).

Structurally both cyclooxygenase isoforms are very similar with a molecular weight at around 70 000 and a length of about 600 amino acids that share a 63% identical sequence. The 3D X-ray crystallographic structure of COX2 can be superimposed on that of COX1 revealing that the residues that form the substrate binding site, the catalytic region and the residues immediately adjacent are all identical except for two variations. To be specific, in COX1 the binding pocket consists of amino acid ILE at positions 434 and 523, while on the other hand COX2 shows amino acid VAL in those positions instead. This results not only in some biochemical differences – for example that COX2 accepts a wider range of fatty acids as substrates than COX1 – but also makes selective inhibition of COX2 possible (Vane, Bakhle & Botting, 1998). Several binding modes exist for ligands interacting with cyclooxygenases. The classic NSAIDs typically bind via ionic interactions to ARG120 and via hydrogen bonding to TYR355. Another mode would be binding to TYR385 and SER530 via hydrogen bonding, as exhibited by diclofenac. Both of these binding modes are observed when arachidonic acid binds to COX2 (Xu et al., 2014).

Cofactor(s): The single crystal structure for human COX1 (PDB code 6Y3C) does not contain a cofactor. The crystal structures for sheep (Ovis aries) COX1 are co-crystallized with the cofactor HEME (PDB 3-letter code HEM). Crystal structures for human COX2 contain protoporphyrin IX containing CO as a cofactor (PDB 3-letter code COH) and crystal structures for mouse (Mus musculus) COX2 are also co-crystallized with the cofactor HEME.

Analysed structures: For human COX1 and COX2 the PDB was queried for Enzyme Classification Number = 1.14.99.1 AND Scientific Name of Source Organism = Homo sapiens. This query results in one hit for COX1 and seven hits for COX2, however, the single COX1 structure does not contain a ligand and was therefore not further analysed. The remaining seven hits for COX2 were all kept for further research. Visibly more structures are available for sheep COX1 where the PDB was queried for Enzyme Classification Number = 1.14.99.1 AND Scientific Name of Source Organism = Ovis aries. From the available structures 25 were selected for further analysis. Similarly for COX2 a lot more structures exist for mouse COX2 – the PDB was queried for Enzyme Classification Number = 1.14.99.1 AND Scientific Name of Source Organism = Mus musculus and from the resulting hits 44 structures were subsequently used. Furthermore, docking and scoring was based on the PDB structure 4O1Z (Xu et al., 2014).

It should be noted here that docking and scoring results are only available for sheep COX1 but frequency analysis was carried out for both cyclooxygenase isoforms (with exception of human COX1 as there are no structures with ligands publicly available on the PDB). Complete lists of utilized structures can be found in the respective data folders for COX1 and COX2 in the GitHub repository or in the appendix section of this thesis.

2.1.5 Targets: Dipeptidyl peptidase IV

Basic information:

- EC number: 3.4.14.5
- Encoding gene name: DPP4
- Encoding gene location: 2q24.3
- Organism: Homo sapiens
- Number of residues: 766
- Molecular weight: 88277.935
- Cellular location: Secreted

Data taken from DrugBank (DrugBank - P27487, 2021).

DPP4 is a multifunctional cell surface protein and serine protease that is expressed in most cell types and is involved in the inactivation of glucagon-like peptide 1 (GLP-1) and glucose-dependent insulinotropic peptide (GIP), two insulin-sensing hormones, by cleaving the N-terminal dipeptides from these and other polypeptides with proline or alanine in the penultimate position (Havre et al., 2008; Chen, 2006). Furthermore, this ability allows DPP4 to also regulate the activity of numerous other cytokines and chemokines and DPP4 can therefore act as a tumor suppressor or activator and is involved in many different cancer types. Manipulation of DPP4 by specific cDNA-carrying plasmids, siRNA and monoclonal antibodies resulted in inhibition of cell growth, enhanced sensitivity to selected chemotherapeutic agents and enhanced survival rates in mouse xenograft models, proving the potential of these targeted therapies for specific cancers expressing DPP4 (Havre et al., 2008). On the other hand, because of its involvement with GLP-1 and GIP, the inhibition of DPP4 has been proposed as an effective approach for the treatment of type 2 diabetes and several structurally diverse DPP4 inhibitors have been established and approved therapeutically in the past. Among them sitagliptin, vildagliptin, saxagliptin, linagliptin and alogliption, to name some examples (Berger et al., 2017).

Structurally DPP4 is made up by a S1, S2, S1’ and S2’site. The S1 site is categorized as a hydrophobic pocket near a catalytic SER630 where – assuming an active substrate compound – the substrates P1 region binds. Secondly, the substrates P2 position is anchored by interactions with GLU205 and GLU206 in the S2 pocket of DPP4. The S2 pocket is also mostly hydrophobic and features residues ARG125, PHE357 and TYR547 of which specifically ARG125 forms a hydrogen bond with the substrates P1’ residue. The S1’ pocket is flat and not very well defined and the interactions with the substrates P1’ residues are mostly nonspecific Van der Waals interactions. Last but not least the S2’ pocket of DPP4 contains a TRP629

residue forming a hydrophobic wall that interacts with the lipophilic P2' region of the substrate. Most importantly however, the primary residues involved in substrate recognition and binding are located in the S2 pocket, namely the above mentioned GLU205, GLU206 and ARG125 (Berger et al., 2017).

Cofactor(s): None.

Analysed structures: The respective PDB query for DPP4 was Enzyme Classification Number = 3.4.14.5 AND Scientific Name of Source Organism = Homo sapiens. From the resulting hits 98 structures were eligible for further analysis and PDB entry 2G5T (Longenecker et al., 2006) was chosen for docking and scoring. Complete lists of all used structures are again available in the respective data folder for DPP4 in the GitHub repository and in the appendix section.

2.1.6 Targets: Monoamine oxidase B

Basic information:

- EC number: 1.4.3.4
- Encoding gene name: MAOB
- Encoding gene location: Xp11.23
- Organism: Homo sapiens
- Number of residues: 520
- Molecular weight: 58762.475
- Cellular location: Mitochondrion outer membrane

Data taken from DrugBank (DrugBank - P27338, 2021).

Monoamine oxidase A (MAOA) and MAOB are both mitochondrial outer membrane flavoenzymes involved in the pathways for controlling amine neurotransmitter levels in the cell by oxidation. Additionally to the oxidation of traditional amines such as dopamine and serotonin, MAOA and MAOB are also responsible for oxidation of ingested amines such as phenethylamine and tyramine to prevent their functioning as false neurotransmitters. Monoamine oxidase (MAO) inhibitors were originally discovered to be great antidepressants but side effects of covalently bound drugs that showed up during clinical application reduced the attractiveness of MAO as therapeutic target. However, MAOB has regained interest of the research and medical community after the observation of an age-related increase of MAOB levels in humans and a possible connection to neurodegenerative diseases such as Parkinson's disease. Henceforth the selective inhibition of MAOB with non-covalently binding agents has become of vital interest (Edmondson, Binda & Mattevi, 2007).

Human MAOB is crystallized as a dimer with two cavities important for substrate binding. Firstly, the so called "entrance cavity" that is very hydrophobic in nature and exhibits a volume of 290 $Å^3$. Secondly, separated from the entrance cavity by ILE199 the also hydrophobic "substrate cavity" is situated with a volume of 390 $Å^3$. The ILE between the two cavities serves as a gate and the substrate cavity can therefore exit in either an open or closed form – which

has been shown to be important for inhibitor specificity. Furthermore, at the end of the substrate cavity resides the flavin-adenine dinucleotide cofactor which is covalently bound to CYS397. Additionally the two nearly parallel residues TYR398 and TYR435 form what has been termed an "aromatic cage" which has catalytic significance by polarizing the amine moiety of the substrate to make it more nucleophile and by providing a path for guiding the substrate amine towards the reactive positions on the flavin ring (Edmondson, Binda & Mattevi, 2007).

Cofactor(s): Human MAOB is co-crystallized with flavin-adenine dinucleotide (PDB 3-letter code FAD).

Analysed structures: The according PDB query for human MAOB was Enzyme Classification Number = 1.4.3.4 AND Scientific Name of Source Organism = Homo sapiens and of the resulting hits 47 structures were further analysed. The PDB entry for docking and scoring was 2XCG (Bonivento et al., 2010). The complete list of utilized structures can be found in the data folder for MAOB in the GitHub repository and in the appendix section.

2.1.7 Targets: P38 mitogen-activated protein kinase 14

Basic information:

- EC number: 2.7.11.24
- Encoding gene name: MAPK14
- Encoding gene location: 6p21.3-p21.2
- Organism: Homo sapiens
- Number of residues: 360
- Molecular weight: 41292.885
- Cellular location: Cytoplasm

Data taken from DrugBank (DrugBank - Q16539, 2021).

MAPK14 (or p38α) is one of the four p38 mitogen-activated protein kinases (MAPK) in mammals together with MAPK11 (p38β), MAPK12 (p38γ) and MAPK13 (p38δ). MAPK14 is usually highly expressed in all cells while MAPK11 is expressed at lower levels and MAPK12 and MAPK13 have more restricted expression patterns (Segalés, Perdiguero & Muñoz-Cánoves, 2016). MAPKs are part of the MAPK signalling pathway where various extracellular stimuli – usually resulting from stress – are converted to activate specific cellular response mechanisms through the activation of the individual p38 proteins. Several environmental stressors have been identified to activate p38 responses, such as UV light, heat shock, osmotic stress, inflammatory cytokines like interleukin 1 and tumor necrosis factor alpha, as well as growth factor stimulation. Downstream products of the MAPK signalling pathway are several transcription factors and molecules of the translational machinery. Therefore p38 kinases are capable of regulating many diverse biological processes like cell growth and differentiation, cell cycle arrest, apoptosis, cardiomyocyte hypertrophy, inflammation, senescence and tumor progression. Two chemical mechanisms are known to regulate p38 MAPK activity, firstly, protein phosphorylation by certain dual kinases called mitogen-activated protein kinase kinases

(MKK), particularly MKK3 and MKK6. Secondly, the interaction of p38 with TAB1 (mitogen-activated protein kinase kinase kinase 7-interacting protein 1) which leads to autophosphorylation of the enzyme (Pillai et al., 2011). Inhibition of MAPKs has seen therapeutic application especially in the treatment of autoimmune disorders due to the involvement of p38 in inflammatory cell signalling (Goldstein & Gabriel, 2005).

Residues LYS53 and LYS152 have been identified as key amino acids for binding and regulating the activity of p38. Specifically LYS53 is important for ATP binding while LYS152 plays an essential role in substrate binding of p38 (Pillai et al., 2011).

Cofactor(s): None.

Analysed structures: The corresponding PDB query for MAPK14 was Enzyme Classification Number = 2.7.11.24 AND Gene Name = MAPK14 AND Scientific Name of Source Organism = Homo sapiens. In total 199 structures were selected for further research from the resulting hits. The full list of analysed structures can be found in the data folder for MAPK14 in the GitHub repository and in the appendix section.

2.1.8 Targets: Phosphodiesterase 5

Basic information:

- EC number: 3.1.4.35, 3.1.4.17 (PDB, UniProt)
- Encoding gene name: PDE5A
- Encoding gene location: 4q25-q27
- Organism: Homo sapiens
- Number of residues: 875
- Molecular weight: 99984.14
- Cellular location: Cytoplasm

Data taken from DrugBank (DrugBank - O76074, 2021).

Phosphodiester (PDE) enzymes play a key role in all cellular functions involving cyclic nucleotides as second messengers by hydrolysing the phosphodiester bonds of cyclic adenosine monophosphate (cAMP) and cyclic guanine monophosphate (cGMP). Among the 11 known PDE families PDE5 is the predominantly metabolizing cGMP PDE in cavernosal tissue and the penile arteries, however, it is also active in vascular smooth muscle cells, in platelets, and other tissues, such as the lung (Bischoff, 2004). Because cGMP controls the relaxation of vascular smooth muscles and therefore is able to allow increased blood flow, the inhibition of PDE-mediated degradation of cGMP was first considered for therapeutic use in systemic hypertension and angina. However, a first selective PDE5 inhibitor named sildenafil proved to be unsuccessful in cardiovascular disease trials. Instead patients reported increased erectile function which eventually led to a refocusing of the clinical program and ultimately the approval of sildenafil as a drug for treating erectile dysfunction (Ravipati et al., 2007).

The active site of PDE5 is approximately 15 Å deep and has on opening of about 20 Å times 10 Å. Generally it can be subdivided into three pockets, the metal binding pocket (M pocket) consisting of dimetal ions as well as polar and hydrophobic residues, a solvent-filled side pocket (S pocket) and a pocket containing a purine-selective glutamine and a hydrophobic clamp (Q pocket). In PDE5A specifically the purine-selective glutamine GLN817 in the Q pocket is of importance as it is involved in nucleotide recognition and is a key residue for the selective inhibition of PDE5 where inhibitors usually bind via hydrogen bonds (Card et al, 2004).

Cofactor(s): None.

Analysed structures: The respective PDB query to retrieve structures for PDE5 was Enzyme Classification Number = 3.1.4.35 AND Scientific Name of Source Organism = Homo sapiens of which 32 entries were selected for subsequent analysis. An exhaustive list of used structures can be found in the data folder for PDE5A in the GitHub repository and in the appendix section.

2.1.9 Targets: Protein-tyrosine phosphatase 1B

Basic information:

- EC number: 3.1.3.48
- Encoding gene name: PTPN1
- Encoding gene location: 20q13.1-q13.2
- Organism: Homo sapiens
- Number of residues: 435
- Molecular weight: 49966.44
- Cellular location: Endoplasmic reticulum membrane

Data taken from DrugBank (DrugBank - P18031, 2021).

Phosphorylation of proteins is an important process in the regulation of many cellular functions in eukaryotes. Specifically two different families of proteins are involved in this process, protein tyrosine kinases and protein tyrosine phosphatases. Protein tyrosine kinases catalyse the phosphorylation of phosphotyrosine residues in proteins while on the other hand protein tyrosine phosphatases catalyse the dephosphorylation of phosphotyrosine residues in proteins. When functioning properly, these two classes of enzymes provide dynamic control of cellular responses to external stimuli and regulation of cell internal mechanisms. PTP1B was the first protein tyrosine phosphatase that was cloned and fully characterized and today it is one of the best validated biological targets for non-insulin dependent diabetes and obesity. PTP1B catalyses the dephosphorylation of the insulin receptor as well as insulin receptor substrates involved in insulin signalling and therefore negatively regulates the actions of insulin. Furthermore, several research groups have found PTP1B to be also involved in cancer as experiments in mice showed that an overexpression of PTP1B is sufficient to drive tumorigenesis. Inhibition of PTP1B might therefore be a promising approach in cancer therapy (Combs, 2010).

PTP1B consists of 435 amino acids of which residues 30 – 278 correspond to the catalytic domain while the 35 C-terminal residues are responsible for guiding the protein to the cytosolic face of the endoplasmic reticulum where the catalytic reaction takes place. The recognition of the substrate binding sequence and binding of the phosphotyrosine are mediated by residues HIS214, CYS215, SER216, ALA217, GLY218, ILE219, GLY220 and ARG221. In detail, a TRP-PRO-ASP loop closes down on the substrate and positions the thiolate of CYS215 for nucleophilic attack upon the phosphotyrosine. The phosphate is then cleaved from the phosphotyrosine residue and the dephosphorylated substrate can diffuse from the active side and allows water to take its place. As a result PTP1B is left with the phosphorylated CYS215 which is hydrolysed by a catalytic reaction with ASP181 to regenerate the active form of the phosphatase and complete the catalytic cycle (Combs, 2010).

Cofactor(s): None.

Analyses structures: The according PDB query for PTP1B was Enzyme Classification Number = 3.1.3.48 AND Gene Name = PTP1B AND Scientific Name of Source Organism = Homo sapiens. For subsequent analysis 102 structures were selected of which all can be found in the data directory for PTP1B in the GitHub directory or in the appendix section (as PDB codes).

2.1.10 Targets: Soluble epoxide hydrolase

Basic information:

- EC number: 3.3.2.10
- Encoding gene name: EPHX2
- Encoding gene location: 8p21-p12
- Organism: Homo sapiens
- Number of residues: 555
- Molecular weight: 62615.22
- Cellular location: Cytoplasm

Data taken from DrugBank (DrugBank - P34913, 2021).

SEH has two distinct enzyme activities, namely it functions as an epoxide hydrolase and as a phosphatase. Structurally the SEH protein is a homodimer and each monomer features two separate domains responsible for one of the two enzymatic activities. The C-terminal exerts epoxide hydrolase activity and the N-terminal phosphatase activity. Moreover, the N-terminal hydrolyses phosphate esters in a magnesium-dependent reaction while the C-terminal is responsible for the biological roles associated with SEH, namely the metabolism of arachidonic acid epoxides that play an important part in blood pressure, cell growth, inflammation and pain. Pharmacological inhibition of the C-terminal active site has seen use in anti-inflammatory, anti-hypertensive, neuroprotective and cardioprotective drugs (Morisseau et al., 2013).

Several residues have been identified to be of importance in the binding process of ligands with SEH, specifically the residues ASP335 and TYR383 in the active site as well as residues TRP336, LEU499 and HIS524 from hydrophobic pockets. Furthermore, it has been shown that TYR383, TYR466 and ASP335 form hydrogen bonds that are important for inhibitor binding (Karami et al., 2016).

Cofactor(s): None.

Analysed structures: The corresponding PDB query for SEH was Enzyme Classification Number = 3.3.2.10 AND Scientific Name of Source Organism = Homo sapiens. Of the resulting hits 104 structures were included in the analysis. PDB entry 6HGV (Kramer et al., 2018) was used for docking and scoring. Again a complete list of utilized structures can be found in the data directory of SEH in the GitHub repository as well as in the appendix section.

2.1.11 Ligands: DUD-E

Additionally to the ligands that were co-crystallized in the PDB entries which were used to get an overview of interactions present in the respective target, molecules from the Directory of Useful Decoys: Enhanced (DUD-E) were used for docking and interaction-frequency-based scoring. The DUD-E is a benchmark dataset based on its predecessor DUD, the Directory of Useful Decoys (Huang, Shoichet & Irwin, 2006). The DUD-E features 22 886 active compounds for 102 targets, an average of 224 ligands per target. Furthermore it contains 50 decoys for each active compound where each decoy has similar physico-chemical properties but a dissimilar 2D topology to its corresponding active compound (Mysinger et al., 2012).

DUD-E actives and decoys were used for four of the five targets that were evaluated with interaction-frequency-based scoring. The following list contains the names of the targets as well as the name of the respective DUD-E directory in parentheses.

- Acetylcholinesterase (ACES)
- Cyclooxygenase 1 (PGH1)
- Dipeptidyl peptidase IV (DPP4)
- Monoamine oxidase B (AOFB)

2.1.12 Ligands: SEH active and inactive compounds

The fifth target for interaction-frequency-based scoring was soluble epoxide hydrolase. Since SEH is not one of the 102 targets included in the DUD-E, a separate dataset of active and inactive molecules was used. Specifically an internal dataset of the Institute of Pharmacy of the Paracelsus Medical University Salzburg – which was previously established for the discovery of potent SEH inhibitors by pharmacophore-based virtual screening (Waltenberger et al., 2016) – was used. Although this dataset is not publicly available, an SDF file containing the docked ligands can be found in the scoring directory of SEH in the GitHub repository.

2.1.13 Data partitioning

For each of the five docked and scored targets the data was split into distinct training, validation and test partitions by random sampling. The training partition was used to calculate the optimal

cut-off value for discrimination of active and inactive ligands, the validation partition was used for optimization of hyperparameters and the purpose of the test partition was to have an unbiased estimate of the scoring performance (Xu & Goodacre, 2018). Particularly the training dataset contained 64% of the ligands, the validation dataset 16% and the test dataset 20%.

The structures from the Protein Data Bank that were considered for interaction frequency analysis have also been randomly assigned to a training and test partition using a 80%–20% split respectively. This was done because scoring of these structures was originally also considered, however it was finally not carried out because mining the experimental binding affinities from the web would have been a time consuming and error prone process. Therefore the structures that were assigned to the test partitions were never analysed as part of this thesis.

2.2 Protein-Ligand Interaction Profiler

To identify the interactions occurring in a specific protein-ligand complex the Protein-Ligand Interaction Profiler (PLIP version 2.1.8, PharmAI GmbH, https://plip.biotec.tu-dresden.de) was applied. PLIP is available as a web service, command-line tool and as a python package which enables high-throughput computation and the integration into existing workflows. The expected input is a protein-ligand complex in PDB format (file ending = ".pdb") which can either be from the Protein Data Bank itself or from docking or molecular dynamics software, for example. Subsequently the output is a list of detected interactions on single atom level for each binding site with a small molecule. Furthermore PLIP also offers 2D and 3D interaction diagrams. PLIP is able to identify eight different interaction types, namely hydrogen bonds, hydrophobic contacts, pi-stacking, pi-cation interaction, salt bridges, water bridges, halogen bonds and metal complexation. To characterize these interactions a rule/knowledge-based approach is applied which is founded in literature, mostly large-scale studies of analyses of high-quality protein structures. However, to also account for low-quality structures and structural errors some thresholds are modified to be more permissive. PLIP was validated on a set of 30 diverse literature-validated protein-ligand complexes (Salentin et al., 2015).

2.2.1 PLIP algorithm

The PLIP algorithm responsible for detecting and reporting relevant interactions can be categorized into four steps, which are structure preparation, functional characterization, rule-based matching and filtering of interactions. Firstly, in the preparation step, the input structure is hydrogenated and ligands are extracted along with their binding sites. Secondly, in the functional characterization step, functional groups, atoms and molecules are detected in the following procedure:

- **Detection of binding site atoms:** A binding site distance cut-off value is defined by adding 8.5 Å to the maximum extent of the ligand (which is the maximum distance of a ligand atom to ligand centroid). If a protein atom is within this distance cut-off value to any binding site atom it is characterized as belonging to the binding site.
- **Detection of hydrophobic atoms:** An atom is labelled as hydrophobic if it is a carbon atom and only has carbon or hydrogen atoms as neighbours.

- **Detection of aromatic rings:** The software Open Babel (O'Boyle et al., 2011) is used to identify rings and their aromaticity. If Open Babel does not report any aromaticity the ring is checked for planarity by calculating the normals of each atom to its neighbours in the ring. If the angles between each of pair of normals are less than 7.5° the ring is also considered to be aromatic.
- **Detection of hydrogen bond donors and acceptors:** This task is also carried out by Open Babel. Furthermore, halogen atoms are excluded as hydrogen bond acceptors.
- **Detection of charged groups:** For proteins the positive charges are assigned to the side chain nitrogen atoms of ARG, HIS and LYS while negative charges are attributed to the carboxyl groups in ASP and GLU. For ligands the positive charges are assigned to quaterny ammonium groups, tertiary amines with the assumption that the nitrogen could pick up a hydrogen and thus get charged, sulfonium and guanidine groups while negative charges are defined for phosphate, sulfonate, sulfonic acid and carboxylate. The detection of charged groups is only exhaustive for the binding site and not the ligand.
- **Detection of halogen bond donors and acceptors:** The assumption is made that halogen atoms are not present in proteins and therefore halogen bond donors are searched for only in ligands. An atom qualifies as a halogen bond donor if it is a fluorine, chlorine, bromide or iodine atom connected to a carbon atom. On the other hand an atom is considered as a halogen bond acceptor in a protein if it is a proximal oxygen, nitrogen or sulphur atom connected to a carbon, nitrogen, phosphor or sulphur atom.
- **Detection of water:** Water molecules are considered if the respective oxygen atoms are within 8.5 Å to the maximum extent of the ligand.

In the third step of the PLIP algorithm rule-based matching is applied to detect the interactions between the protein and the ligand. The rules are mostly checking for geometric constraints like distance or angle between atoms. The approach is described in more detail below:

- **Detection of hydrophobic interactions:** Hydrophobic interactions are reported between all pairs of hydrophobic atoms within a distance of 4.0 Å.
- **Detection of hydrogen bonds:** A hydrogen bond between a hydrogen bond donor and a hydrogen bond acceptor is reported if the distance between donor and acceptor is less than 4.1 Å and the angle at the donor group X–H is above 100°.
- **Detection of aromatic stacking:** A pi-stacking interaction is given whenever the centres of the aromatic rings are within a distance of 7.5 Å and the angle deviates no more than 30° from the optimal angle. Moreover, the centre of each aromatic ring is projected onto the opposing ring's plane and the distance between centre and projected point has to be less than 2.0 Å.
- **Detection of pi-cation interactions:** A pi-cation interaction is present if there exists a positively charged entity and an aromatic ring where the charge centre and the aromatic

ring centre is less than 6.0 Å away. If the pi-cation interaction is putative with a tertiary amine an additional angle criterion is applied.

- **Detection of salt bridges:** A salt bridge is reported whenever two centres of opposite charge are located within a distance of 5.5 Å.
- **Detection of water bridges:** Even though residues can be bridged by more than one water molecule, PLIP only considers the case of one water molecule bridging ligand and protein via hydrogen bonding. A water bridge is reported in this case if two conditions are fulfilled. The first condition is that the water molecule is positioned between hydrogen bond donor and hydrogen bond acceptor pairs of ligand and protein with distances of the water oxygens within 2.5 Å and 4.0 Å to the corresponding polar atoms of the donor or acceptor groups. The second condition is that the angle between the acceptor atom, the water oxygen and the donor hydrogen is between 75° and 140° and the angle between the water oxygen, the donor hydrogen and the donor atom is larger than 100°.
- **Detection of halogen bonds:** A halogen bond is detected if a halogen bond acceptor and halogen bond donor is within 4.0 Å, the angle of the donor group deviates no more than 30° from 165° and the angle of the acceptor group deviates no more than 30° from 120°.

The final step of the PLIP algorithm is the so called filtering or reduction step where redundant and overlapping interactions are eliminated. The process of filtering is dependent on the interaction type and is described as follows:

- **Filtering of hydrophobic interactions:** Hydrophobic contacts between rings interacting via pi-stacking are removed because pi-stacking already involves hydrophobic interactions. Additionally if a ligand atom forms hydrophobic interactions with several binding site atoms in the same residue, only the interaction with the closest distance is kept. Vice versa if a protein atom forms hydrophobic interactions with several neighbouring ligand atoms, again only that interaction is kept that exerts the shortest distance.
- **Filtering of hydrogen bonds:** Hydrogen bonds are removed if one of the atoms already belongs to a group that forms a salt bridge. Furthermore, since a hydrogen bond donor can only take part in one hydrogen bond, only that hydrogen bond where the donor angle is closest to 180° is kept.
- **Filtering of water bridges:** A water molecule is only allowed to participate as a hydrogen bond donor in two hydrogen bonds and in any case where there are more than two hydrogen bonds possible, only the two interactions with a water angle closest to 110° are kept.

The output of the PLIP algorithm is a set of residues and the specific interactions they are forming for every binding site and small molecule. Additionally the PLIP web service offers visual results in JSMol that can be download in PNG format or as PyMOL session files (Salentin

et al., 2015). The residue and interaction data has been used as a basis for the research described in the upcoming sections.

2.3 PIA: Protein Interaction Analyzer

One key component of this research was the development of PIA (Protein Interaction Analyzer). PIA is a python package and collection of scripts and workflows to extract interaction frequencies of protein-ligand complexes from PDB and SDF files, to compare the interaction frequencies of active and inactive molecules, and ultimately to score protein-ligand complexes and predict if they are active or not. PIA is completely written in python and builds upon PLIP for the extraction of interactions, BioPandas (Raschka, 2017) for PDB structure manipulation, and RDKit (RDKit, 2021) to handle and merge molecules from SDF and PDB files. The complete source code as well as a configuration file to setup an Anaconda environment containing all requirements are available in the GitHub repository. Additionally a Docker image can be pulled from DockerHub via michabirklbauer/protein_docking.

The particular functions of PIA are described in more detail in the corresponding subsections below.

2.3.1 Extracting interaction frequencies

There are two possible input modes for analysing interaction frequencies, one has to either supply a list of PDB files of one target e.g. if one has downloaded structures from the Protein Data Bank and wants to analyse them, or supply a SDF file containing ligand coordinates and a PDB file that will serve as the host structure e.g. if one wants to analyse docking results (that are written to SDF format). In the first case the PDB structures will be directly analysed by PLIP. In the second case PIA will first remove any small molecules from the host structure and then write every ligand into a separate instance of the cleaned host PDB file. Following from that, if a SDF file contains *N* ligands it will result in *N* created PDB files. Each of these PDB files will then be supplied to PLIP for detection of interactions.

The result of the analysis by PLIP is a set of interactions for every small molecule in every protein-ligand structure that was supplied. To calculate the frequency for every interaction several aspects were considered:

- **Dealing with artefacts, suspicious ligands and other unwanted co-crystallized small molecules:** PLIP returns all interactions found in a protein-ligand complex, which includes interactions of possibly unwanted small molecules that were co-crystallized with the ligand. To filter out these interactions the BioLiP list of suspicious ligands is applied – which is also available in PLIP (Yang, Roy & Zhang, 2013; Salentin et al., 2015).
- **Dealing with cofactors:** Many proteins depend on and are co-crystallized with a cofactor, however, the interactions between cofactor and protein are unwanted when looking at interaction frequencies due to the fact that they are present in (almost) all structures. Cofactor-protein interactions would supersede interactions happening

between the protein and the ligand which are important for protein-ligand binding and therefore would make the results more ambiguous. For this reason a list of known cofactors has been compiled using the CoFactor database (EMBL-EBI, 2011) and cofactors/cofactor interactions are excluded from further analysis. Furthermore, this list is user extensible to enable users the exclusion of cofactors that are not mentioned in this list. It should also be noted here that interactions between cofactor and ligand would be of interest, however PLIP is not able to detect them and therefore no further research was done in that regard.

- **Dealing with hydrophobic interactions:** PLIP often reports multiple hydrophobic interactions for the same residue. Counting all these interactions would inherently lead to very high hydrophobic interaction frequencies and displace non-hydrophobic interaction frequencies. This behaviour is unwanted because hydrophobic interactions are considerably weaker and less impactful to protein-ligand binding than other interaction types. As a result only one hydrophobic interaction per residue was kept.
- **Dealing with multiple docking poses:** Many docking programs will not return a single pose but multiple docking poses per ligand. For interaction frequency analysis only the "best" pose was considered for every ligand where "best" was denoted as that pose that showed the most protein-ligand interactions.

After filtering out all the unwanted interactions based on the above criteria, a set of unique interactions was created from all interactions of all ligands. Each interaction was denoted by its interaction type, its residue number and the corresponding residue chain. In the experiments of this thesis only binding sites in chain A were considered. The absolute interaction frequencies are then calculated by counting for each interaction in how many structures it is present. The last step consists of calculating the relative frequencies by normalizing with the total number of analysed structures. The final result and output is a list of interactions with the corresponding frequencies. A summarised overview of the workflow can be seen in **Fig. 1**.

Extraction of interaction frequencies has been carried out for all 11 targets using structures from the Protein Data Bank as described in the subsections of the specific targets.

2.3.2 Comparing interaction frequencies between active and inactive molecules

To compare active molecules with inactive ones the workflow is extended by creating two sets of interactions, one for all active molecules and one for all inactive molecules. Naturally the input has to be complemented by the according structure information, this can either be in the form of having two separate SDF files for active and inactive compounds, labelled ligand names (PIA recognizes names containing "inactive" or "decoy" as inactive), or available IC50 values in the SDF file – for the later also a condition of what is considered active (or inactive) has to be given. For comparison of interactions the union of the two sets is taken and a list of all interactions in the union with corresponding active and inactive frequencies as well as the differences between the two (sorted by decreasing difference) is returned. For convenience a

plotting function is also implemented which shows the frequencies of active and inactive compounds in a grouped bar chart.

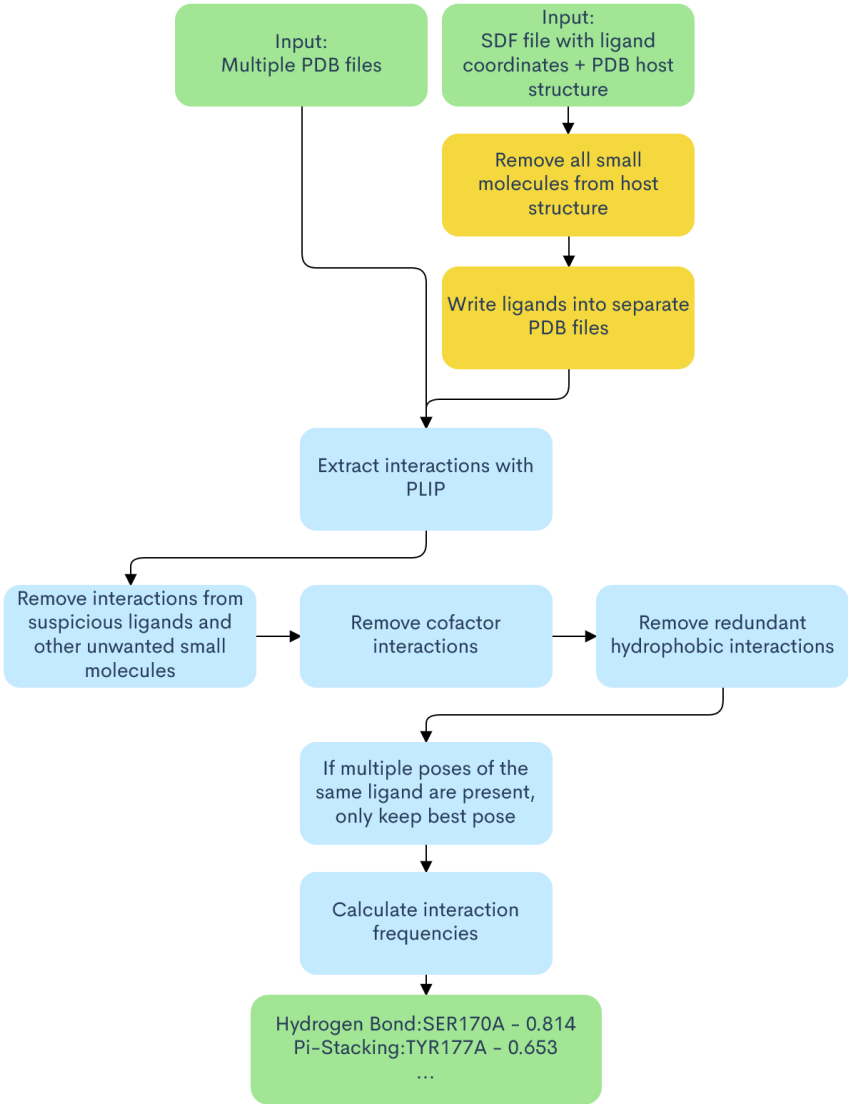


Fig. 1: Overview of the workflow for extracting interaction frequencies.

2.3.3 Scoring

The scoring workflow further extends the approach taken in the comparison procedure. Although the input is the same, PIA will first split the data into a training, validation and test partition – as described in 2.1.13 – before doing any data manipulation or analysis. Interactions are extracted from the training partition while the validation and test partition are only checked for interactions that appear in the training partition. Therefore an interaction that appears in the validation and/or test partition but not in the training partition will not be picked up by PIA and has no influence on the scoring. Furthermore, after the extraction of interactions a comparison between the interactions in active molecules and the interactions in inactive molecules (in the training partition) is made. On the basis of the comparison data a subset of interactions is selected for scoring. An interaction is part of this subset if and only if:

- The difference in interaction frequencies between active and inactive molecules is greater or equal to *D*.
- The interaction frequency in active molecules is greater or equal to *A*.
- The interaction frequency in inactive molecules is greater or equal to *I*.

Parameters *D*, *A* and *I* are determined by a grid search that looks for the optimal values in terms of maximizing the accuracy of the scoring on the validation partition. This procedure also returns the optimal scoring strategy (more on that at the end of this section).

This subset of interactions is then further divided into a subset *P* that contains all interactions that have a positive impact on the score, and a subset *N* that contains all interactions that negatively impact the score. Specifically that means if an interaction is more frequent in active molecules it is assigned to *P*, if it is more frequent in inactive molecules it is assigned to *N*.

Based on these subsets *P* and *N*, four different scoring strategies have been established. Consider a ligand with interactions $i_1, i_2, \ldots, i_n$ where each interaction belongs to either *P*, *N* or neither – in which case the interaction is discarded. Let *p* be the set of ligand interactions that belongs to *P* with interactions $p_1, p_2, \ldots, p_x$ and vice versa let *n* be the set of ligand interactions that belongs to *N* with interactions $n_1, n_2, \ldots, n_y$. Moreover the absolute frequency of an interaction p_i in the ligand shall be denoted as $f(p_i)$ – or in the negative case of an interaction n_i as $f(n_i)$. In most cases the frequency of an interaction in a ligand is one, nevertheless it can be greater, for example if a residue forms hydrogen bonds with several ligand atoms. Using the described notation, the first scoring strategy, herein after named "Strategy 1" or "Strategy +", can be denoted as follows:

$$S_1 = \sum_{i=1}^{x} 1$$

S_1 is the score of the first strategy and is the defined as the number of elements in *p*. Strategy 1 therefore does not account for negative interactions and interactions that happen multiple times in a single protein-ligand complex are counted only once.

The second scoring strategy named "Strategy 2" or "Strategy +-" is defined as the following:

$$S_2 = \sum_{i=1}^{x} 1 - \sum_{j=1}^{y} 1$$

S_2 is the score of the second strategy and is defined as the difference between the number of elements in *p* and the number of elements in *n*. Strategy 2 is an extension of Strategy 1 that also takes into account negative interactions but still counts multiple occurrences of an interaction only once.

The third strategy is referred to as “Strategy 3” or “Strategy ++” and can be denoted as:

$$S_3 = \sum_{i=1}^{x} f(p_i)$$

S_3 is the score of the third strategy and is defined as the sum of all absolute frequencies in p. Strategy 3 is an extension of Strategy 1 that also considers multiple occurrences of an interaction.

Last but not least the fourth strategy is called “Strategy 4” or “Strategy ++--“ and can be described as:

$$S_4 = \sum_{i=1}^{x} f(p_i) - \sum_{j=1}^{y} f(n_j)$$

S_4 is the score of the fourth strategy and is defined as the sum of all absolute frequencies in p minus the sum of all absolute frequencies in n. Strategy 4 is an extension of Strategy 2 that also considers multiple occurrences of an interaction. Strategy 4 is the only strategy that utilizes all information available in p and n.

The optimal scoring strategy (maximization of accuracy on the validation set) can be determined via grid search which is also applied for the determination of D, A and I. Nevertheless, in the standard workflow each of the scoring functions is applied to every ligand in the training dataset and an optimal cut-off value for discriminating between active and inactive complexes is determined for all four approaches. Optimal here refers to optimal for the maximization of the prediction accuracy on the training dataset. Additionally, the performance of each strategy has also been evaluated on the validation and test partition of the data using the performance metrics described in the following section 2.4. A summarised overview of the scoring workflow can be seen in **Fig. 2**.

In total five targets have been scored using the described approach, namely:

- Acetylcholinesterase (ACHE)
- Cyclooxygenase 1 (COX1)
- Dipeptidyl peptidase IV (DPP4)
- Monoamine oxidase B (MAOB)
- Soluble epoxide hydrolase (SEH)

In all cases the ligands were first docked using the software GOLD (Jones et al., 1997) applying standard docking workflows for ACHE, COX1, DPP4 and MAOB and a specialized workflow that was known from previous experiments for SEH. The resulting SDF files containing 10 poses for every ligand were then processed and scored with PIA.

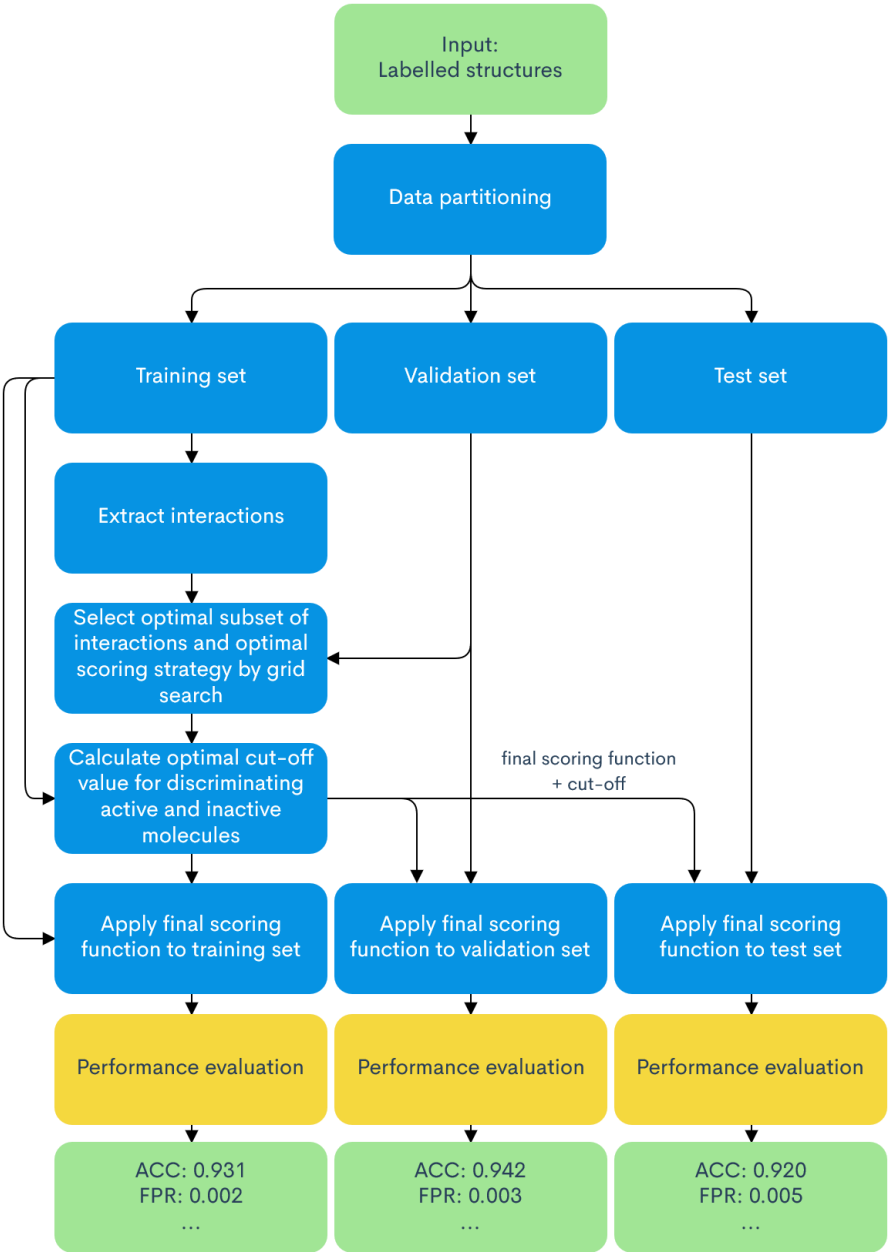


Fig. 2: Overview of the scoring workflow when applying a single scoring strategy. When evaluating more than one scoring strategy the bottom three layers of the workflow are repeated for every additional scoring function.

2.4 Performance metrics

The predictive power of the scoring workflows has been measured in terms of six metrics, namely the prediction accuracy (ACC), the false positive rate (FPR), the area under the receiver operating characteristic curve (AUC), the yield of actives (Ya), the enrichment factor (EF) and the relative enrichment factor (REF). PIA additionally returns a confusion matrix and the receiver operating characteristics (ROC) curve for visual inspection.

Let *TP* be the number of true positives, *TN* the number of true negatives, *FP* the number of false positives and *FN* the number of false negatives, then the metrics are defined as follows:

The **accuracy *ACC*** is the fraction of samples that is correctly predicted and is defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

(Lopes et al., 2017)

The **false positive rate *FPR*** is the number of samples wrongly predicted as active in relation to the number of all inactive molecules in the dataset and is defined as:

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

(Lopes et al., 2017)

The **yield of actives *Ya*** is the fraction of true actives among all predicted actives and is defined as:

$$Ya = \frac{TP}{TP + FP}$$

(Güner, 2000)

The **enrichment factor *EF*** is the proportion of how much more frequent true actives are in the set of predicted actives compared to the complete dataset. The enrichment factor can be any positive real number and is defined as:

$$EF = \frac{\frac{TP}{TP + FP}}{\frac{TP + FN}{TP + TN + FP + FN}}$$

(Lopes et al., 2017)

The **relative enrichment factor *REF*** denotes the percentage that the *EF* takes up of the maximum achievable *EF*. In other words, the relative enrichment factor is the *EF* normalised by the maximum *EF*. The relative enrichment factor is defined as:

$$REF = \frac{100 * TP}{\min(TP + FP,\ TP + FN)}$$

(Lopes et al., 2017)

The **area under the receiver operating characteristic curve *AUC*** represents the overall accuracy of a scoring workflow with a value close to 1.0 indicating high sensitivity and high specificity. The ROC curve is defined by a series of points, each point representing the predictive power of a specific cut-off value. The x-coordinate of the point denotes the false positive rate and the y-coordinate denotes the true positive rate of that cut-off value (Lopes et al., 2017). PIA calculates the AUC with the trapezoidal rule using the python package scikit-learn (scikit-learn version 0.24.2, scikit-learn: Machine Learning in Python, https://scikit-learn.org/stable/index.html).

3. Results

In the following the results are described separately for every target, applying the same order used in 2. Methods. High resolution plots and tables as well as all the results presented hereinafter are available in the GitHub repository in the respective data directory of the target.

3.1 11β-hydroxysteroid dehydrogenase type 1

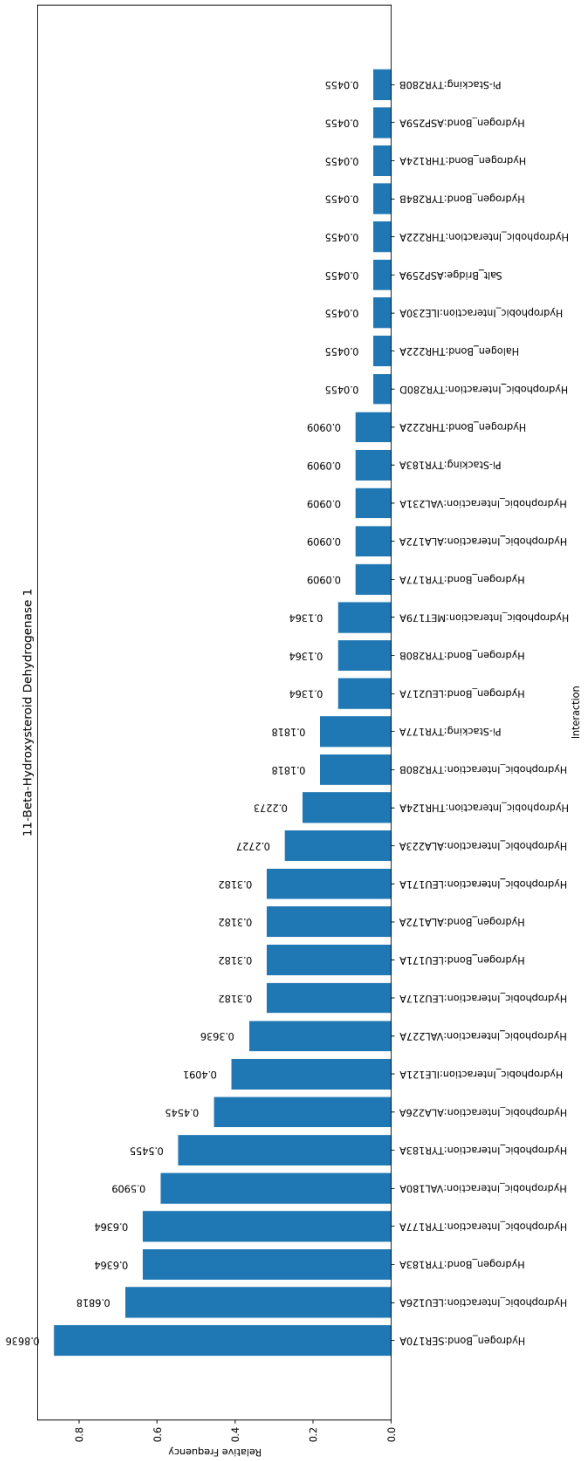


Fig. 3: Interaction frequencies of selected HSD11B1 structures from the PDB.

Fig. 3 shows interactions and their relative frequencies of the selected structures from the PDB. In total 22 structures were used for this analysis and the result shows that interactions known from literature are among the list that is returned by PIA. In more detail and to rehearse the known interacting residues mentioned in 2.1.1, the following interactions shall be highlighted:

- ILE121A: A hydrophobic interaction is present in 41% of the structures.
- THR124A: The threonine residue shows a hydrophobic interaction in 23% of the structures and forms a hydrogen bond in 5% of the structures.
- LEU126A: Hydrophobic interaction that is present in 68% of the structures and is therefore the second most frequent interaction.
- SER170A: A hydrogen bond is formed with this residue in 86% of the structures. It represents the most frequent interaction.
- LEU171A: This leucine residue forms two different interactions, namely a hydrogen bond in 32% of the structures and hydrophobic interactions also in 32% of the structures.
- ALA172A: Alanine at position 172 also interacts in two ways, it forms a hydrogen bond in 32% of the structures and hydrophobic interactions in 9% of the structures.
- TYR177A: This residue interacts with ligands in three different ways, namely by hydrophobic interaction in 64% of the cases, by pi-stacking in 18% of the cases, and by hydrogen bonding in 9% of the cases.

- MET179A: A hydrophobic interaction occurs in 14% of the structures.
- VAL180A: A hydrophobic interaction is present in 59% of the structures.
- TYR183A: Tyrosine 183 interacts with ligands via hydrogen bonding in 64% of the structures, via hydrophobic interactions in 55% of the structures, and via pi-stacking in 9% of the structures.
- LEU217A: Two interactions were detected for this leucine residue, hydrophobic interactions in 32% of the structures and hydrogen bonds in 14% of the structures.
- THR222A: Threonine at position 222 forms a hydrogen bond in 9% of the cases, a halogen bond in 5% of the cases, and hydrophobic interactions also in 5% of the cases.
- ALA223A: A hydrophobic interactions is present in 27% of the cases.
- ALA226A: This alanine shows hydrophobic interactions with ligands in 45% of the structures.
- VAL227A: A hydrophobic interaction is detected in 36% of the structures.
- VAL231A: In 9% of the structures this valine exhibits hydrophobic interactions.

Interacting residues that were mentioned in literature but were either not present or not detected in the analysed structures were THR122, ASN123, SER125, VAL175, PRO178, GLY216, THR220 and MET233.

3.2 Acetylcholinesterase

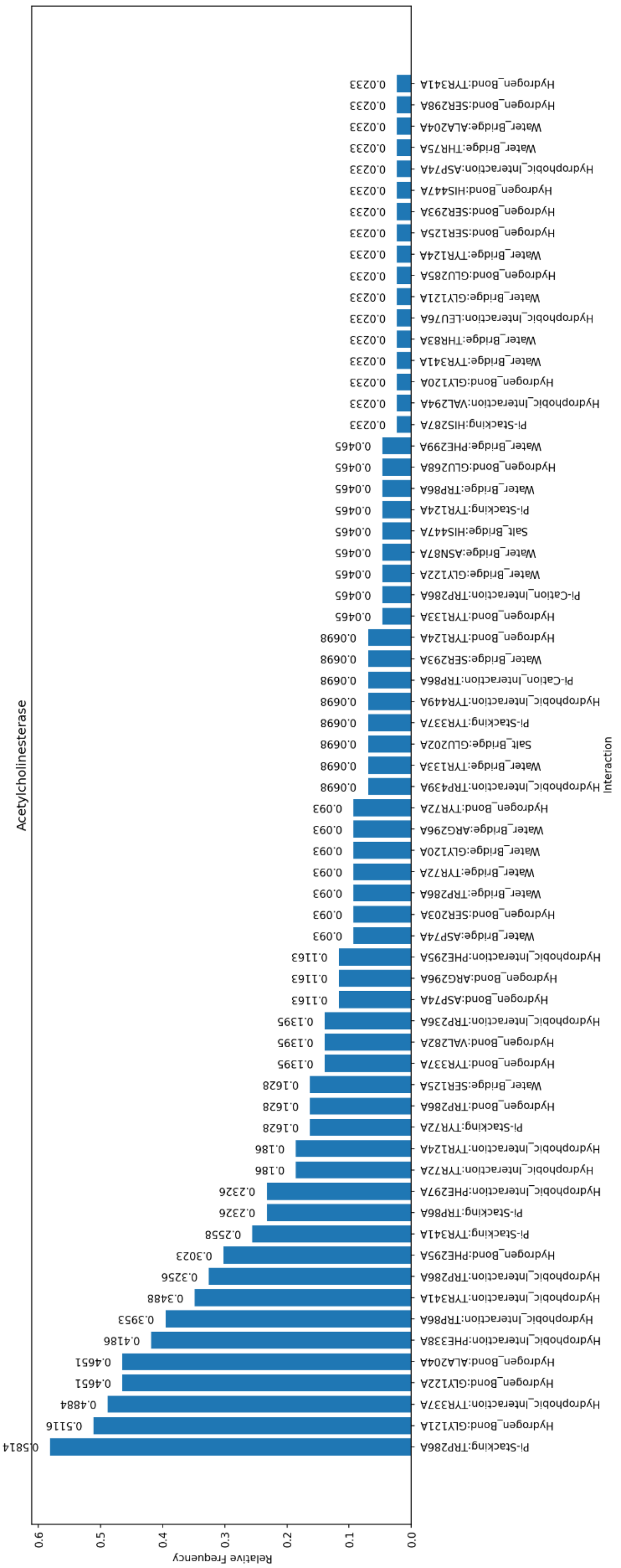
ACHE was the first of five targets that has been scored additionally to the analysis of interaction frequencies in the available PDB structures. The latter is described first.

3.2.1 Interaction frequencies

Interactions and their frequencies were extracted for 43 structures from the PDB. A graphical representation of all detected interactions and their corresponding frequencies can be seen in **Fig. 4**.

The resulting list of interactions included none of the residues known from literature. The top 5 interactions were:

- Pi-stacking with TRP286A in 58% of the structures.
- Hydrogen bonding with GLY121A in 51% of the structures.
- Hydrophobic interactions with TYR337A in 49% of the structures.
- Hydrogen bonding with GLY122A in 47% of the structures.
- Hydrogen bonding with ALA204A in 47% of the structures



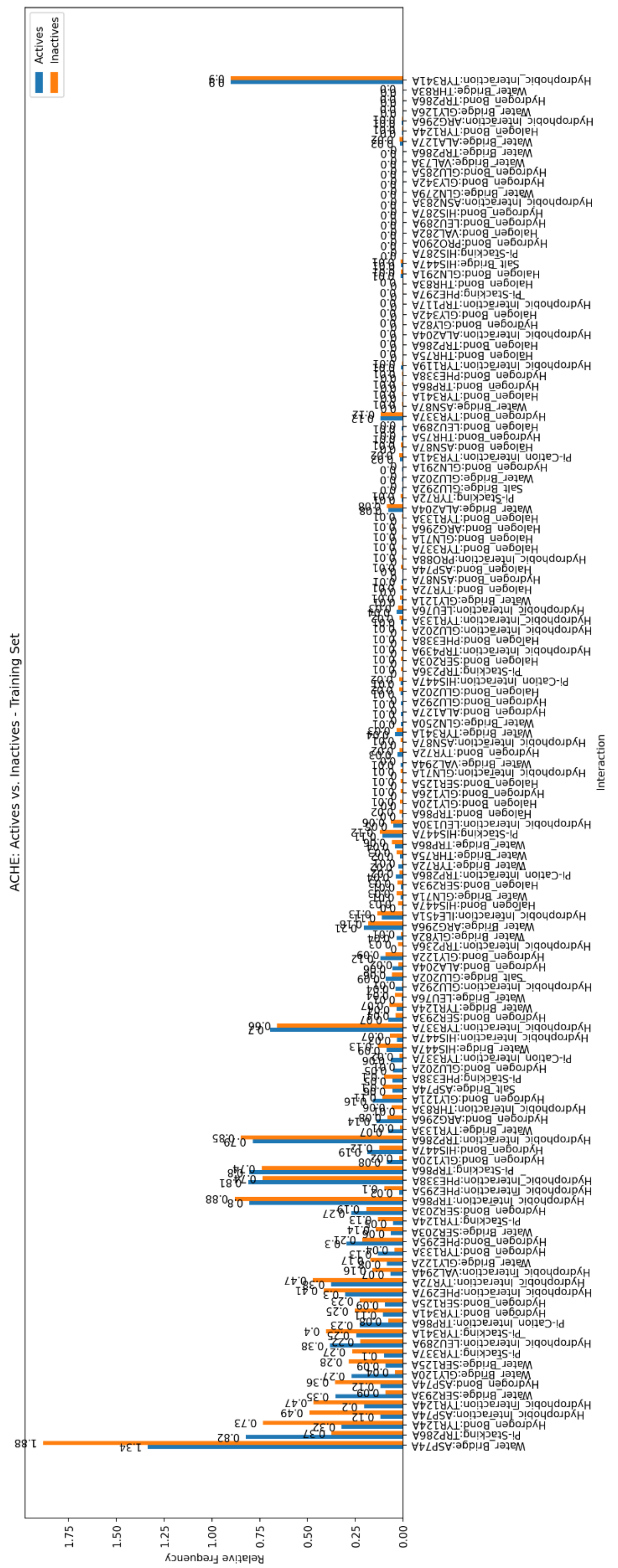


Fig. 4: Interaction frequencies of selected ACHE structures from the PDB.

Fig. 5: Distribution of interaction frequencies of active and inactive ACHE ligands in the training partition in comparison.

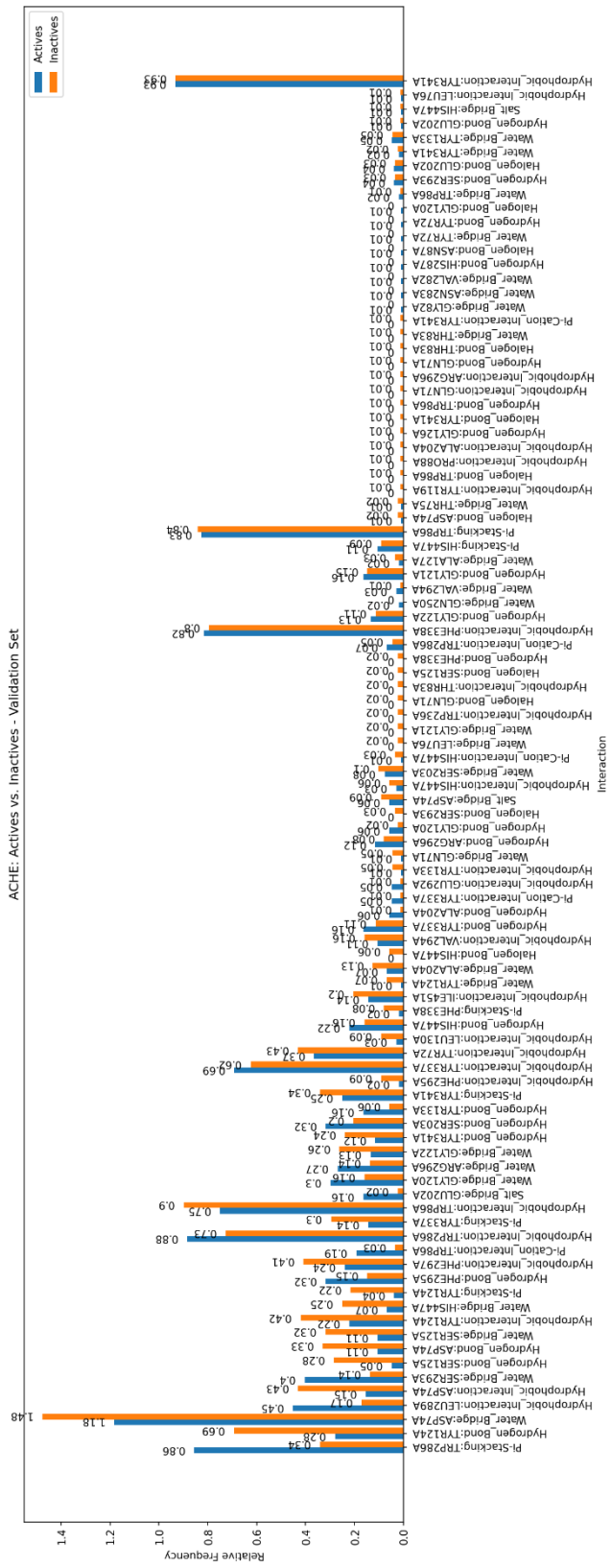


Fig. 6: Distribution of interaction frequencies of active and inactive ACHE ligands in the validation partition in comparison.

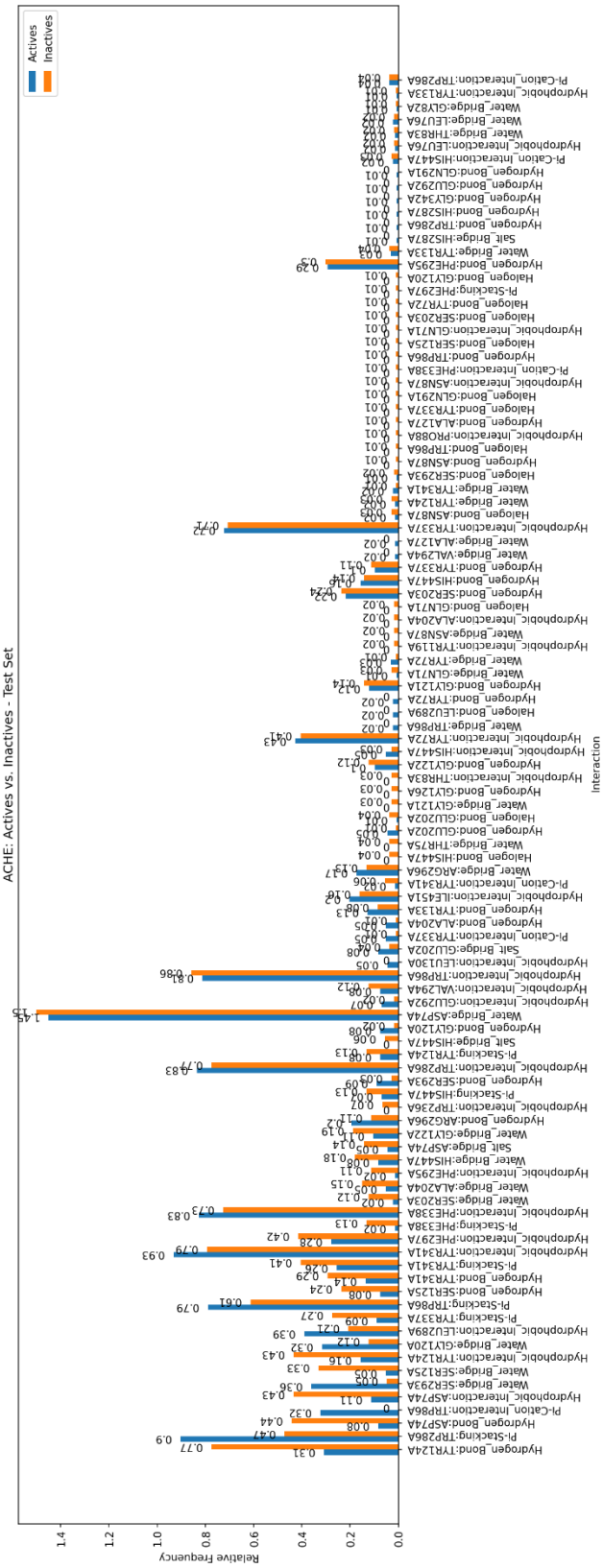


Fig. 7: Distribution of interaction frequencies of active and inactive ACHE ligands in the test partition in comparison.

3.2.2 Scoring

The ACHE dataset for scoring consisted of 1195 compounds, 664 of them active and 531 inactive. The baseline prediction accuracy was therefore 55.6%. All of the ligands were assigned into one of training, validation or test partition and docked in PDB structure 4EY7. The docking result was 10 poses for every ligand, meaning 11 950 structures to be analysed. Subsequently all structures were analysed and scored with PIA as described in 2.3.3. The best-on-validation (best accuracy on the validation partition) scoring strategy was strategy +-. Furthermore, the respective cut-off values were 1 for strategy +, 2 for strategy ++, -2 for strategy +-, and -2 for strategy ++--.

Results on the training partition:

Interaction frequencies of active and inactive molecules in the training partition can be seen in **Fig. 5**. The best-on-validation scoring strategy achieved a classification accuracy of 74.9% on the training dataset. A full overview of all scoring strategies and their corresponding metrics for the training partition can be seen in **Table 1**, the confusion matrix of the best-on-validation strategy in **Fig. 8**, and the ROC curve of the best-on-validation strategy in **Fig. 9**.

Table 1: Performance metrics for all scoring strategies evaluated on the training partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.640	0.484	0.684	0.659	1.179	73.770
++	0.652	0.184	0.697	0.782	1.400	78.246
+-	0.749	0.306	0.819	0.766	1.371	79.157
++--	0.737	0.220	0.813	0.802	1.435	90.214

Results on the validation partition:

In **Fig. 6** the interaction frequencies of active and inactive ligands of the validation partition are shown. The best-on-validation strategy achieved a prediction accuracy of 76.6% on the validation data. A complete list of performance metrics of all scoring strategies for the validation partition can be viewed in **Table 2**. The confusion matrix and ROC curve of the best-on-validation strategy are described in **Fig. 10** and **Fig. 11** respectively.

Table 2: Performance metrics for all scoring strategies evaluated on the validation partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.703	0.420	0.751	0.694	1.282	80.769
++	0.661	0.205	0.749	0.76	1.403	76
+-	0.766	0.318	0.831	0.757	1.397	83.654
++--	0.734	0.273	0.812	0.762	1.407	76.238

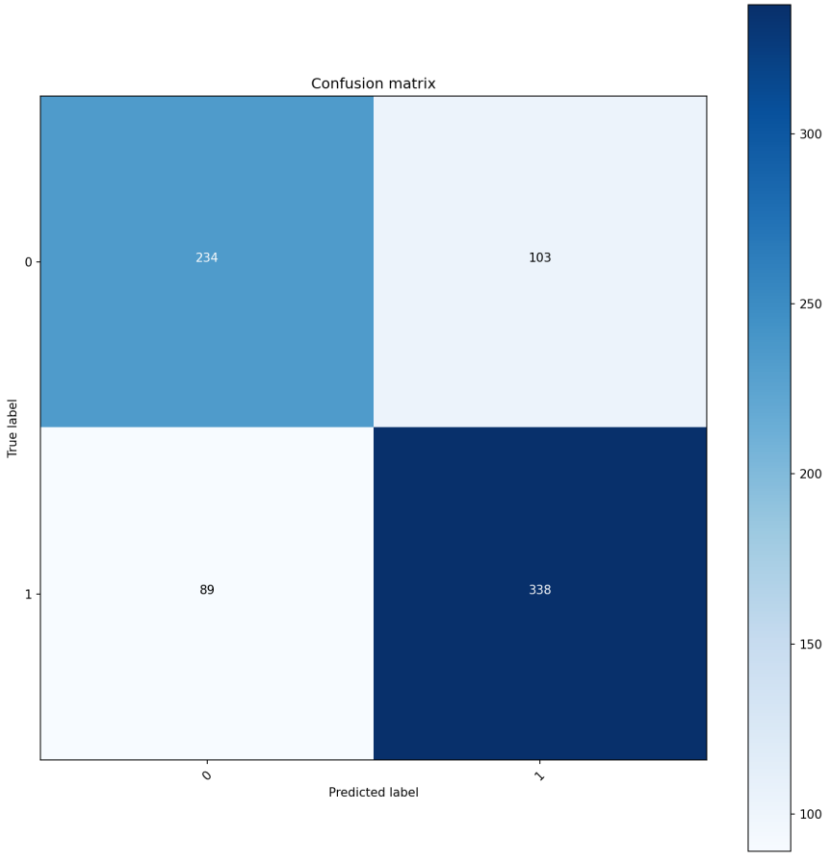


Fig. 8: Confusion matrix of the best-on-validation scoring strategy on the training data.

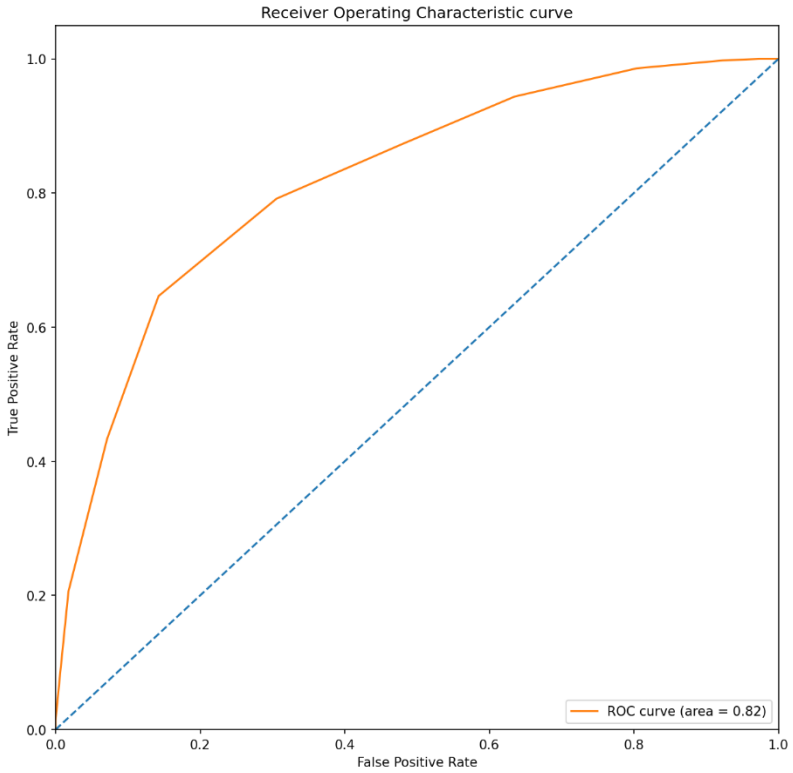


Fig. 9: ROC curve of the best-on-validation scoring strategy on the training data.

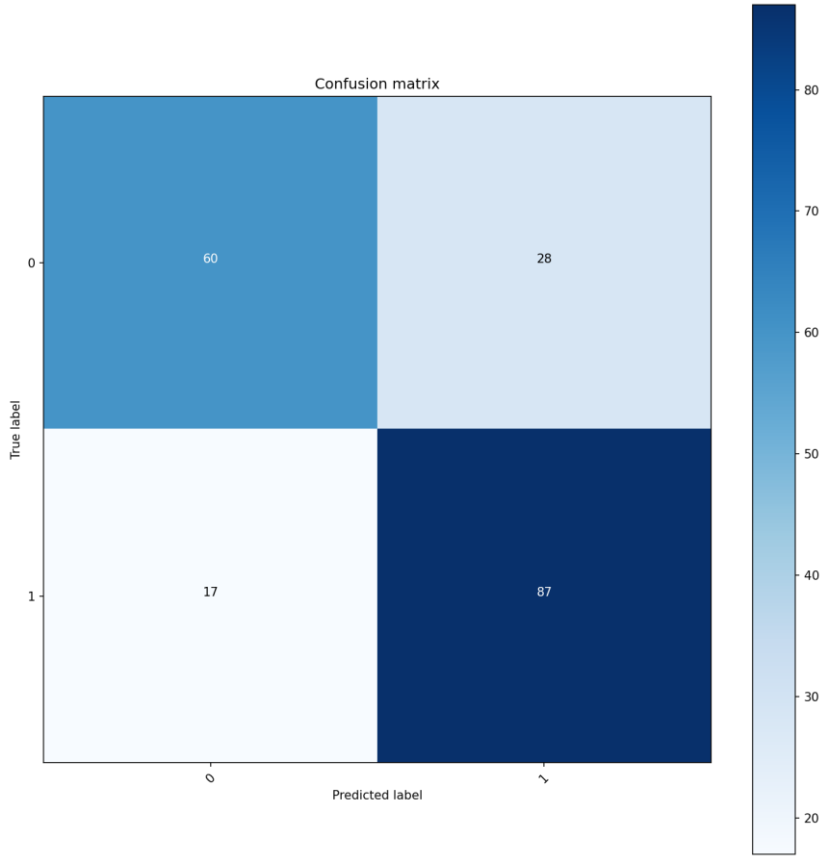


Fig. 10: Confusion matrix of the best-on-validation scoring strategy on the validation data.

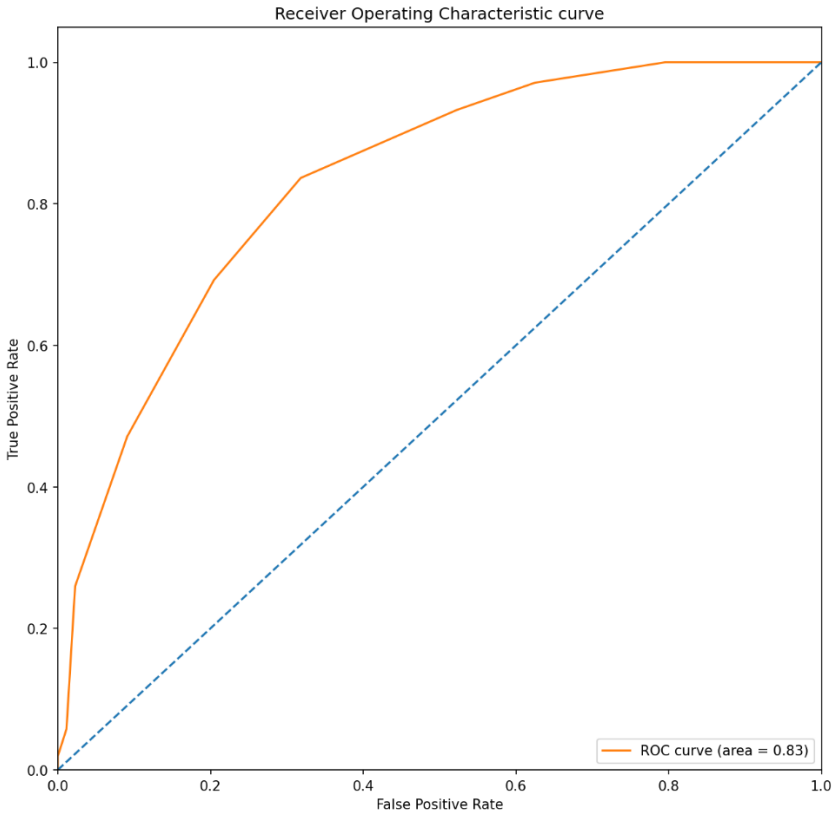


Fig. 11: ROC curve of the best-on-validation scoring strategy on the validation data.

Results on the test partition:

The interaction frequencies of active and inactive ACHE ligands in the test partition are show in **Fig. 7**. Out of all ligands 74.9% were classified correctly as active or inactive by the best-on-validation scoring strategy. An exhaustive list of performance metrics for all scoring strategies evaluated on the test partition is shown in **Table 3**. Confusion matrix and ROC curve of the best-on-validation scoring strategy are available in **Fig. 12** and **Fig. 13**.

Table 3: Performance metrics for all scoring strategies evaluated on the test partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.690	0.472	0.733	0.686	1.232	81.955
++	0.665	0.208	0.747	0.773	1.389	77.320
+-	0.749	0.302	0.831	0.766	1.377	78.947
++--	0.745	0.226	0.812	0.8	1.438	80

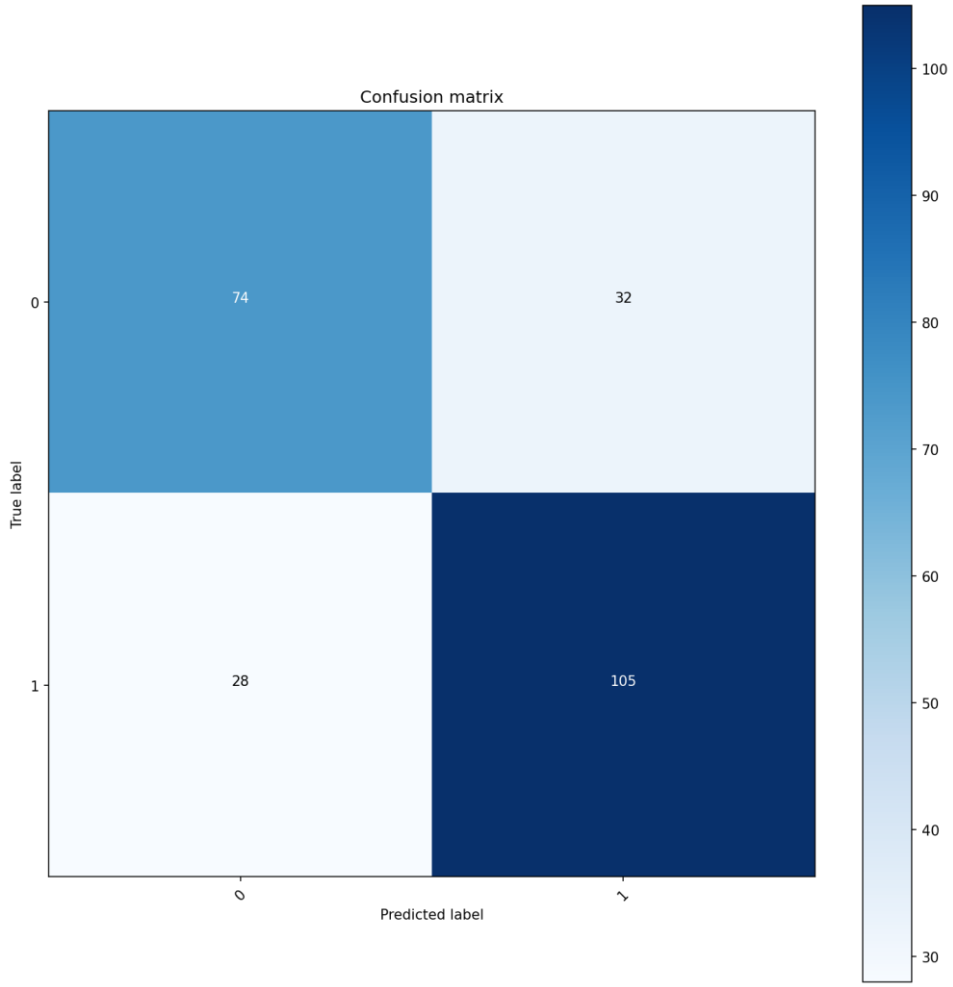


Fig. 12: Confusion matrix of the best-on-validation scoring strategy on the test data.

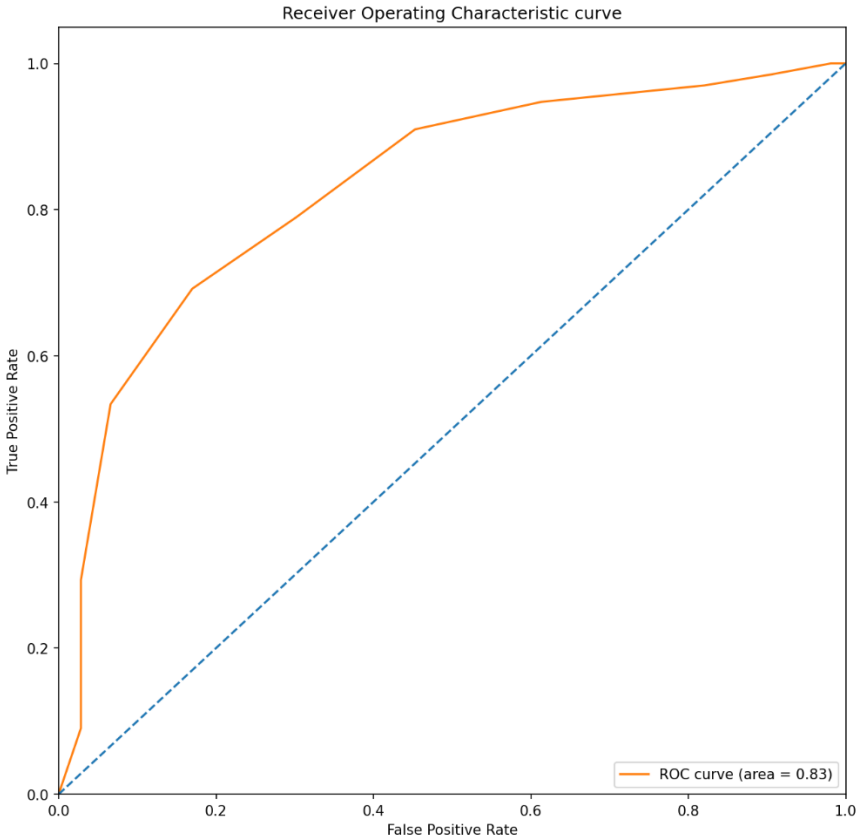


Fig. 13: ROC curve of the best-on-validation scoring strategy on the test data.

3.3 Coagulation factor Xa

In **Fig. 14** the interactions and frequencies extracted from selected PDB structures for FXA are shown. The selected set of PDB structures contained 103 protein-ligand complexes and all three residues that were known to be involved in binding are present and detected by PLIP/PIA.

- ASP189A: This aspartic acid forms a hydrogen bond in 11% of the structures or a water-mediated hydrogen bond (water bridge) in 1% of the structures.
- ALA190A: This alanine residue shows hydrophobic interactions in 31% of the structures and interacts via hydrogen bonding in 9% of the structures.
- GLN192A: Four different interaction modes are possible with this residue, namely hydrogen bonding in 48%, hydrophobic interactions in 9%, water bridges in 3%, and halogen bonding in 1% of the structures.

The top 5 interactions were:

- Hydrogen bonding with GLY216A in 73% of the structures.
- Pi-stacking with TRP215A in 65% of the structures.
- Metal complexation with GLU80A in 64% of the structures.
- Hydrophobic interactions with TRP215A in 63% of the structures.
- Metal complexation with ASP70A and ASN72A in 61% of the structures.

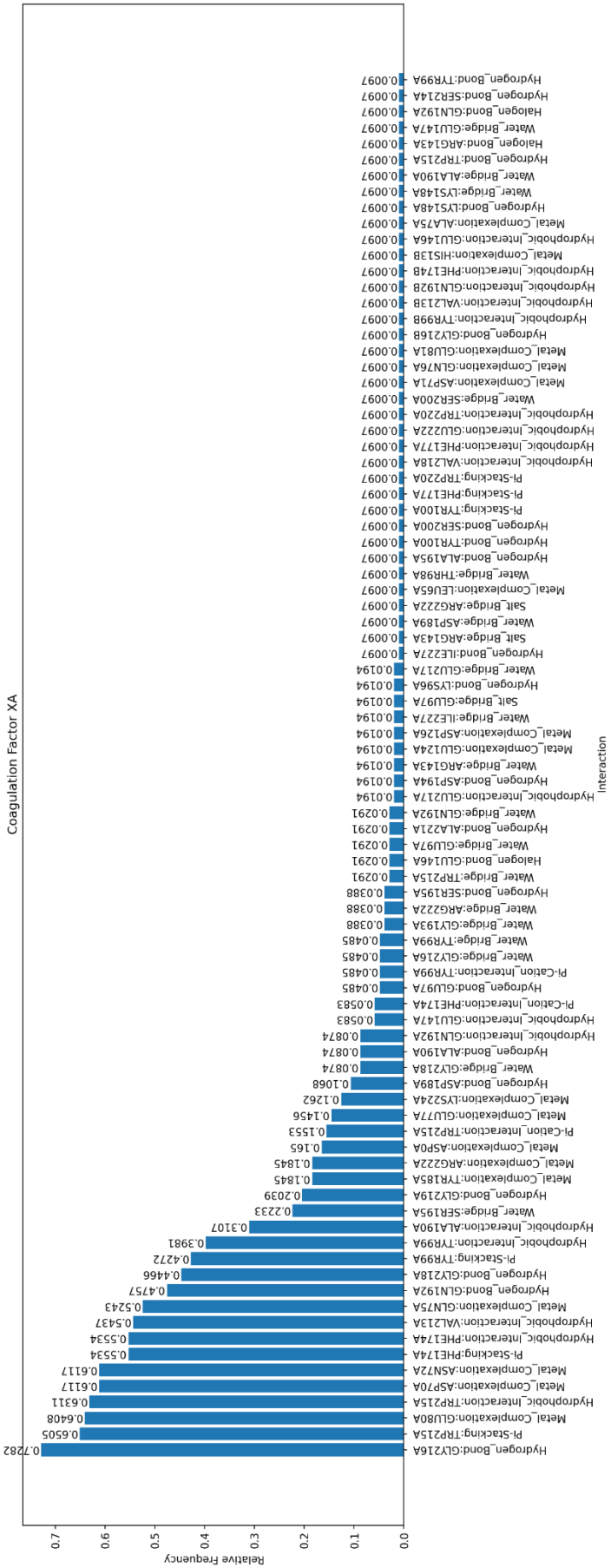


Fig. 14: Interaction frequencies of selected FXA structures from the PDB.

3.4 Cyclooxygenase 1

COX1 was the second target that was scored and analysed in terms of interactions present in the selected structures of the Protein Data Bank. Again the interaction frequencies will be described first.

3.4.1 Interaction frequencies

The plot in **Fig. 15** shows interactions and frequencies of the 20 structures that were selected from the PDB. Almost all known interactions and residues can be found among these results:

- ILE523A: A hydrophobic interaction is present in 45% of the structures.
- ARG120A: Two binding modes are present for this residue, the more frequent salt bridge occurs in 70% of the structures while the less frequent pi-cation interaction appears in 5% of the structures.
- TYR355A: This tyrosine residue exhibits hydrophobic interactions in 70% of the structures. In 50% of the structures it forms a hydrogen bond.
- TYR385A: A hydrophobic interaction is present in 45% of the structures.
- SER530A: Serine at position 530 forms a hydrogen bond in 20% of the structures. In 5% of the structures there is a hydrophobic interaction detected at this residue.

Two of the known interactions were not present or detected during the analysis: Interactions with ILE434 and hydrogen bonds with TYR385.

The top 5 interactions in terms of frequency were:

- Hydrophobic interactions with TRP387A in 75% of the structures.
- Salt bridges with ARG120A in 70% of the structures.
- Hydrophobic interactions with ALA527A, LEU352A, VAL349A and TYR355A also in 70% of the structures.

3.4.2 Scoring

The COX1 dataset for scoring consisted of 357 active compounds and 879 inactive compounds, meaning 1236 compounds in total. The distribution of active and inactive molecules was skewed in favour of the inactive molecules and the baseline classification accuracy was therefore 71.1%. The 1236 compounds were randomly assigned to training, validation and test partition and each ligand was docked in the PDB structure 4O1Z which resulted in 10 poses for each ligand. Consequentially 12 360 structures had to be analysed and each best pose was scored with PIA. The best-on-validation accuracy was achieved with the scoring strategy +-. Moreover, the respective cut-off values for the specific strategies were 5 for strategy +, 5 for strategy ++, 4 for strategy +-, and 4 for strategy ++--.

Results on the training partition:

A comparison of interaction frequencies of active and inactive molecules of the training partition is shown in **Fig. 16**. The best-on-validation scoring strategy achieved a prediction accuracy of 72.3% on the training data. A full overview of all applied scoring strategies and

their respective performance on the training dataset can be seen in **Table 4**. Additionally the confusion matrix and ROC curve of the best-on-validation strategy is shown in **Fig. 19** and **Fig. 20**.

Table 4: Performance metrics for all scoring strategies evaluated on the training partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.719	0.037	0.607	0.512	1.812	51.163
++	0.716	0.062	0.616	0.493	1.746	49.275
+-	0.723	0.025	0.633	0.563	1.993	56.25
++--	0.724	0.032	0.636	0.561	1.987	56.098

Results on the validation partition:

The interaction frequencies calculated for the active and inactive molecules of the validation partition are plotted in **Fig. 17**. For this dataset the best-on-validation scoring strategy achieved a classification accuracy of 71.7% and the corresponding confusion matrix and ROC curve can be seen in **Fig. 21** and **Fig. 22**. A complete list of performance metrics of all scoring strategies evaluated on the validation partition can be viewed in **Table 5**.

Table 5: Performance metrics for all scoring strategies evaluated on the validation partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.702	0.007	0.521	0.667	2.200	66.667
++	0.687	0.043	0.534	0.4	1.32	40
+-	0.717	0.014	0.600	0.75	2.475	75
++--	0.697	0.043	0.626	0.5	1.65	50

Results on the test partition:

Interaction frequencies of active and inactive COX1 ligands of the test partition are presented in **Fig. 18**. Of the 248 ligands in the test partition 70.2% were correctly predicted as active or inactive by the best-on-validation scoring strategy. The calculated performance metrics for all applied scoring strategies are described in **Table 6**. Furthermore, **Fig. 23** and **Fig. 24** show the respective confusion matrix and ROC curve of the best-on-validation strategy.

Table 6: Performance metrics for all scoring strategies evaluated on the test partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.702	0.063	0.670	0.5	1.676	50
++	0.701	0.092	0.680	0.529	1.774	52.941
+-	0.702	0.034	0.668	0.5	1.676	50
++--	0.714	0.034	0.658	0.6	2.011	60

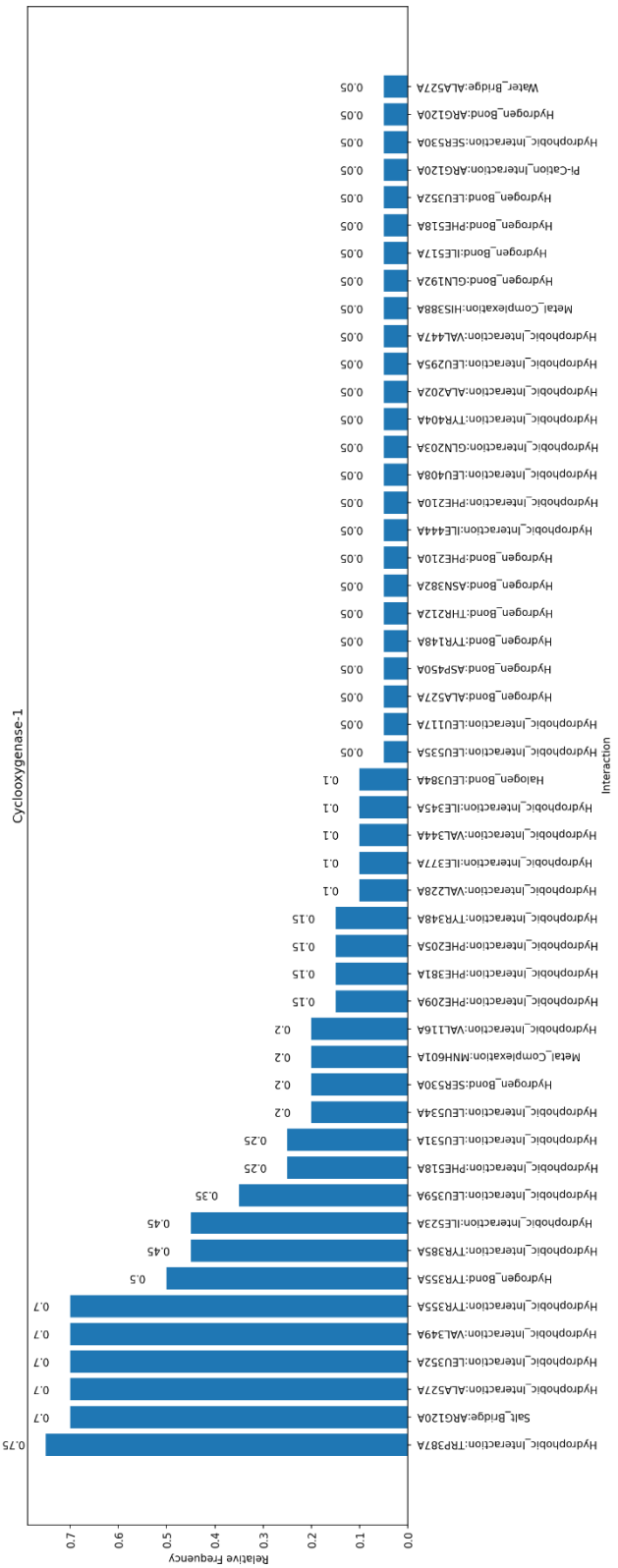


Fig. 15: Interaction frequencies of selected COX1 structures from the PDB.

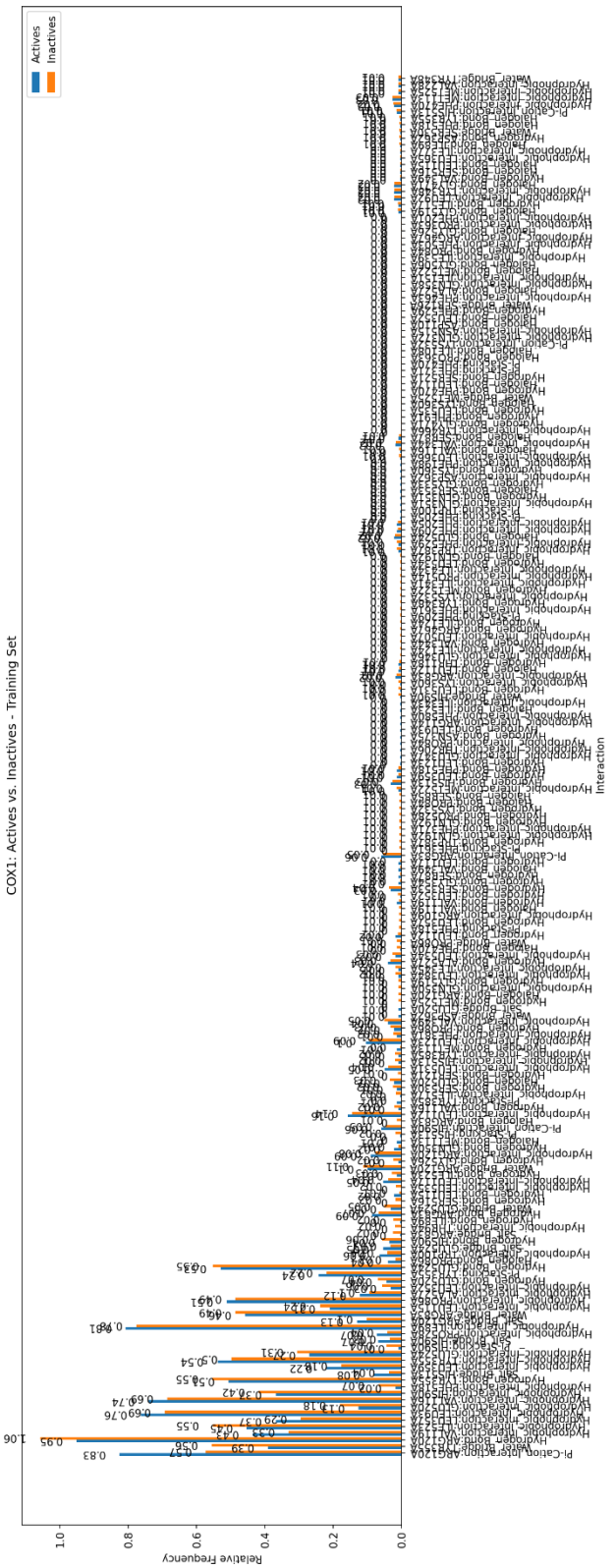


Fig. 16: Distribution of interaction frequencies of active and inactive COX1 ligands in the training partition in comparison.

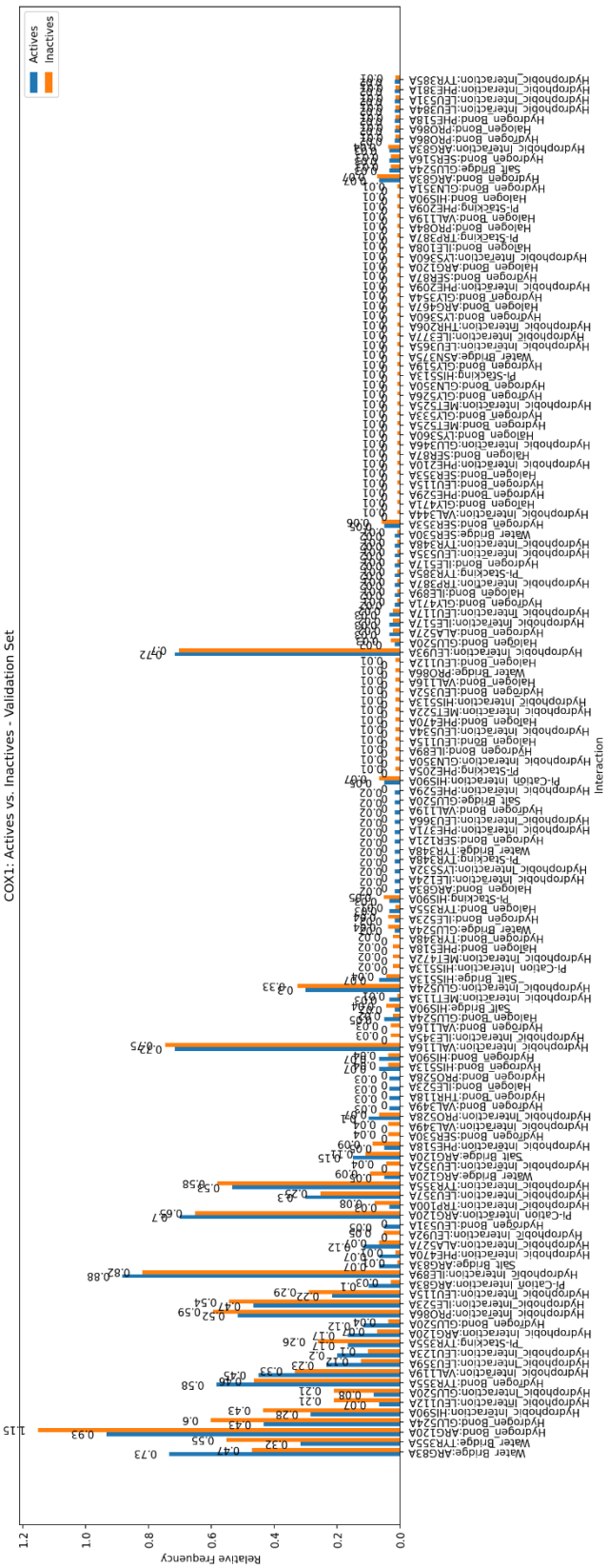


Fig. 17: Distribution of interaction frequencies of active and inactive COX1 ligands in the validation partition in comparison.

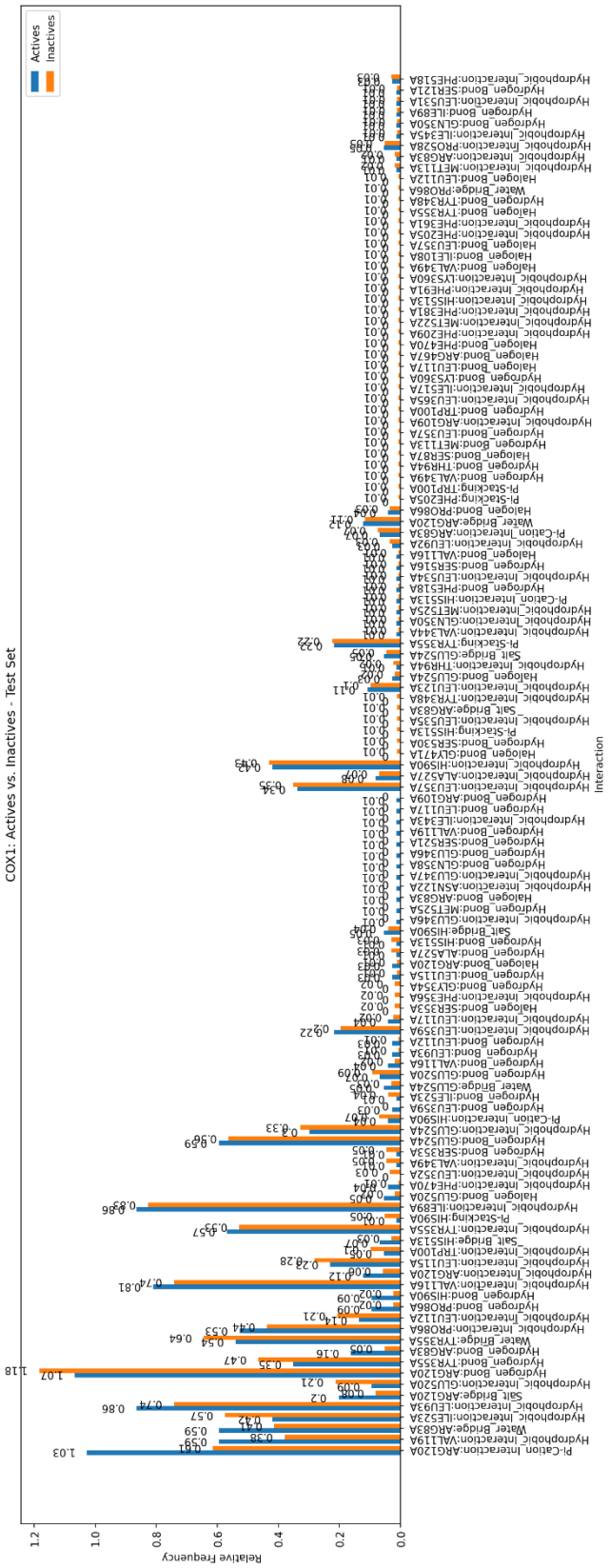


Fig. 18: Distribution of interaction frequencies of active and inactive COX1 ligands in the test partition in comparison.

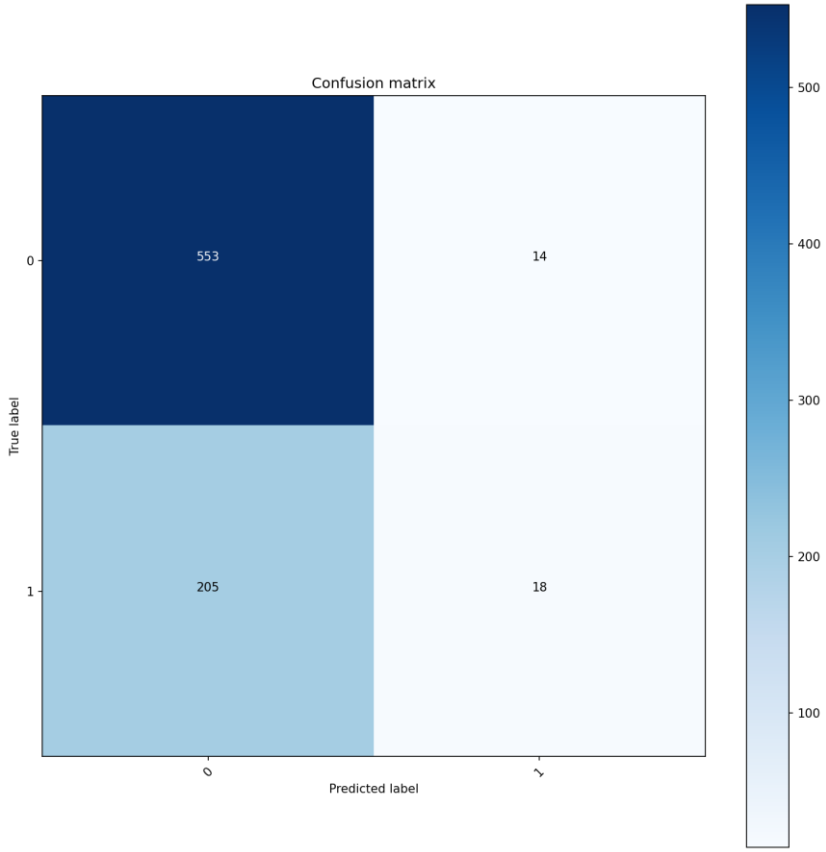


Fig. 19: Confusion matrix of the best-on-validation scoring strategy on the training data.

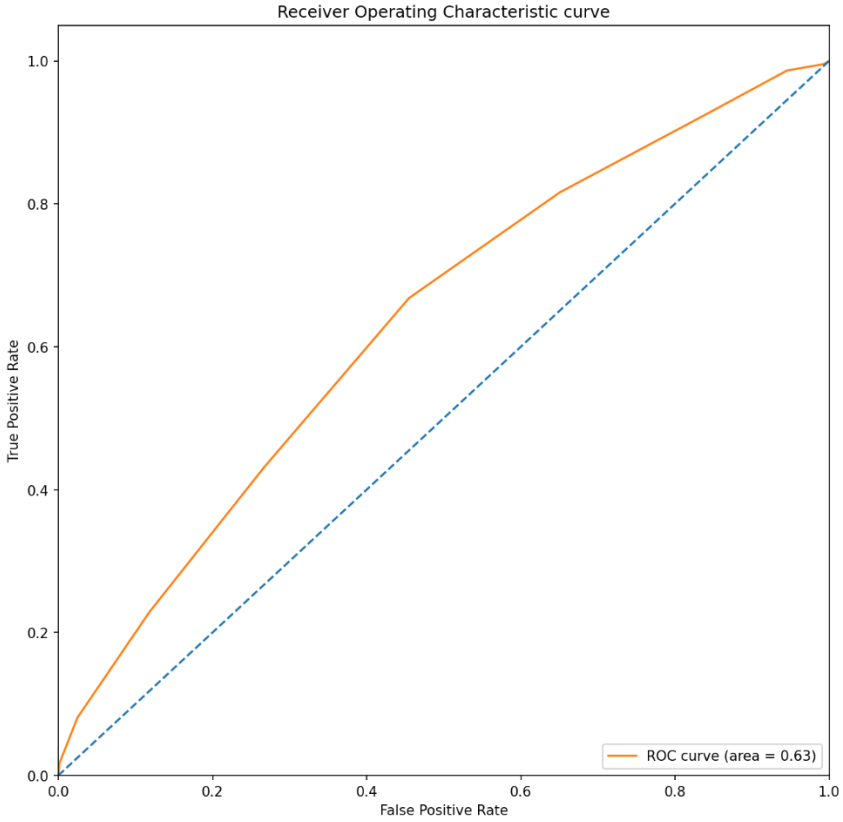


Fig. 20: ROC curve of the best-on-validation strategy on the training data.

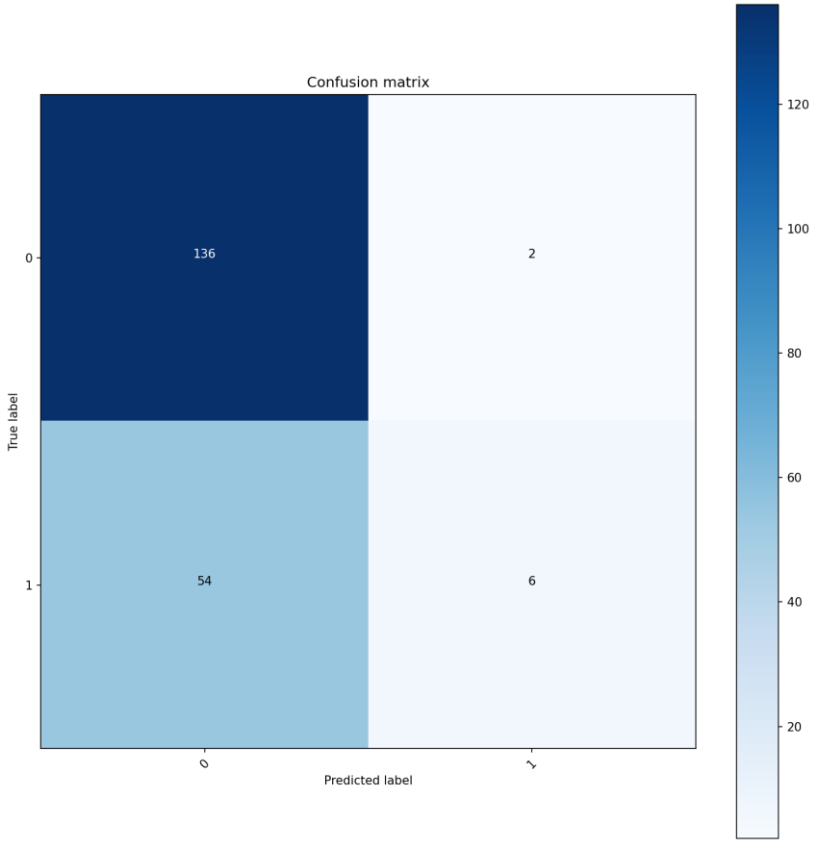


Fig. 21: Confusion matrix of the best-on-validation scoring strategy on the validation data.

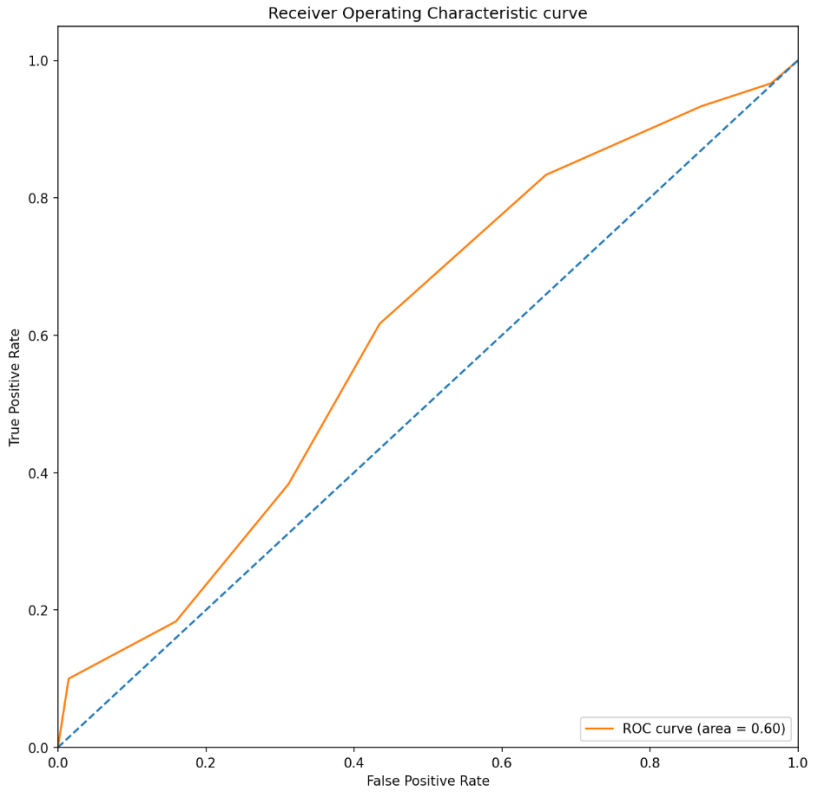


Fig. 22: ROC curve of the best-on-validation scoring strategy on the validation data.

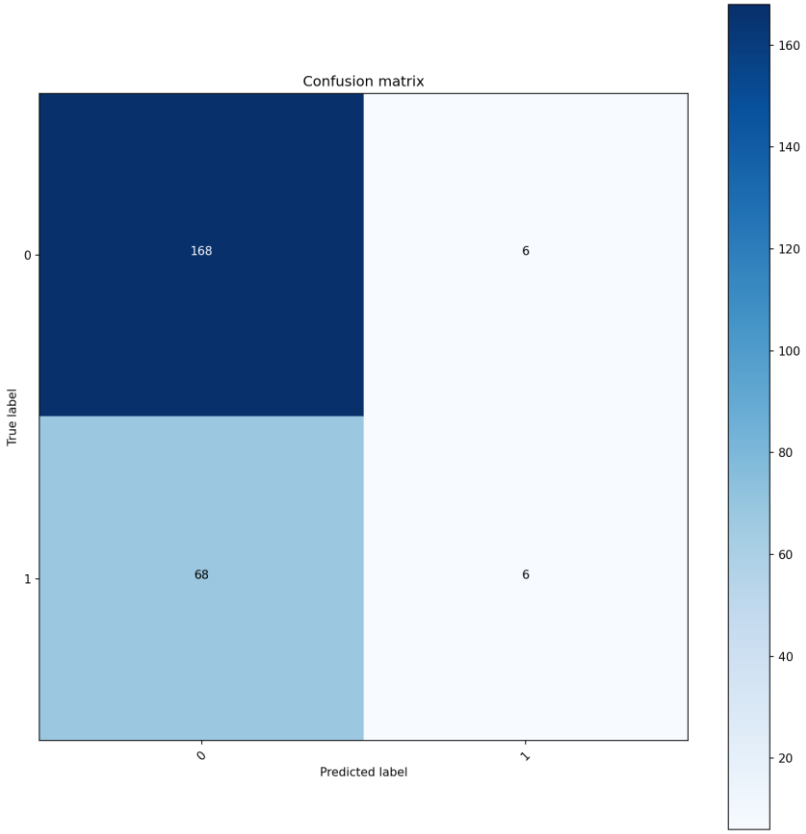


Fig. 23: Confusion matrix of the best-on-validation scoring strategy on the test data.

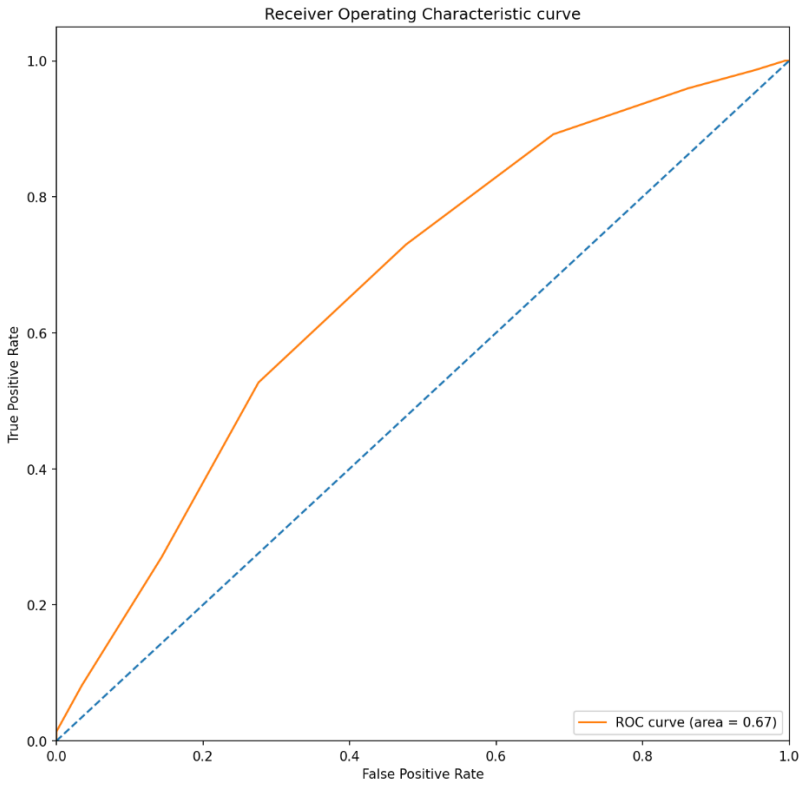


Fig. 24: ROC curve of the best-on-validation scoring strategy on the test data.

3.5 Cyclooxygenase 2

The interaction frequencies of selected PDB structures for human COX2 are shown in **Fig. 25**. The sample size of human COX2 structures available in the PDB is relatively small and in total 6 structures were analysed.

The following interactions were known form literature and also present in the results:

- VAL523A: A hydrophobic interaction is detected in 83% of the structures.
- TYR355A: A hydrophobic interaction is present in 50% of the structures.
- TYR385A: This residue forms a hydrogen bond in all structures. Additionally it reacts via hydrophobic interactions with the ligand in 66% of the structures.
- SER530A: A hydrogen bond is formed in 66% of the structures.

Interactions with residue VAL434 and ARG120 – although described in literature – were not picked up, either because they were not present or not detected.

The top 5 interactions in terms of frequency were:

- The hydrogen bond with TYR385A that was present in all structures.
- Hydrophobic interactions with VAL349A, ALA527A, TRP387A and VAL523A. All occurring at a frequency of 83%.

The interaction frequencies of mouse COX2 structures from the PDB can be seen in **Fig. 26**. More protein-ligand complexes were available compared to human COX2 with a total of 35 structures being analysed.

For mouse COX2 the interactions listed below were described in literature and picked up by PIA:

- VAL523A: The valine residue shows hydrophobic interactions in 57% of the structures.
- ARG120A: Four possible binding modes were detected, a salt bridge in 26% of the structures, a hydrogen bond in 9% of the structures, a pi-cation interaction in 3% of the structures, and a halogen bond also in 3% of the structures.
- TYR355A: In 37% of the structures there is a hydrophobic interaction occurring with this tyrosine, in 34% of the structures a hydrogen bond is formed.
- TYR385A: The most frequent interaction with this residue is a hydrophobic interaction which occurs in 43% of the structures. Secondly, a hydrogen bond is formed with this residue in 20% of the structures.
- SER530A: Serine 530 hydrogen bonds in 40% of the structures. Furthermore, in 3% of the structures it hydrogen bonds via an intermediate water molecule.

Again an interaction with VAL434 was not detected.

The top 5 interactions in terms of frequency were all hydrophobic interactions, either with TRP387A in 57%, with VAL523A in 51%, with VAL349A in 46%, with LEU352A in 46%, or/and with TYR385A in 43% of the structures.

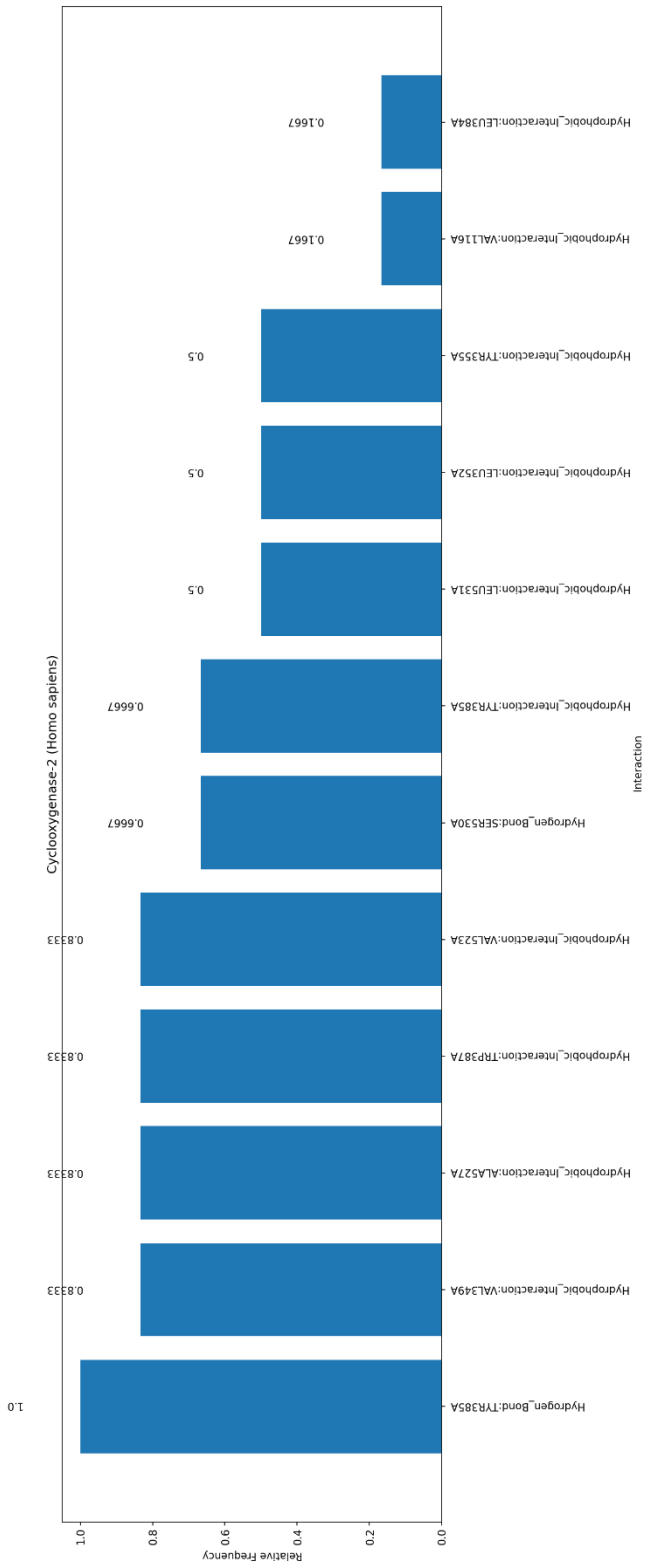


Fig. 25: Interaction frequencies of selected human COX2 structures from the PDB.

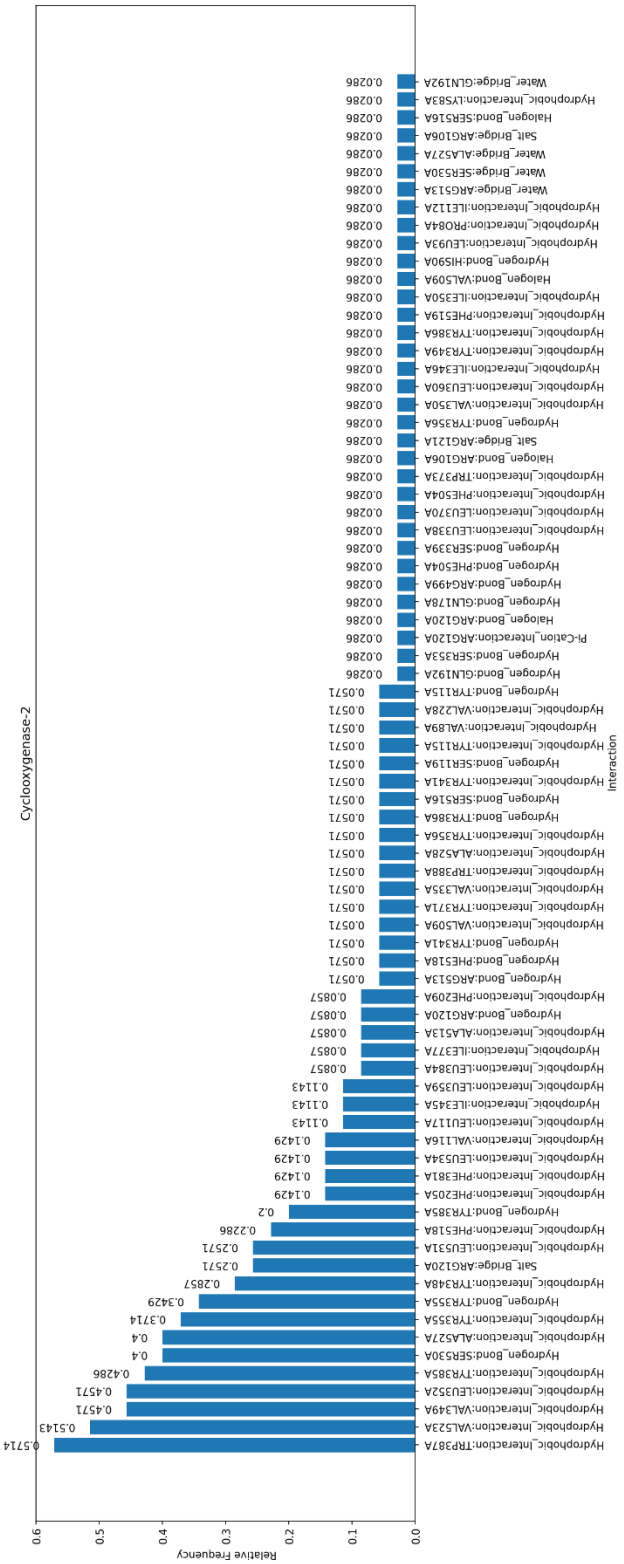


Fig. 26: Interaction frequencies of selected mouse COX2 structures from the PDB.

3.6 Dipeptidyl peptidase IV

DPP4 was the third target that was scored additionally to the interaction frequency analysis of selected PDB structures. The interaction frequencies will be discussed first again and the scoring results afterwards.

3.6.1 Interaction frequencies

Interactions and their respective frequencies have been extracted from 78 PDB structures containing DPP4. A graphical representation of the result can be seen in **Fig. 27**. All of the important and literature established interactions and residues were present in the results generated by PIA:

- GLU205A: Glutamic acid in position 205 forms a hydrogen bond with the ligand in 97% of the structures. In 1% of the structures a hydrophobic interaction is present.
- GLU206A: Three possible binding options have been observed for GLU206A, hydrogen bonding in 37% of the structures, salt bridges in 4% of the structures, and hydrophobic interactions in 1% of the structures.
- ARG125A: The arginine residue also interacts in four different possible ways with the ligand, in 38% of the structures it interacts via hydrogen bonding, in 26% of the structures it forms a water bridge, in 9% of the structures a pi-cation interaction is observed, and in 5% of the structures a salt bridge forms.

The top 5 interactions in terms of frequency were:

- Hydrogen bonding with GLU205A in 97% of the structures.
- Hydrogen bonding with TYR662A in 72% of the structures.
- Hydrophobic interactions with VAL711A in 65% of the structures.
- Hydrophobic interactions with TYR662A in 54% of the structures.
- Pi-stacking with TYR666A in 53% of the structures.

3.6.2 Scoring

A total number of 2018 compounds was used for the scoring workflow of DPP4, of which 1043 were active ligands and 975 inactive. Following from that the baseline prediction accuracy was 51.7%. All of the 2018 compounds were randomly split into a training, validation and test partition and docked into the PDB structure 2G5T. After docking the resulting 20 180 structures were analysed with PIA and subsequently the best pose of each ligand was scored. The best-on-validation scoring strategy was strategy +-. Furthermore, the calculated optimal cut-off values were 6 for strategy +, 7 for strategy ++, 3 for strategy +-, and 4 for strategy ++--.

Results on the training partition:

The comparison of interaction frequencies between active and inactive molecules of the training data is depicted in **Fig. 28**. The best-on-validation scoring strategy achieved a classification accuracy of 66.2% on the training data as shown in **Table 7** together with the performance

metrics of all other evaluated strategies. Furthermore, **Fig. 31** and **Fig. 32** show the confusion matrix and ROC curve of the best-on-validation strategy on the training partition.

Table 7: Performance metrics for all scoring strategies evaluated on the training partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.648	0.509	0.683	0.618	1.220	80.122
++	0.642	0.468	0.679	0.622	1.227	74.924
+-	0.662	0.358	0.724	0.662	1.306	68.196
++--	0.670	0.278	0.727	0.696	1.374	69.588

Results on the validation partition:

Interaction frequencies of active and inactive DPP4 ligands of the validation dataset are described in **Fig. 29**. The prediction accuracy of the best-on-validation scoring strategy was 70.9% on this split of the data and the according confusion matrix and ROC curve of this strategy are shown in **Fig. 33** and **Fig. 34** respectively. Calculated performance metrics of all four strategies based on the validation data are presented in **Table 8**.

Table 8: Performance metrics for all scoring strategies evaluated on the validation partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.678	0.5	0.670	0.671	1.211	82.123
++	0.669	0.458	0.699	0.676	1.221	77.095
+-	0.709	0.292	0.739	0.751	1.356	75.148
++--	0.706	0.243	0.739	0.773	1.394	77.273

Results on the test partition:

The interaction frequencies calculated for the active and inactive molecules of the test partition are plotted in **Fig. 30** and 65.1% of the 404 protein-complexes were correctly predicted as either active or inactive by the best-on-validation scoring strategy. The corresponding confusion matrix and ROC curve of this strategy and particular data split are shown in **Fig. 35** and **Fig. 36**. A complete list of performance metrics of all scoring strategies evaluated on the test partition can be viewed in **Table 9**.

Table 9: Performance metrics for all scoring strategies evaluated on the test partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.626	0.515	0.653	0.614	1.181	75.714
++	0.614	0.454	0.644	0.617	1.188	67.619
+-	0.651	0.330	0.685	0.675	1.299	67.513
++--	0.636	0.263	0.690	0.691	1.329	69.091

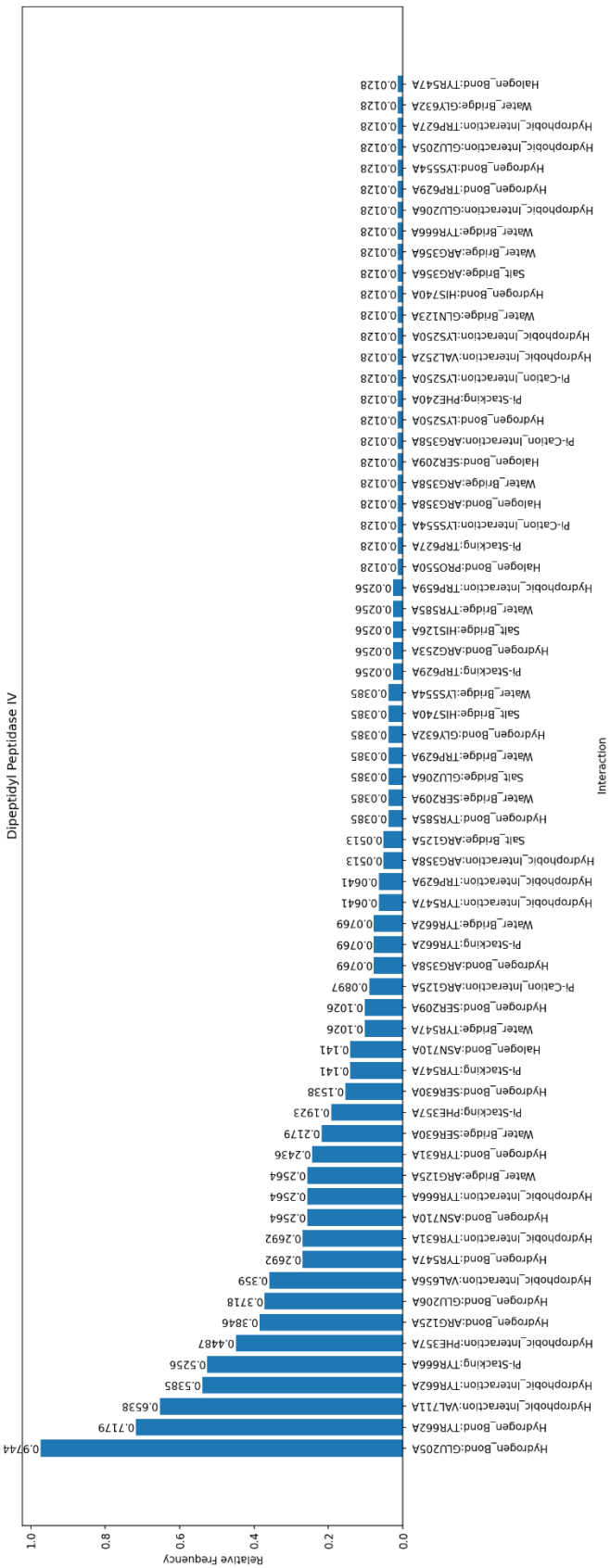


Fig. 27: Interaction frequencies of selected DPP4 structures from the PDB.

Fig. 28: Distribution of interaction frequencies of active and inactive DPP4 ligands in the training partition in comparison.

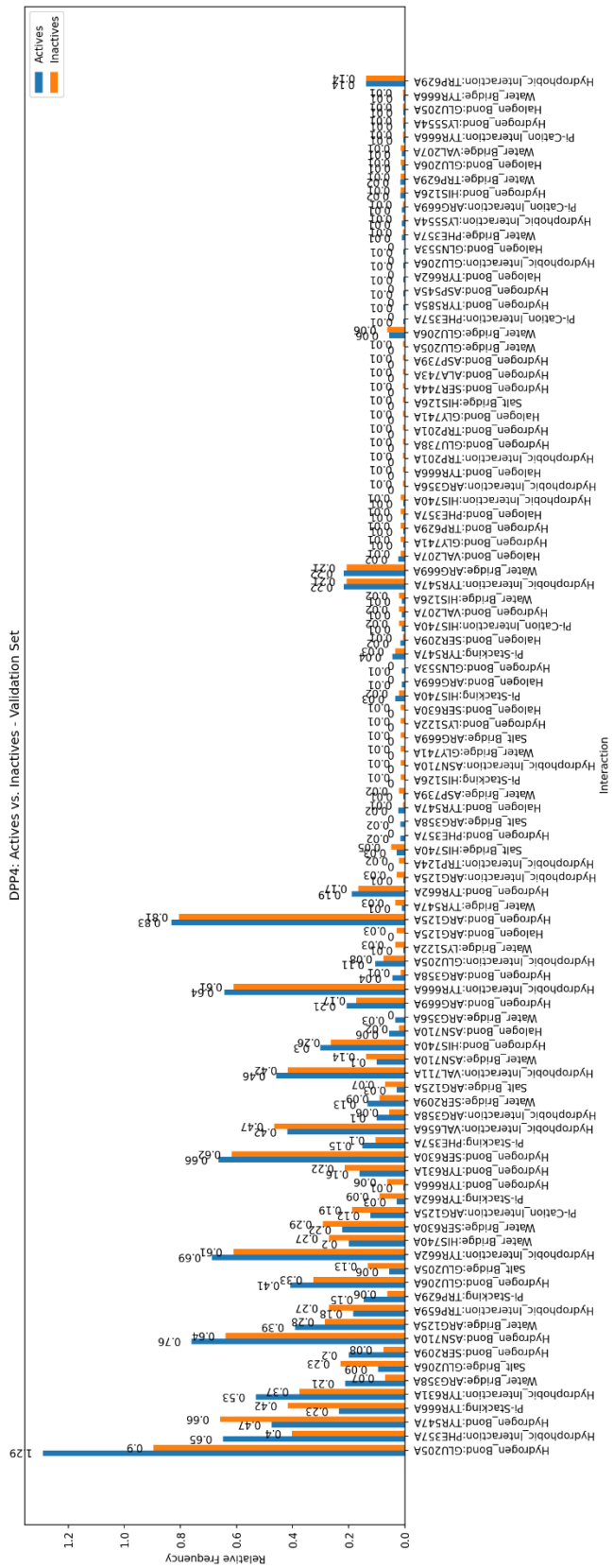


Fig. 29: Distribution of interaction frequencies of active and inactive DPP4 ligands in the validation partition in comparison.

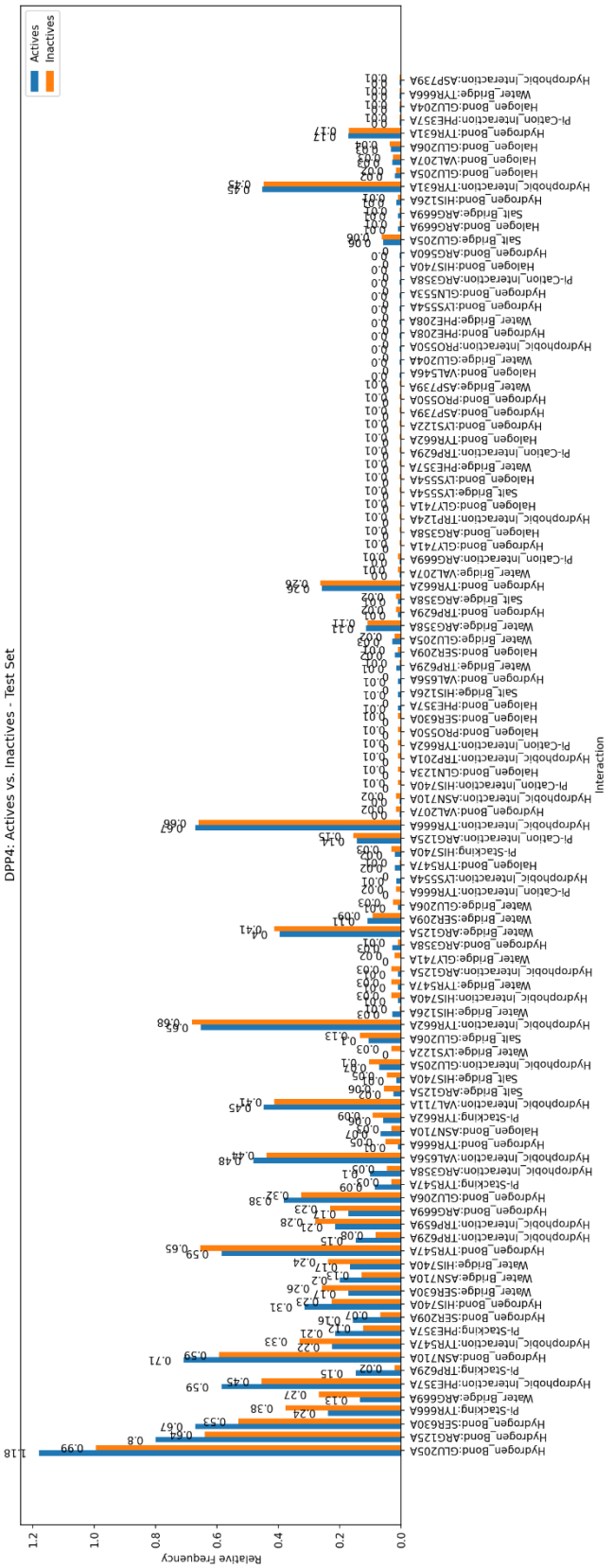


Fig. 30: Distribution of interaction frequencies of active and inactive DPP4 ligands in the test partition in comparison.

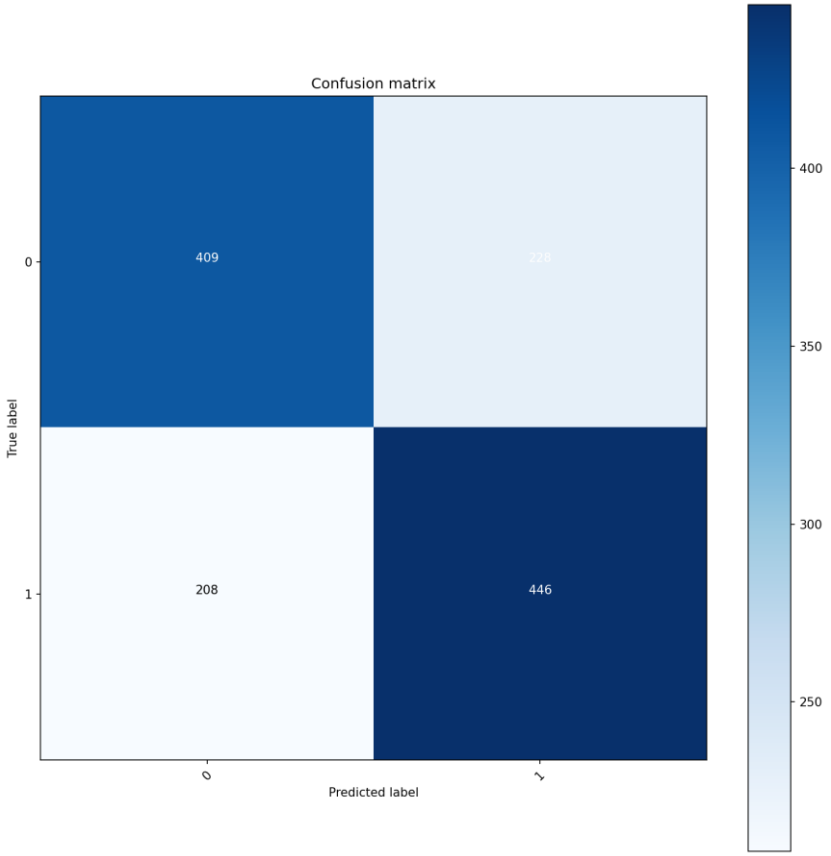


Fig. 31: Confusion matrix of the best-on-validation scoring strategy on the training data.

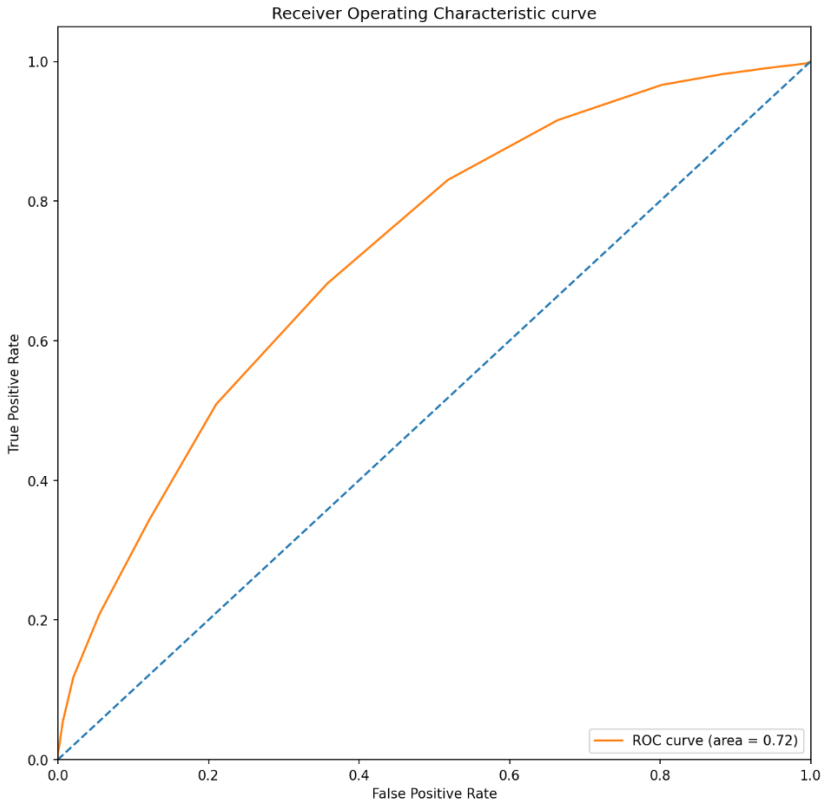


Fig. 32: ROC curve of the best-on-validation scoring strategy on the training data.

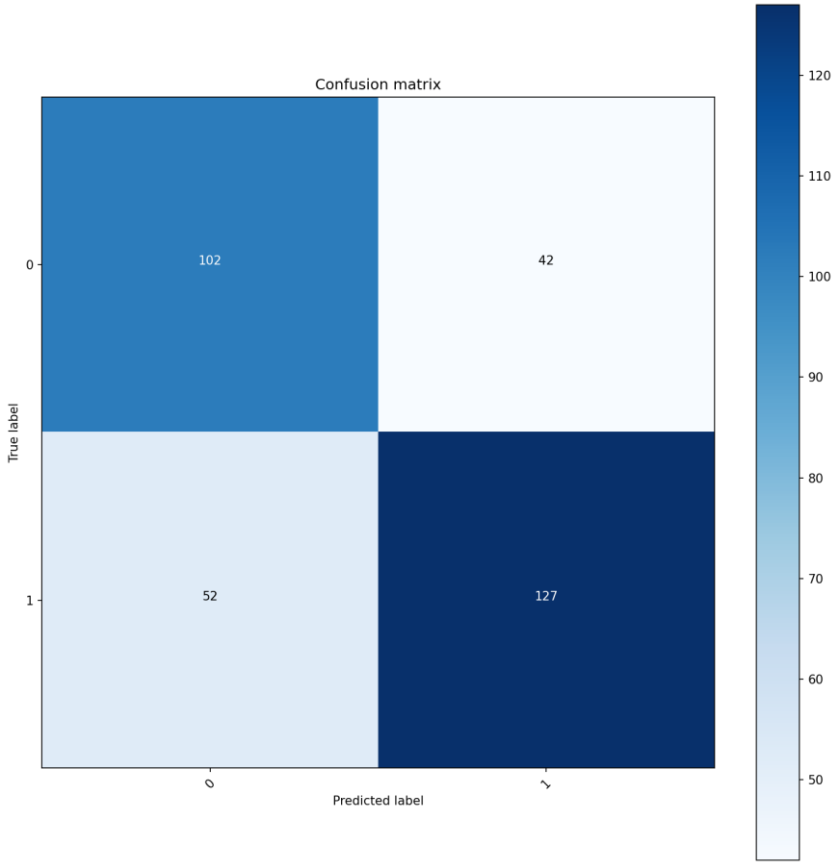


Fig. 33: Confusion matrix of the best-on-validation scoring strategy on the validation data.

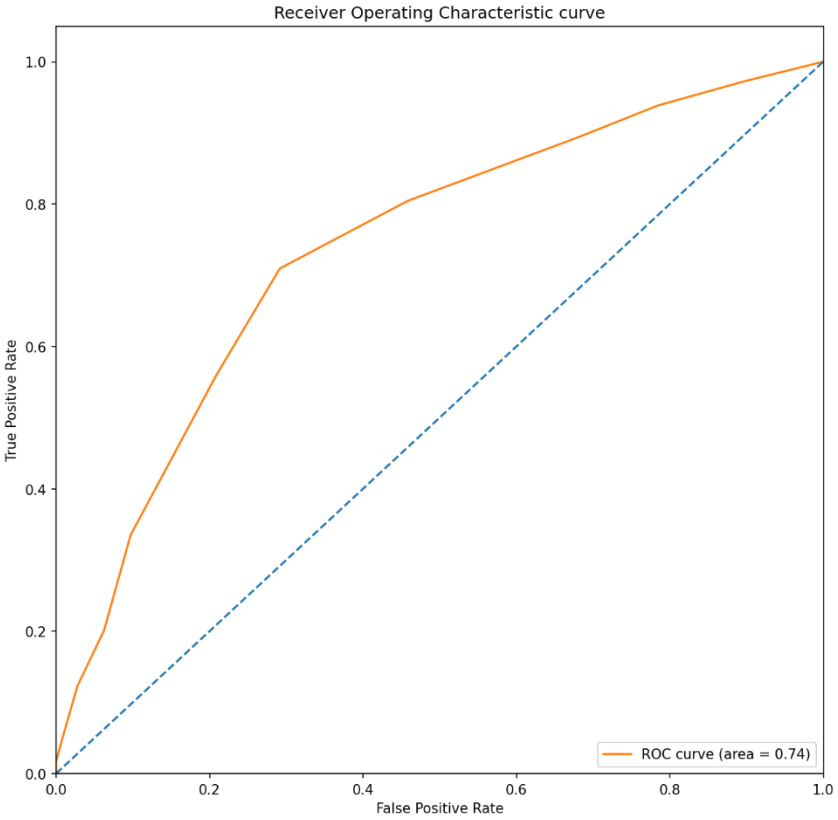


Fig. 34: ROC curve of the best-on-validation scoring strategy on the validation data.

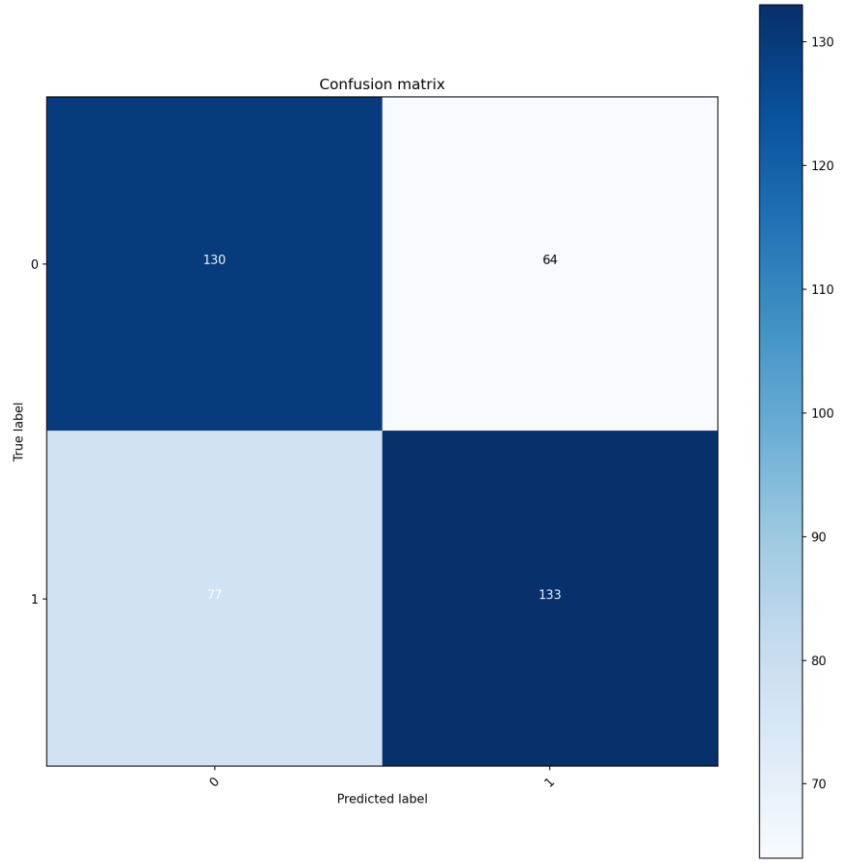


Fig. 35: Confusion matrix of the best-on-validation scoring strategy on the test data.

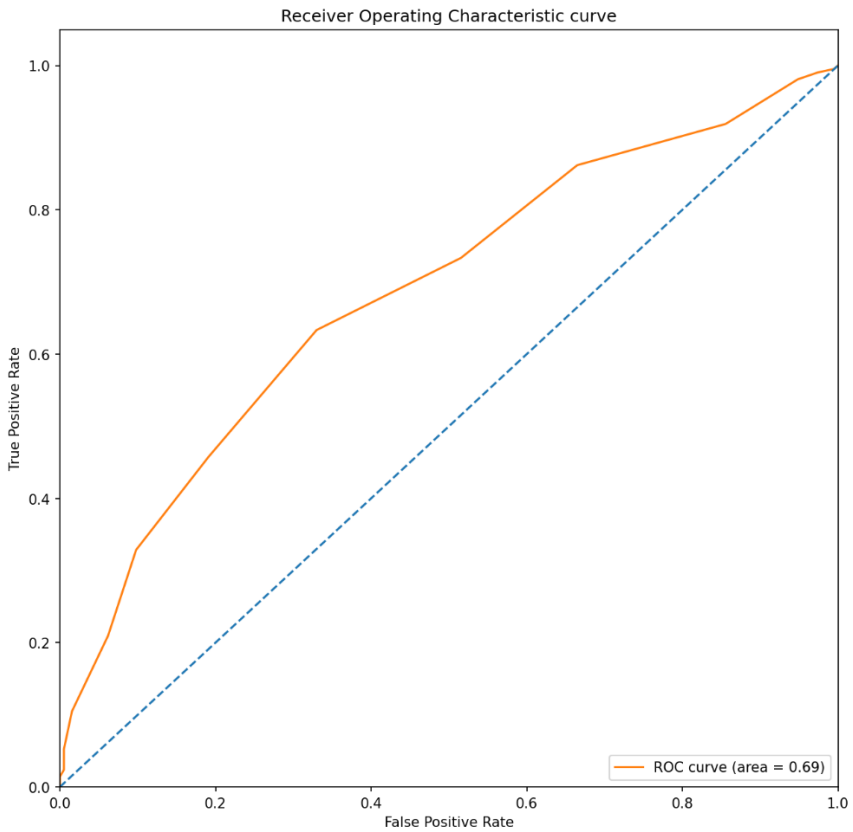


Fig. 36: ROC curve of the best-on-validation scoring strategy on the test data.

3.7 Monoamine oxidase B

MAOB was the fourth of the five targets that was both scored and analysed. The interaction frequency analysis based on the selected PDB structures is presented first.

3.7.1 Interaction frequencies

The bar chart in **Fig. 37** depicts the interactions and their respective frequencies of the 37 structures that were selected from the PDB. The three residues known from literature to be impactful for ligand binding are among the results and interact in the following way:

- ILE199A: A hydrophobic interaction is present in 38% of the structures.
- TYR398A: Three different binding modes are observed with the tyrosine residue, in 32% of the structures the residue takes part in hydrophobic interactions, in 8% of the structures it participates in pi-stacking, and in 3% of the structures it forms a hydrogen bond.
- TYR435A: In 8% of the structures a hydrogen bond is formed with TYR435A, in 3% of the structures a water bridge is detected, and also in 3% of the structures hydrophobic interactions occur with TYR435A.

The top 5 interactions in terms of frequency were:

- Pi-cation interactions with TRP157A in 54% of the structures.
- Hydrophobic interactions with LEU171A in 54% of the structures.
- Hydrophobic interactions with GLN206A and ILE199A in 38% of the structures.
- Hydrophobic interactions with PHE343A in 35% of the structures.

3.7.2 Scoring

The scoring dataset for MAOB featured 442 compounds in total of which 168 were active and 274 were inactive. The baseline prediction accuracy was therefore at 62%. The 442 compounds were randomly assigned to either the training partition, the validation partition or the test partition. Docking into PDB structure 2XCG yielded 10 poses for each ligand and subsequently 4420 structures were analysed by PIA and the best poses were scored. The best-on-validation accuracy was achieved with scoring strategy ++-- this time. The cut-off values for the different scoring strategies were 5 for strategy +, 5 for strategy ++, 4 for strategy +-, and 4 for strategy ++--.

Results on the training partition:

The interaction frequencies calculated for the active and inactive molecules in the training partition are depicted in **Fig. 38**. The best-on-validation scoring strategy achieved a classification accuracy of 68.8% on this split of the data, the corresponding confusion matrix and ROC curve are plotted in **Fig. 41** and **Fig. 42**. A complete overview of all applied scoring strategies and their performance on the training dataset can be viewed in **Table 10**.

Table 10: Performance metrics for all scoring strategies evaluated on the training partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.681	0.130	0.656	0.672	1.676	67.164
++	0.677	0.136	0.655	0.662	1.651	66.176
+-	0.699	0.047	0.672	0.818	2.042	81.818
++--	0.688	0.041	0.678	0.821	2.048	82.051

Results on the validation partition:

The comparison of interaction frequencies of active and inactive molecules of the validation partition is shown in **Fig. 39**. Moreover, the best-on-validation scoring strategy yielded a prediction accuracy of 73.2%. The resulting confusion matrix and ROC curve of that strategy are shown in **Fig. 43** and **Fig. 44**. The complete list of performance metrics of all scoring strategies evaluated on the validation partition is presented in **Table 11**.

Table 11: Performance metrics for all scoring strategies evaluated on the validation partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.634	0.234	0.525	0.45	1.331	45
++	0.648	0.234	0.523	0.476	1.409	47.619
+-	0.690	0.128	0.526	0.571	1.690	57.143
++--	0.732	0.064	0.554	0.727	2.152	72.727

Results on the test partition:

A comparative grouped bar chart of the interaction frequencies of active and inactive compounds in the test partition can be viewed in **Fig. 40**. For this split of the data the best-on-validation scoring strategy performed at a prediction accuracy of 73%. Again the belonging confusion matrix and ROC curve can be viewed in **Fig. 45** and **Fig. 46** respectively. Calculated performance metrics for all applied scoring strategies are depicted in **Table 12**.

Table 12: Performance metrics for all scoring strategies on the test partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.663	0.172	0.638	0.524	1.504	52.381
++	0.674	0.172	0.642	0.545	1.566	54.545
+-	0.719	0.034	0.702	0.8	2.297	80
++--	0.730	0	0.714	1	2.871	100

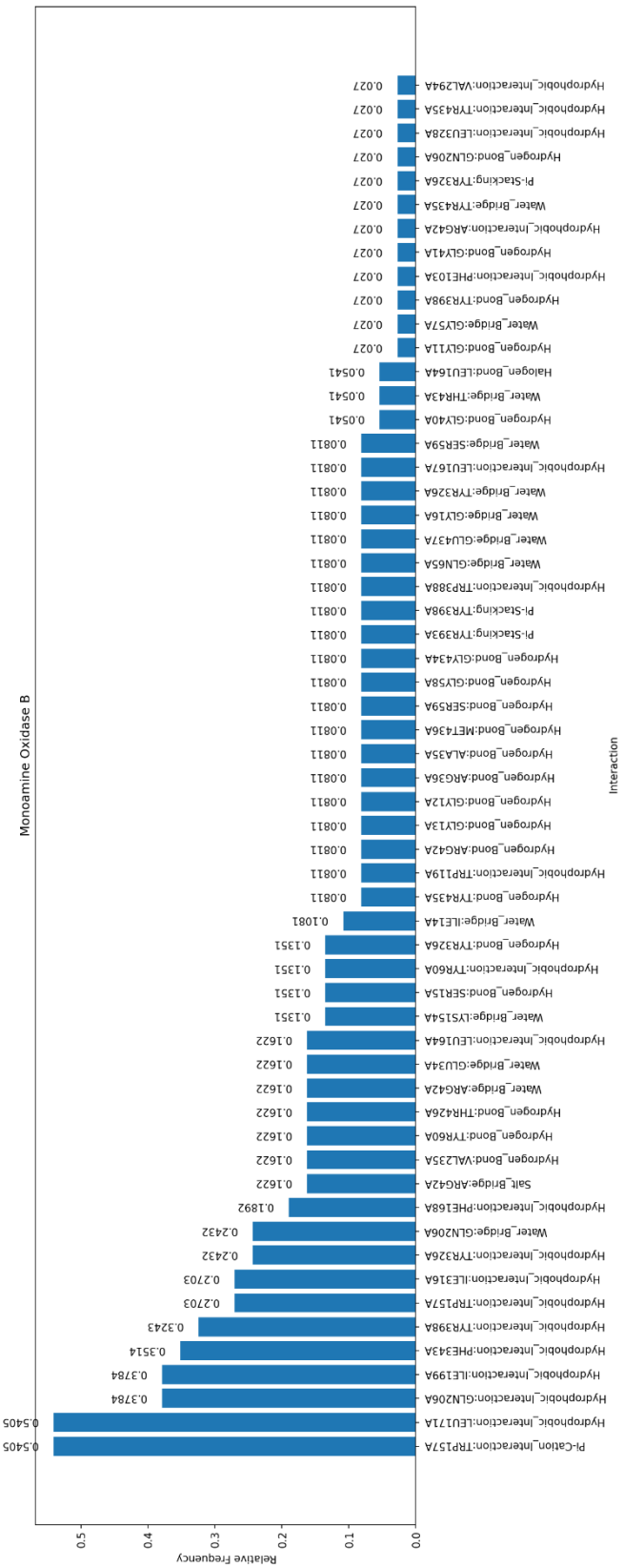


Fig. 37: Interaction frequencies of selected MAOB structures from the PDB.

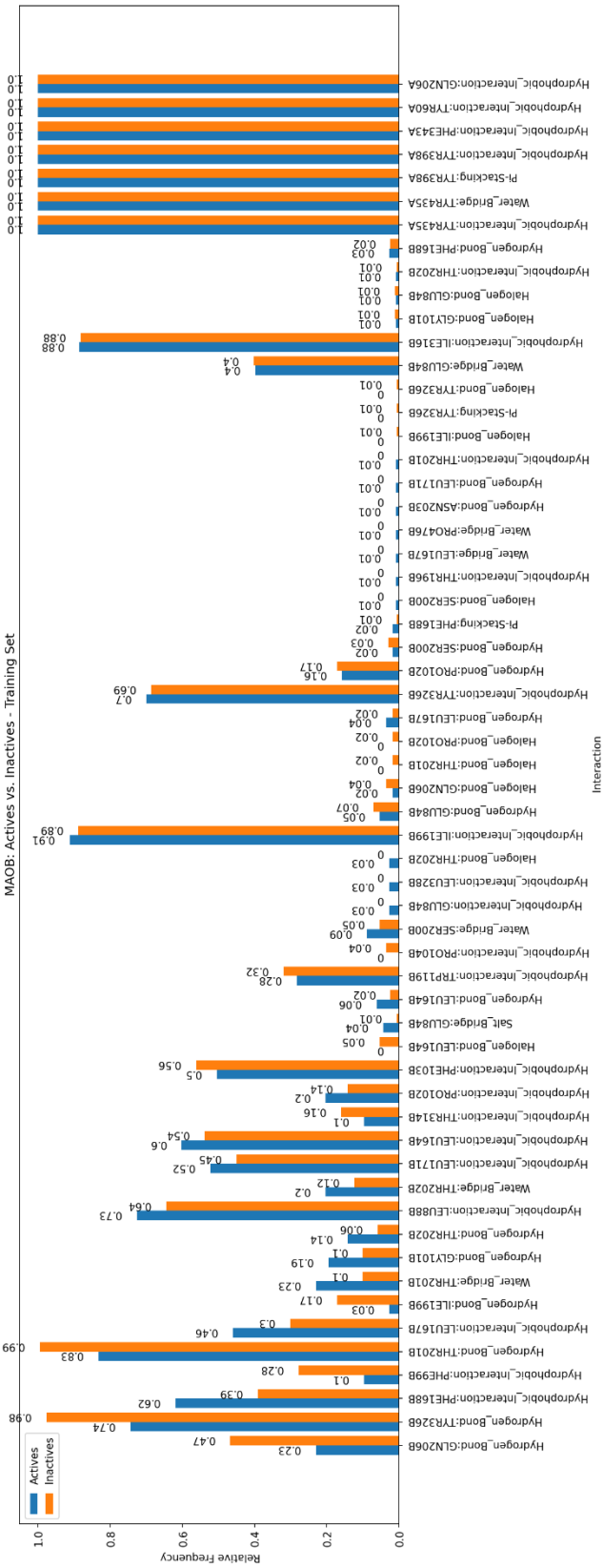


Fig. 38: Distribution of interaction frequencies of active and inactive MAOB ligands in the training partition in comparison.

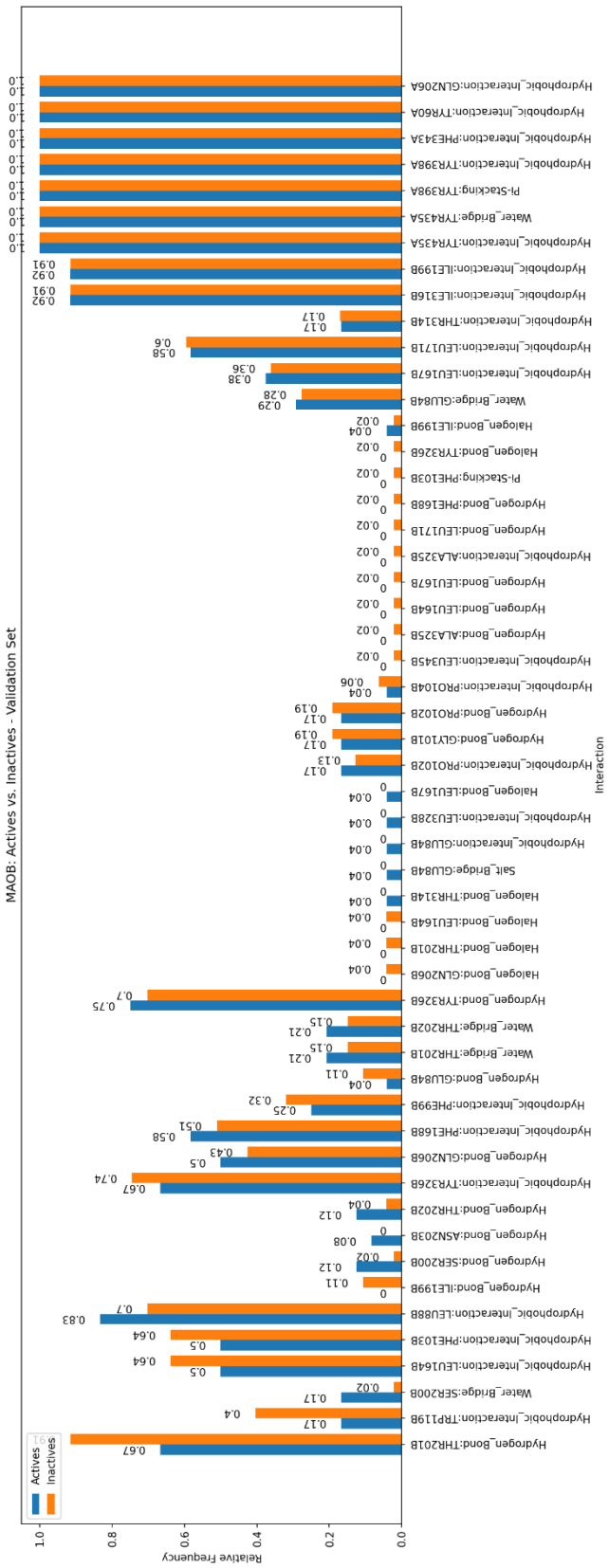


Fig. 39: Distribution of interaction frequencies of active and inactive MAOB ligands in the validation partition in comparison.

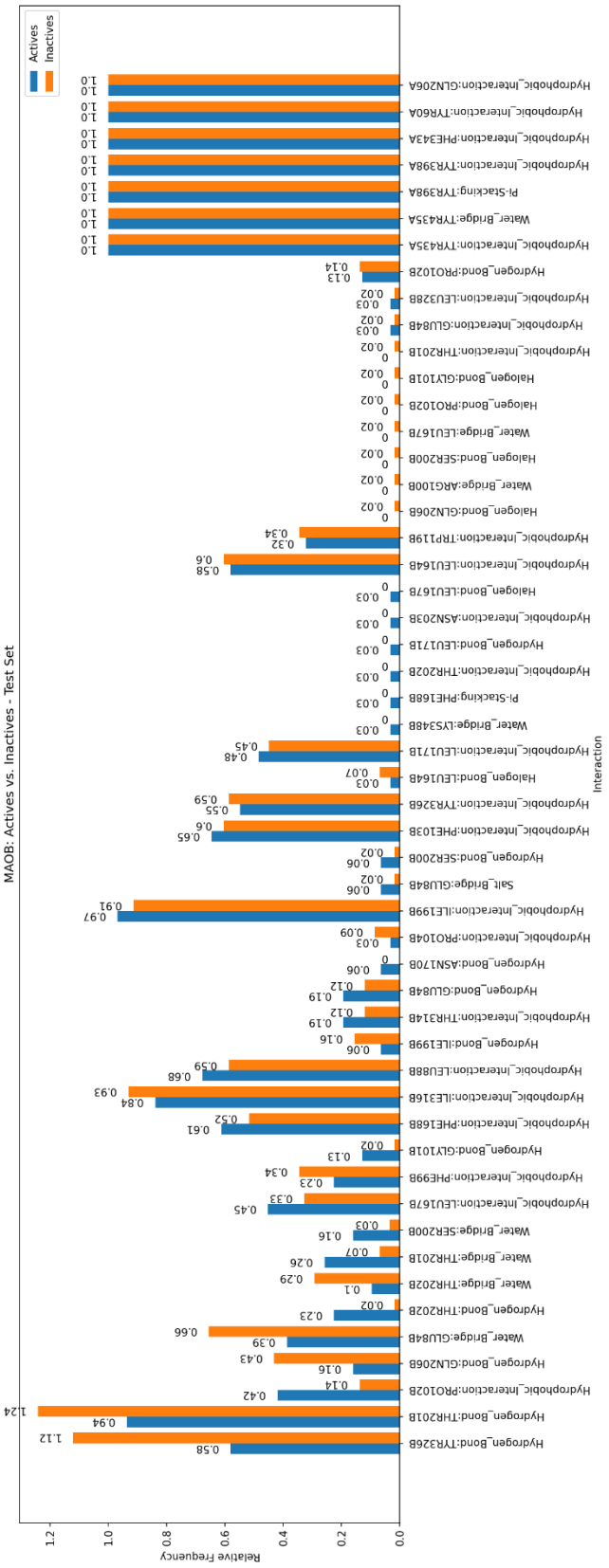


Fig. 40: Distribution of interaction frequencies of active and inactive MAOB ligands in the test partition in comparison.

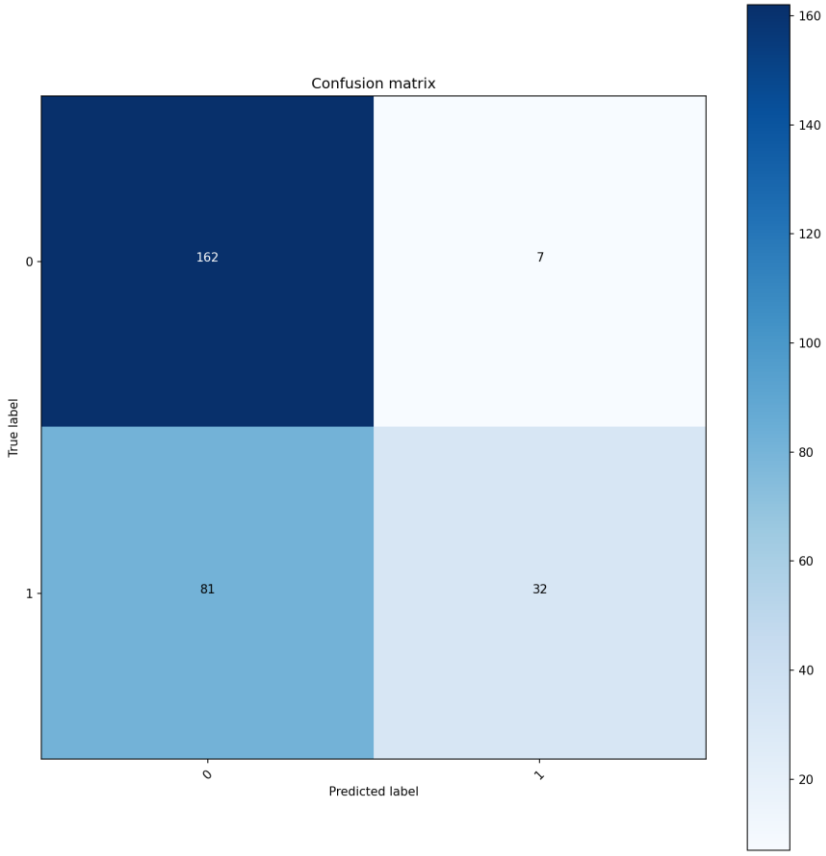


Fig. 41: Confusion matrix of the best-on-validation scoring strategy on the training data.

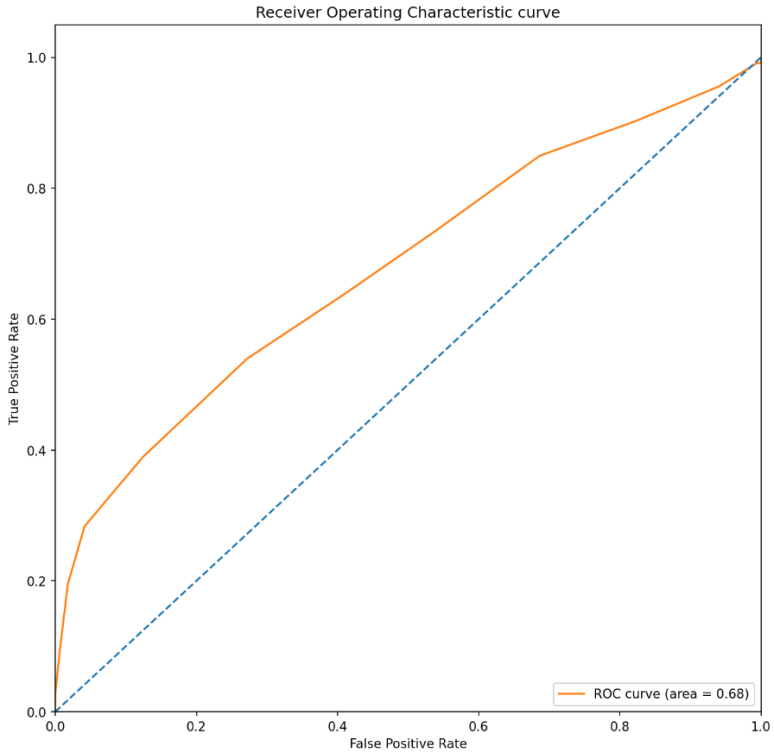


Fig. 42: ROC curve of the best-on-validation scoring strategy on the training data.

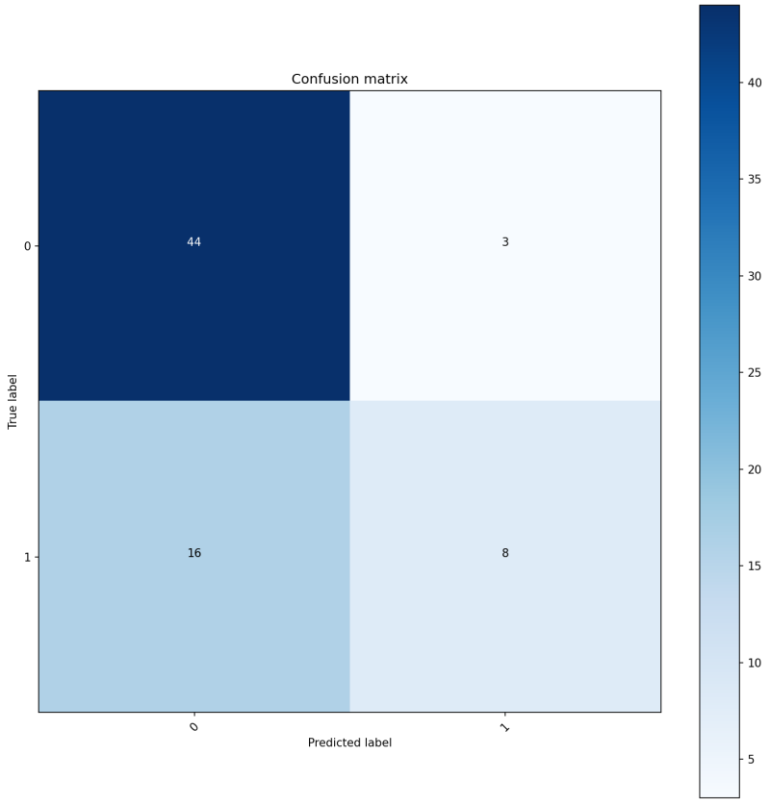


Fig. 43: Confusion matrix of the best-on-validation scoring strategy on the validation data.

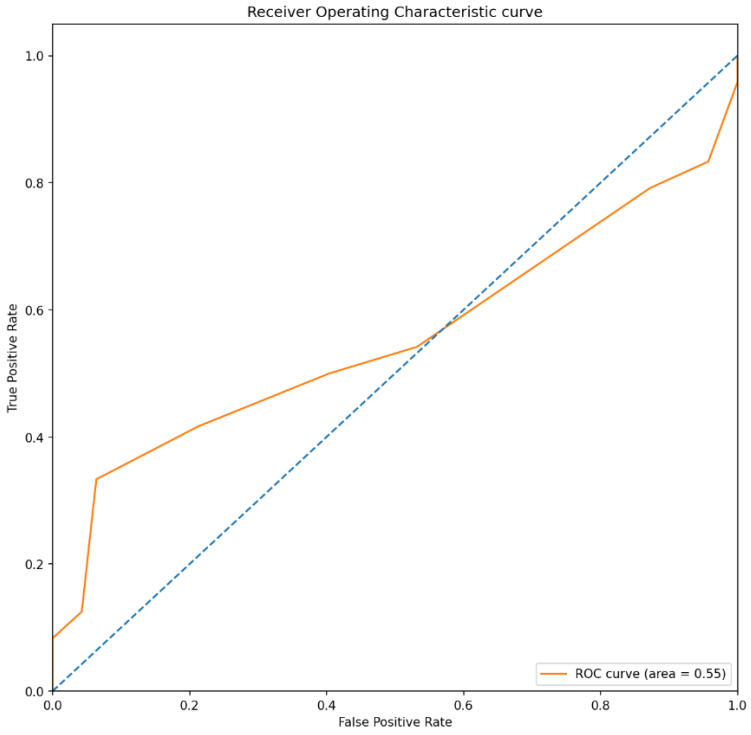


Fig. 44: ROC curve of the best-on-validation scoring strategy on the validation data.

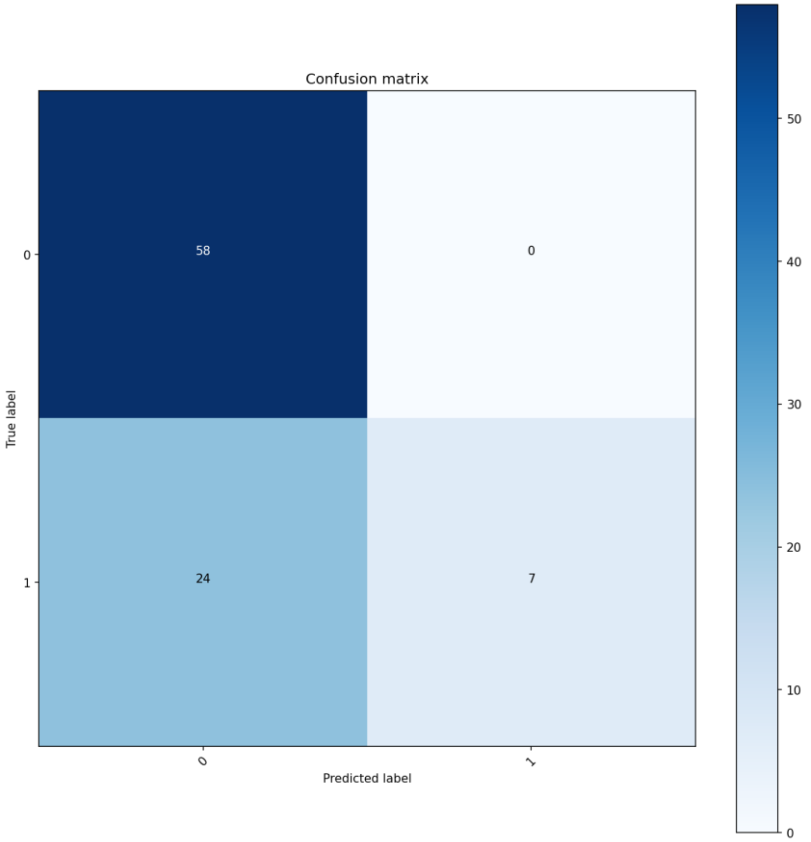


Fig. 45: Confusion matrix of the best-on-validation scoring strategy on the test data.

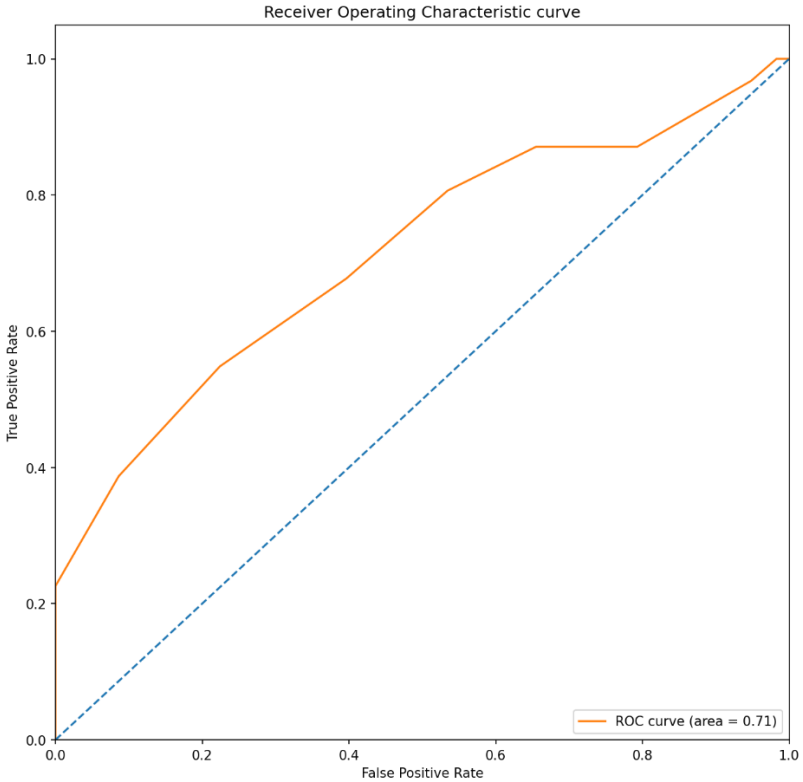


Fig. 46: ROC curve of the best-on-validation scoring strategy on the test data.

3.8 P38 mitogen-activated protein kinase 14

The bar chart depicted in **Fig. 47** shows the interactions and their respective frequencies that were extracted from the selected PDB structures for MAPK14. In total 159 protein-ligand complexes were analysed for that purpose and the results revealed that one of the two residues – that were known from literature to be important for binding – was present. Specifically LYS53A reacted with ligands in four different ways: In 79% of the structures via hydrophobic interactions, in 31% of the structures it formed a hydrogen bond, in 20% of the structures via pi-cation interactions, and in 11% of the structures it formed a water bridge. The other interaction know from literature – LYS152 – was either not present or not detected.

The top 5 interactions in terms of frequency were:

- Hydrogen bonding with MET109A in 91% of the structures.
- Hydrogen bonding with GLU71A in 84% of the structures.
- Hydrophobic interactions with LYS53A in 79% of the structures.
- Hydrophobic interactions with THR106A in 71% of the structures.
- Hydrophobic interactions with LEU75A in 68% of the structures.

3.9 Phosphodiesterase 5

For PDE5 the number of processed and analysed PDB structures was 25 and their interactions and corresponding frequencies are shown in **Fig. 48**. Residue GLN817A – which was mentioned in 2.1.8 – was present in all structures, forming a hydrogen bond in 112% of the cases – which means this residue sometimes forms more than one hydrogen bond in a single protein-ligand complex. In 8% of the structures GLN817A also reacts with the ligand via hydrophobic interactions.

The top 5 interactions in terms of frequency were:

- Metal complexation with ASP654A with a frequency of 128%.
- Hydrogen bonding with GLN817A with a frequency of 112%.
- Pi-stacking with PHE820A in 92% of the structures.
- Hydrophobic interactions with PHE820A in 72% of the structures.
- Metal complexation with HIS617A and HIS653A in 68% of the structures.

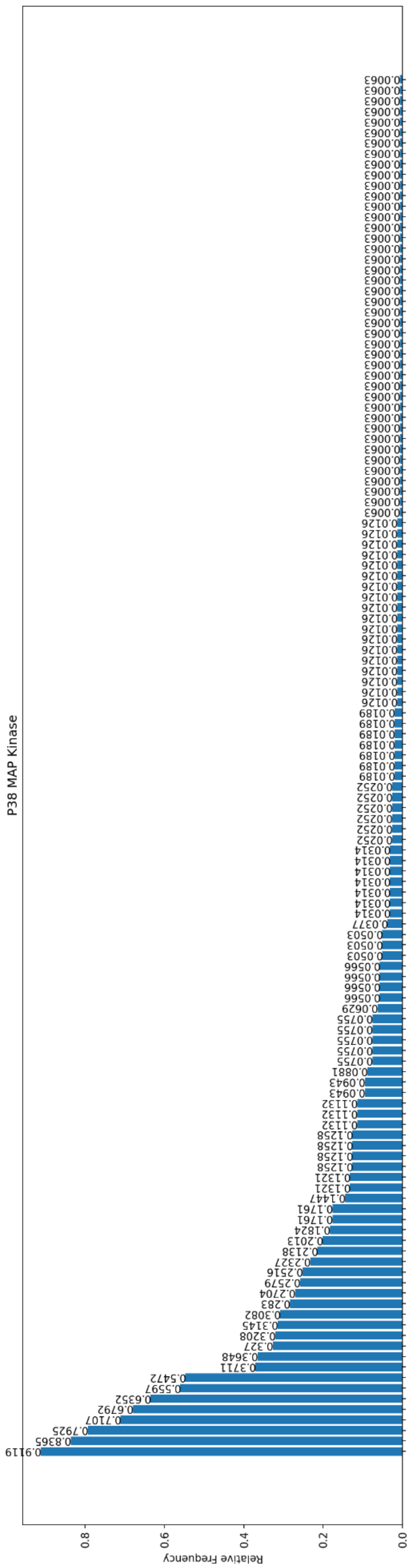


Fig. 47: Interaction frequencies of selected MAPK14 structures from the PDB.

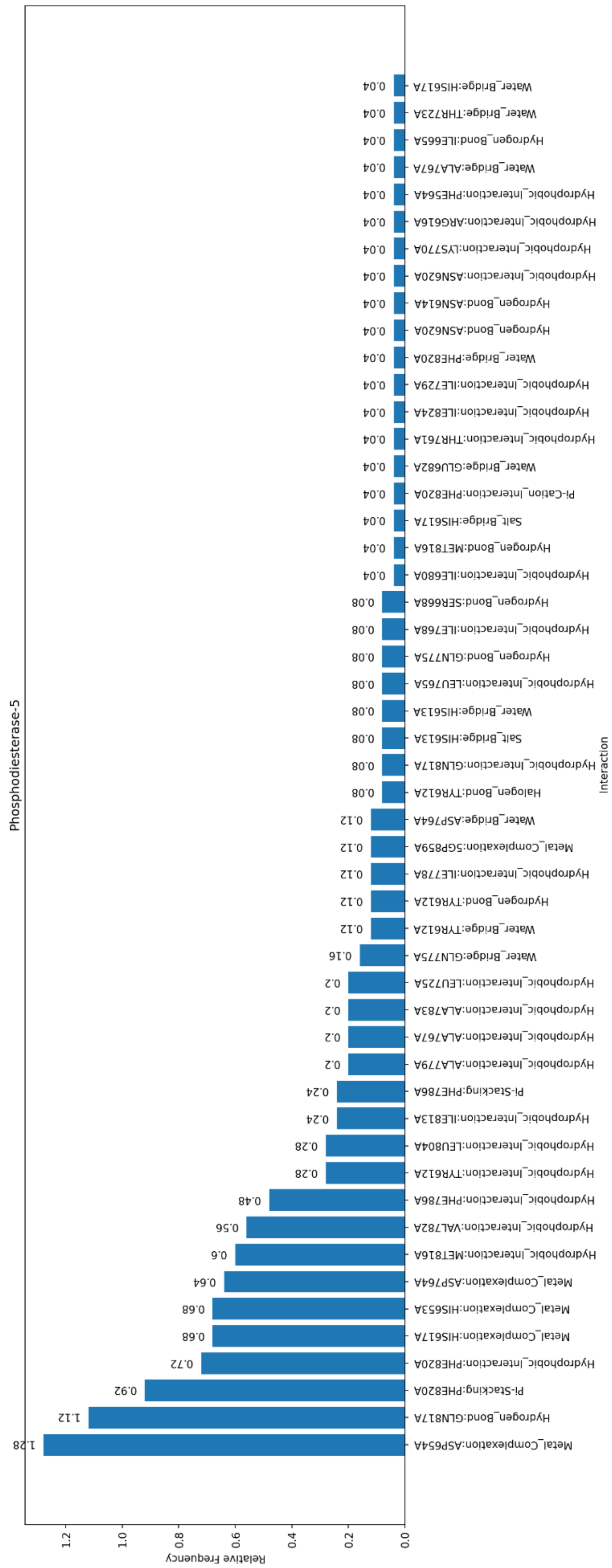


Fig. 48: Interaction frequencies of selected PDE5 structures from the PDB.

3.10 Protein-tyrosine phosphatase 1B

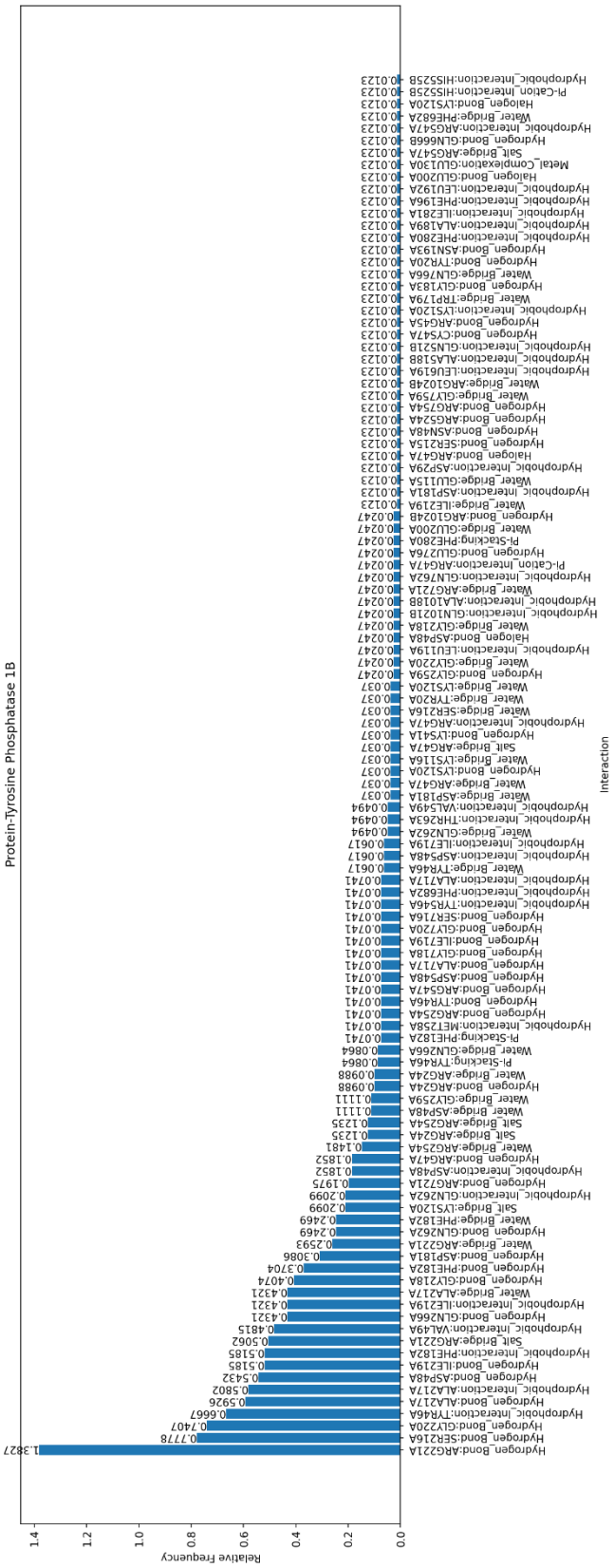


Fig. 49: Interaction frequencies of selected PTP1B structures from the PDB.

Fig. 49 shows interactions and their relative frequencies found in the 81 PTP1B structures that were selected from the PDB. Except for HIS214 and CYS125 all residues mentioned in 2.1.9 are present and detected in the selection. The following interactions occur with these residues:

- SER216A: The serine residue forms a hydrogen bond with the ligand in 78% of the structures. In 4% of the structures the hydrogen bond is water mediated.
- ALA217A: Three different binding modes are observed, namely hydrogen bonding in 59% of the structures, hydrophobic interactions in 58% of the structures, and water bridges in 43% of the structures.
- GLY218A: This residue binds via hydrogen bonding in 41% of the cases, in 2% of the structures it binds via water bridges.
- ILE219A: In 52% of the structures a hydrogen bond is formed with ILE219A and hydrophobic interactions with ILE219A occur in 43% of the structures.
- GLY220A: This residue interacts via hydrogen bonding in 74% of the structures and via water bridges in 3% of the structures.
- ARG221A: Hydrogen bonding with ARG221A is the most frequent interaction in this selection of structures of PTP1B with a frequency of 138%. Alternatively ARG221A binds via salt bridging in 51% of the structures or via water bridging in 26% of the structures.

The top 5 interactions in terms of frequency were:

- Hydrogen bonding with ARG221A with a frequency of 138%.
- Hydrogen bonding with SER216A with a frequency of 78%.
- Hydrogen bonding with GLY220A with a frequency of 74%.
- Hydrophobic interactions with TYR46A in 67% of the structures.
- Hydrogen bonding with ALA217A in 59% of the structures.

3.11 Soluble epoxide hydrolase

SEH was the fifth and final target for both interaction frequency analysis and scoring. The results of the frequency analysis are described below while scoring results will be depicted right after.

3.11.1 Interaction frequencies

Altogether 83 of the selected SEH structures from the PDB were analysed and their interactions with corresponding frequencies are shown in **Fig. 50**. All of the interactions known to be involved in binding have been picked up by PLIP/PIA:

- ASP335A: Three different binding modes are observed with the residue, in 42% of the structures a hydrogen bond is formed, in 4% of the structures the residue is involved in a salt bridge, and in 1% of the structures a water bridge is present.
- TRP336A: In 39% of the structures TRP336A interacts via pi-stacking and in 29% of the structures it is involved in hydrophobic interactions.
- TYR383A: This tyrosine is involved in four different interaction types, in 49% of the structures it takes part in hydrophobic interactions, in 43% of the structures it forms a hydrogen bond, in 4% of the structures it interacts via pi-stacking, and in 2% of the structures it forms a water bridge.
- TYR466A: In 34% of the structures hydrogen bonding with TYR466A is detected, the residue is also involved in hydrophobic interactions in 14% of the structures, in water bridges in 5% of the structures, in halogen bonds in 4% of the structures, and in pi-stacking in 1% of the structures.
- LEU499A: Hydrophobic interactions occur in 20% of the structures and hydrogen bonds form in 2% of the structures.
- HIS524A: The histidine residue can partake in almost all interaction types, most frequently at a rate of 36% it is involved in pi-stacking, in 19% of the structures it shows hydrophobic interactions, in 11% of the structures it forms a hydrogen bond, in also 11% of the structures it appears in a water bridge, in 2% of the structures pi-cation interactions occur, and in 1% of the structures it is part of a salt bridge.

The top 5 interactions in terms of frequency were:

- Hydrophobic interactions with TYR383A in 49% of the structures.
- Hydrophobic interactions with LEU408A in 47% of the structures.

- Hydrophobic interactions with TRP525A in 45% of the structures.
- Hydrogen bonding with TYR383A in 43% of the structures.
- Hydrogen bonding with ASP335A in 42% of the structures.

3.11.2 Scoring

Soluble epoxide hydrolase was the only target where compounds for scoring were not taken from the DUD-E, instead a custom SEH dataset consisting of 236 molecules was used. Of the 236 compounds 58 were active and 178 inactive, leading to a baseline classification accuracy of 75.4%. The molecules were randomly distributed into a training, validation and test dataset and docked into PDB structure 6HGV. Subsequently the 2360 poses were analysed and the best pose for each ligand was scored with PIA. The best-on-validation scoring strategy was strategy +. Moreover, the cut-off values for each strategy were 7 for strategy +, 8 for strategy ++, 6 for strategy +-, and 8 for strategy ++--.

Results on the training partition:

The interaction frequencies of active and inactive ligands in the training partition can be seen in **Fig. 51**. The best-on-validation scoring strategy achieved a prediction accuracy of 78% on the training dataset and a full overview of performance metrics of all scoring functions can be viewed in **Table 13**. The confusion matrix and ROC curve of the best-on-validation strategy calculated from the training data are plotted in **Fig. 54** and **Fig. 55**.

Table 13: Performance metrics for all scoring strategies evaluated on the training partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.78	0.108	0.769	0.6	2.308	60
++	0.76	0.108	0.729	0.556	2.137	55.556
+-	0.8	0.081	0.769	0.667	2.564	66.667
++--	0.753	0.054	0.728	0.571	2.198	57.143

Results on the validation partition:

In **Fig. 52** the distribution of interaction frequencies of active and inactive molecules in the validation partition is displayed. The best-on-validation scoring strategy classified 86.8% of the ligands in the validation partition correctly as active or inactive and the corresponding confusion matrix and ROC curve of these predictions are shown in **Fig. 56** and **Fig. 57**. A complete list of performance metrics of all scoring strategies applied to the validation data is depicted in **Table 14**.

Table 14: Performance metrics for all scoring strategies evaluated on the validation partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.868	0.067	0.794	0.714	3.393	71.429
++	0.842	0.067	0.810	0.667	3.167	66.667
+-	0.789	0.1	0.694	0.5	2.375	50
++--	0.789	0.067	0.652	0.5	2.375	50

Results on the test partition:

The interaction frequencies of active and inactive SEH ligands in the test partition are shown in the grouped bar chart displayed in **Fig. 53**. The best-on-validation scoring strategy achieved a classification accuracy of 77.1% on the test partition of the data. The resulting confusion matrix and ROC curve of this strategy can be seen in **Fig. 58** and **Fig. 59**. An exhaustive list of calculated performance metrics for all scoring strategies applied to the training data is shown in **Table 15**.

Table 15: Performance metrics for all scoring strategies evaluated on the test partition.

STRATEGY	ACC	FPR	AUC	YA	EF	REF
+	0.771	0.162	0.814	0.5	2.182	54.545
++	0.833	0.081	0.774	0.667	2.909	66.667
+-	0.771	0.162	0.792	0.5	2.182	54.545
++--	0.813	0.081	0.784	0.625	2.727	62.5

3.12 Computational performance of PIA

Extraction of interactions and frequencies in a standard workflow with a single PDB structure and ligands in SDF format as input takes PIA about 1-2 hours/1000 ligands. The computation time is strongly dependent on the structural complexity of the protein and ligand as well as the performance of PLIP. During structure preparation the ligands of the SDF file will be written into PDB files and additionally protonated by PLIP, as a result about 3-6 GB of files/1000 ligands will be generated. This has to be kept in mind especially when evaluating large datasets, for example when evaluating and scoring DPP4 more than 20 000 poses were analysed which created roughly 100 GB of data. Once the interactions and their corresponding frequencies are extracted and saved however, scoring works almost instantly since it is just a sequence of enumerative and additive operations. The only time consuming process in scoring is the determination of optimal threshold parameters for feature selection and cut-off values which usually takes between 30-60 minutes per target.

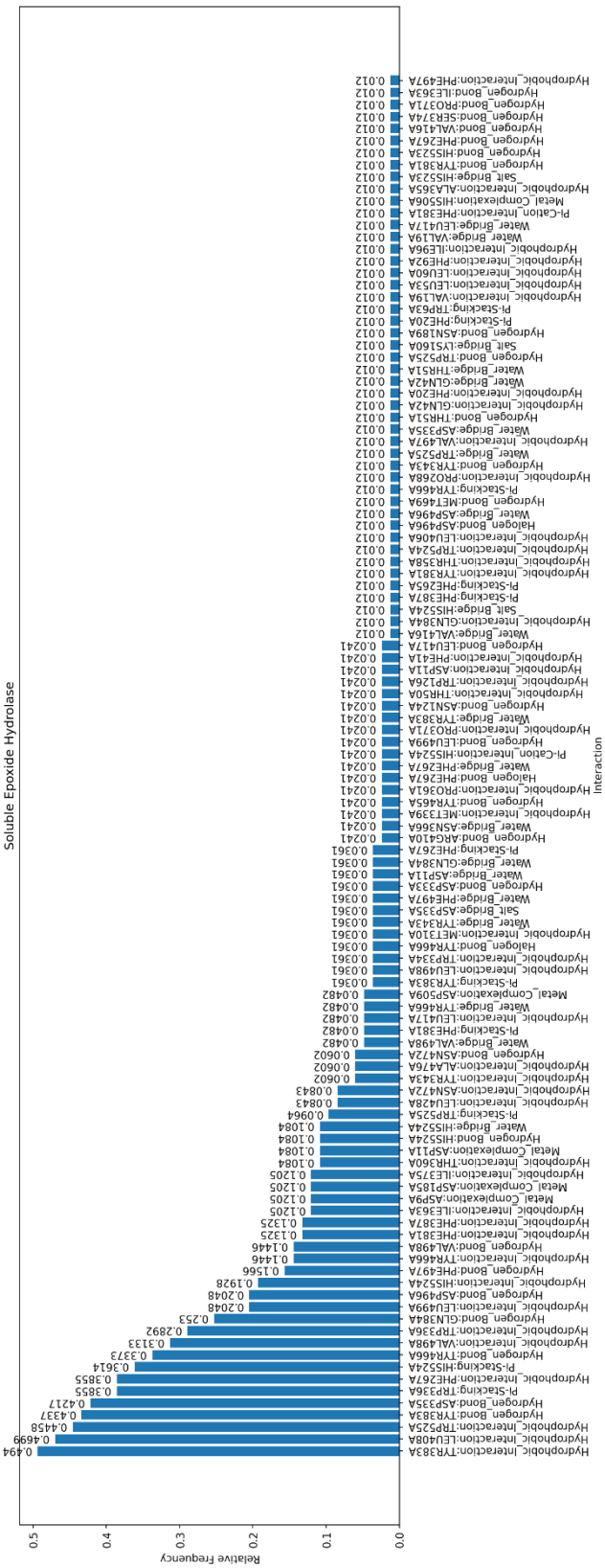


Fig. 50: Interaction frequencies of selected SEH structures from the PDB.

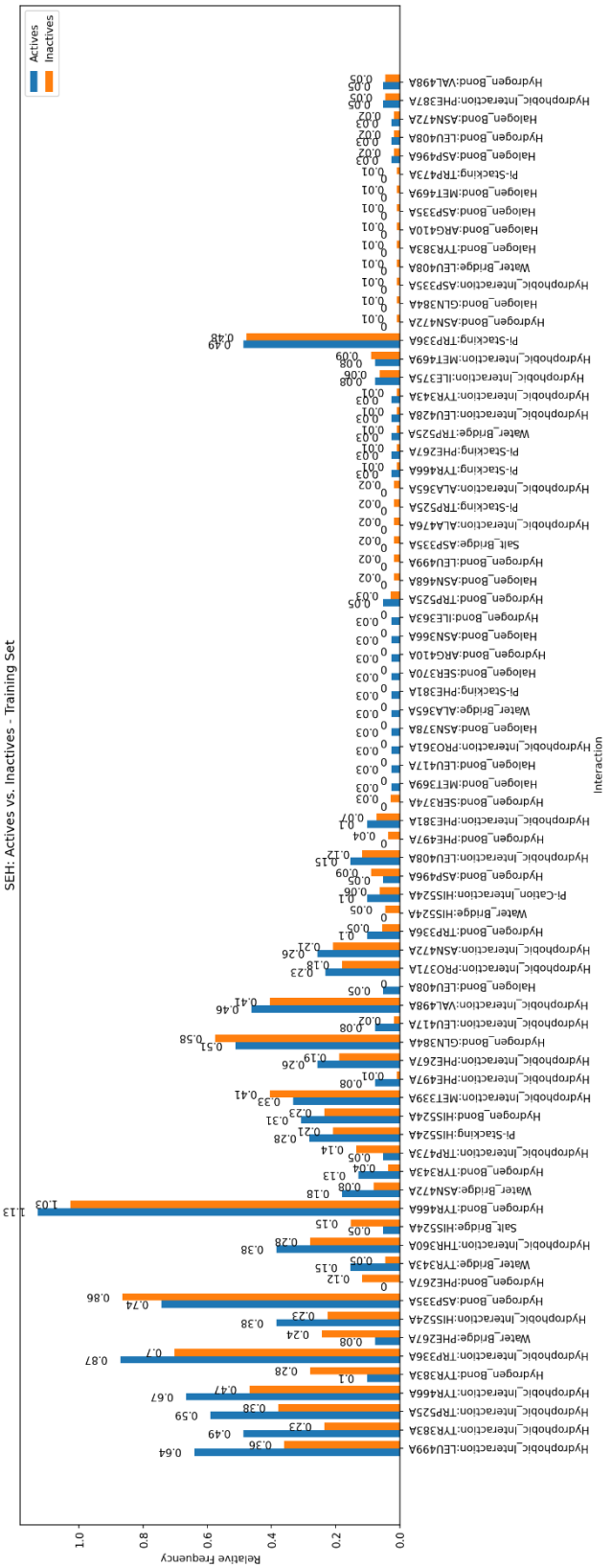


Fig. 51: Distribution of interaction frequencies of active and inactive SEH ligands in the training partition in comparison.

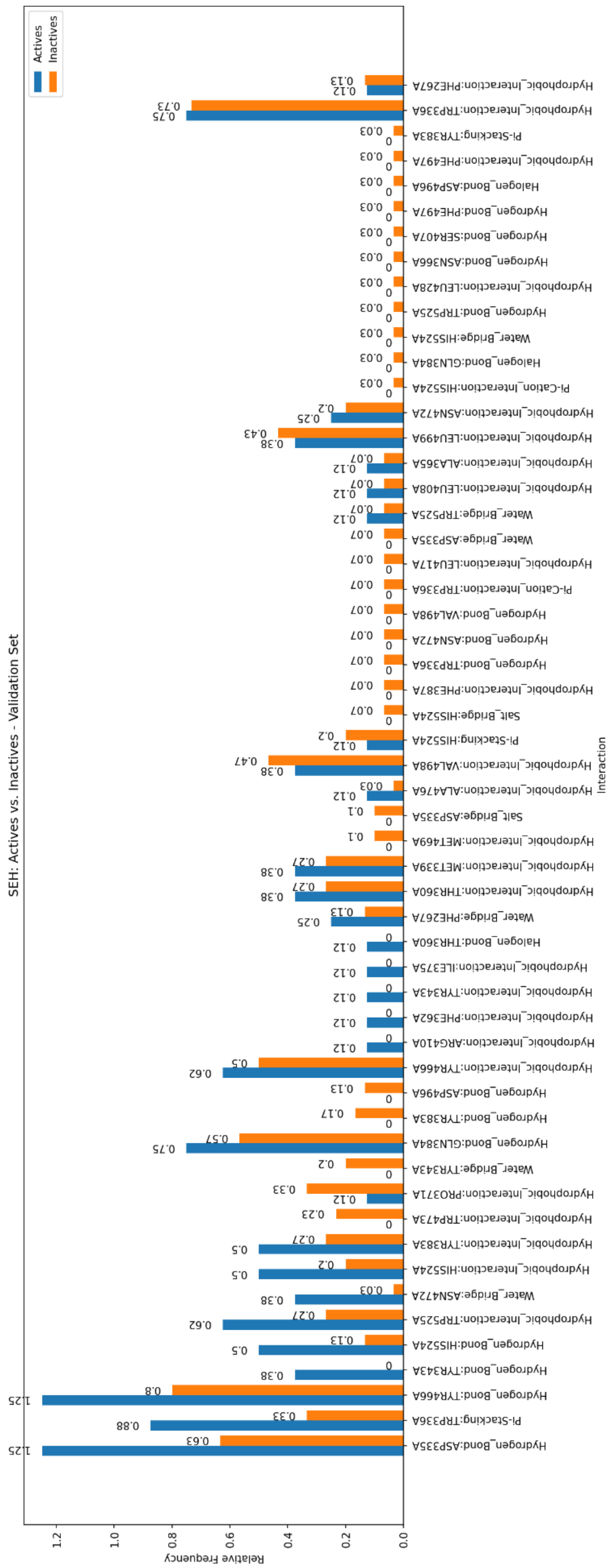


Fig. 52: Distribution of interaction frequencies of active and inactive SEH ligands in the validation partition in comparison.

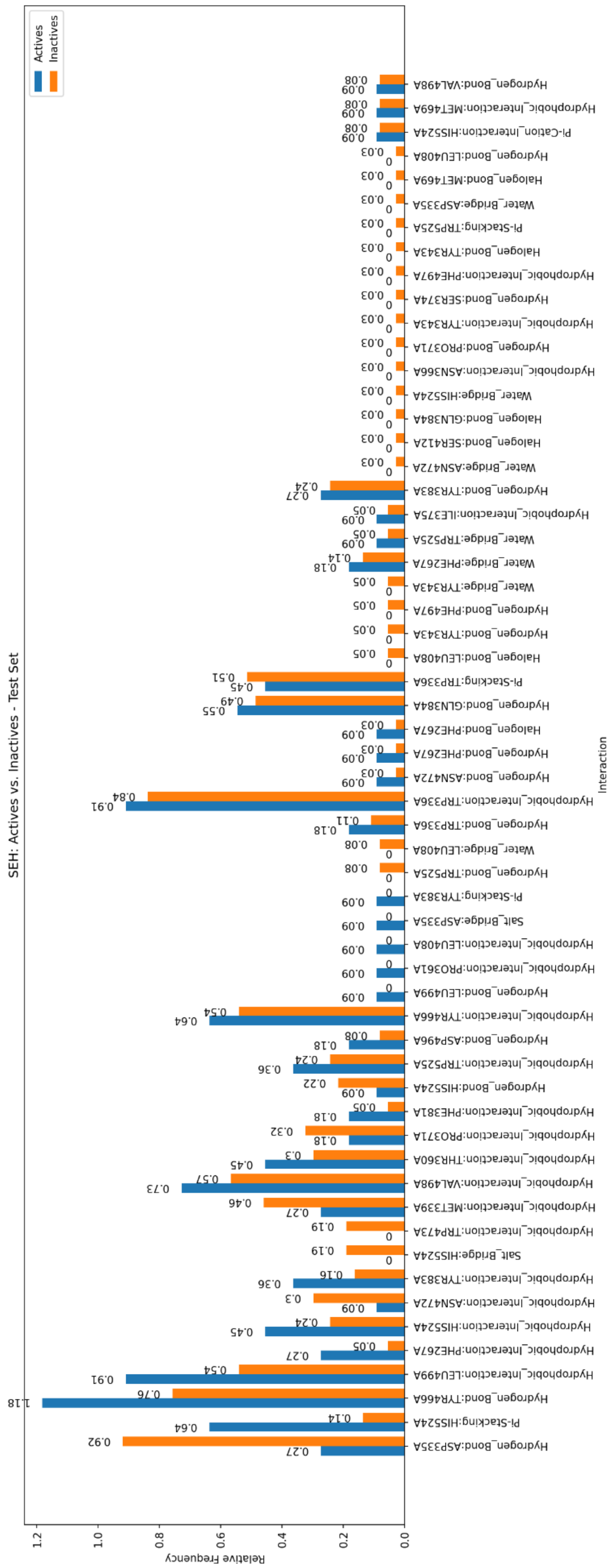


Fig. 53: Distribution of interaction frequencies of active and inactive SEH ligands in the test partition in comparison.

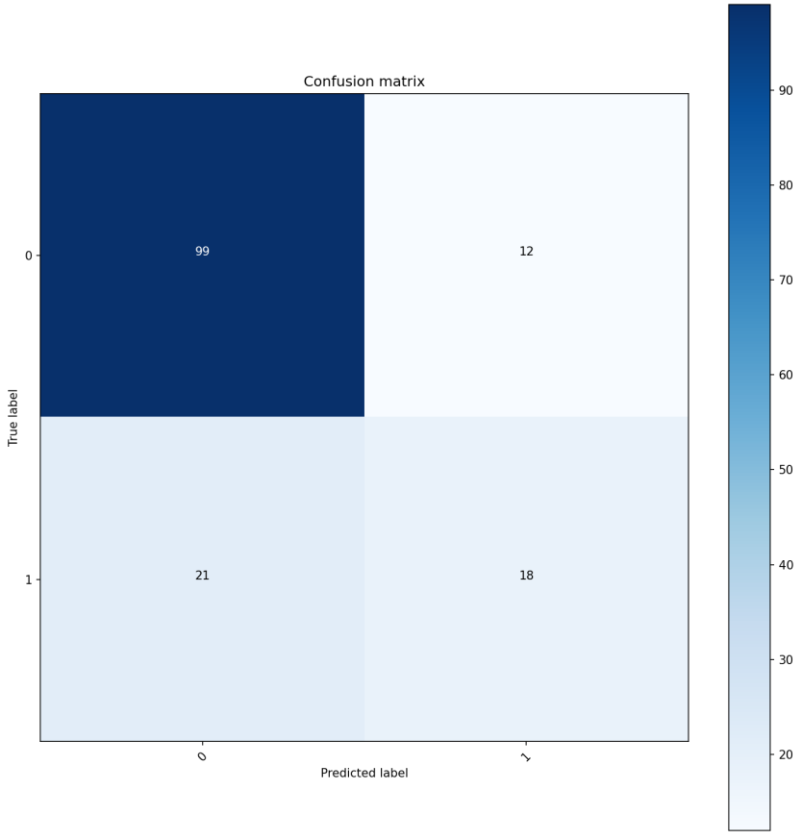


Fig. 54: Confusion matrix of the best-on-validation scoring strategy on the training data.

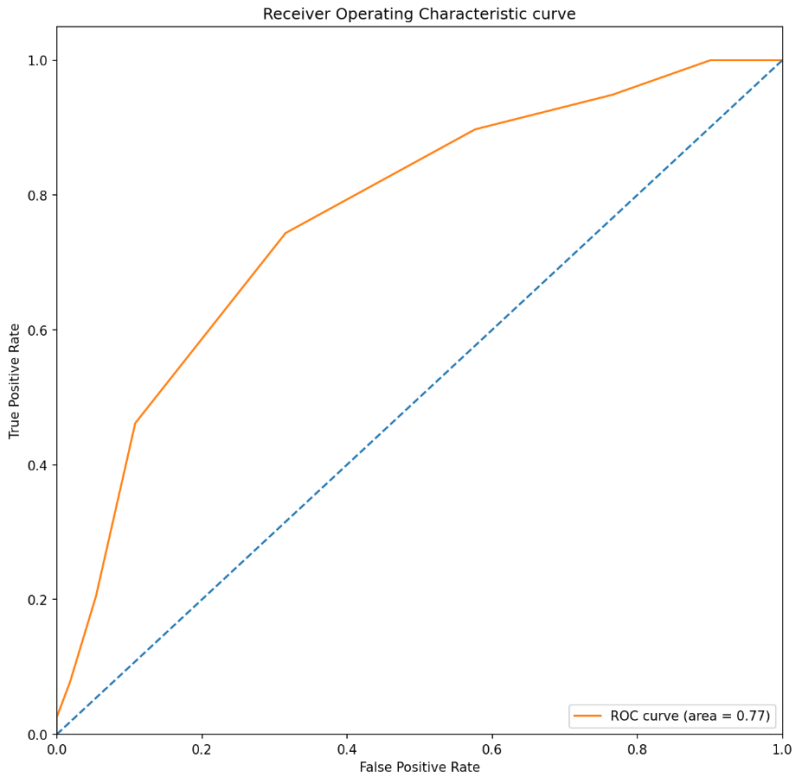


Fig. 55: ROC curve of the best-on-validation scoring strategy on the training data.

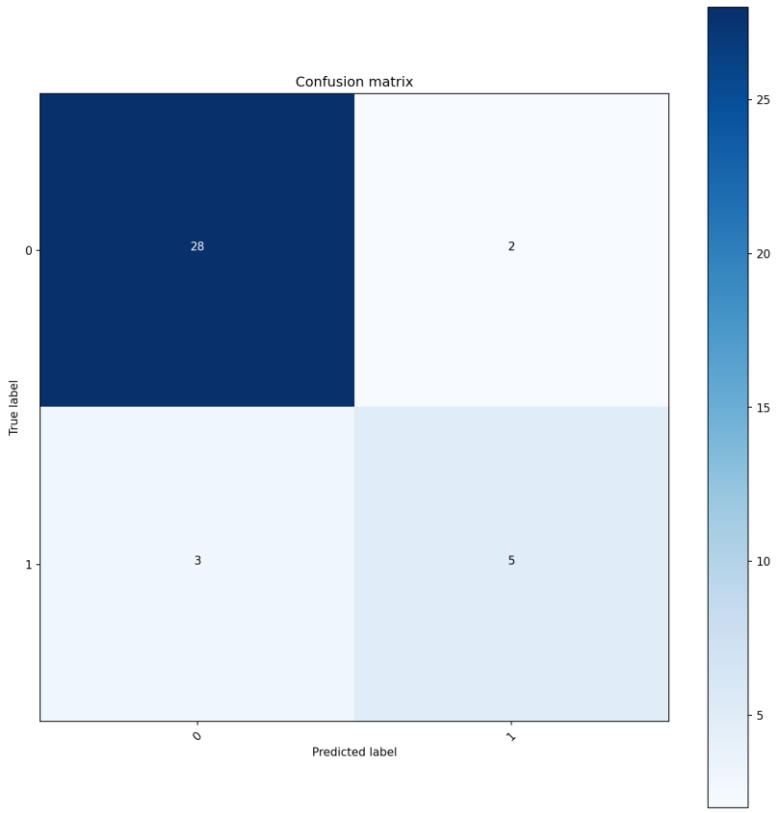


Fig. 56: Confusion matrix of the best-on-validation scoring strategy on the validation data.

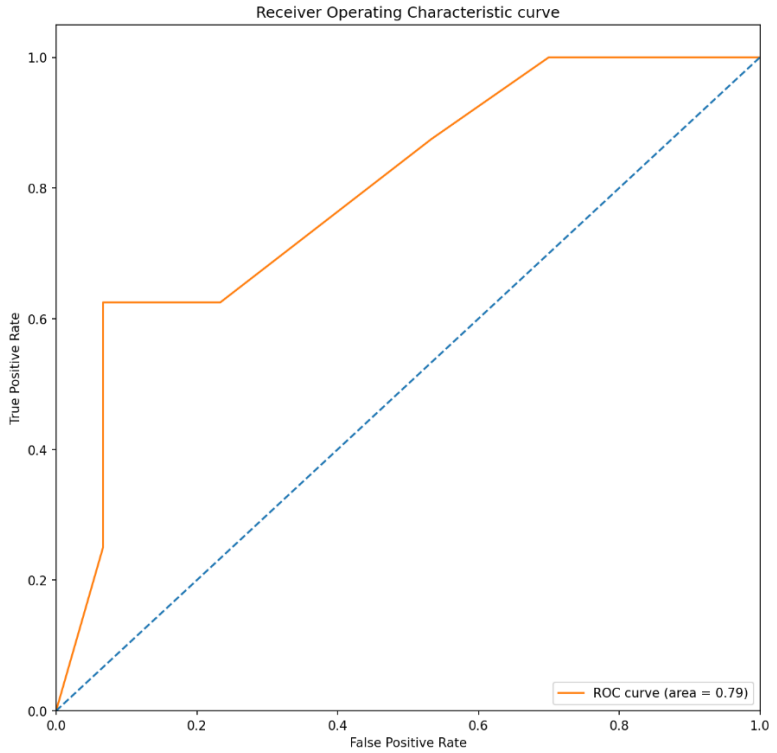


Fig. 57: ROC curve of the best-on-validation scoring strategy on the validation data.

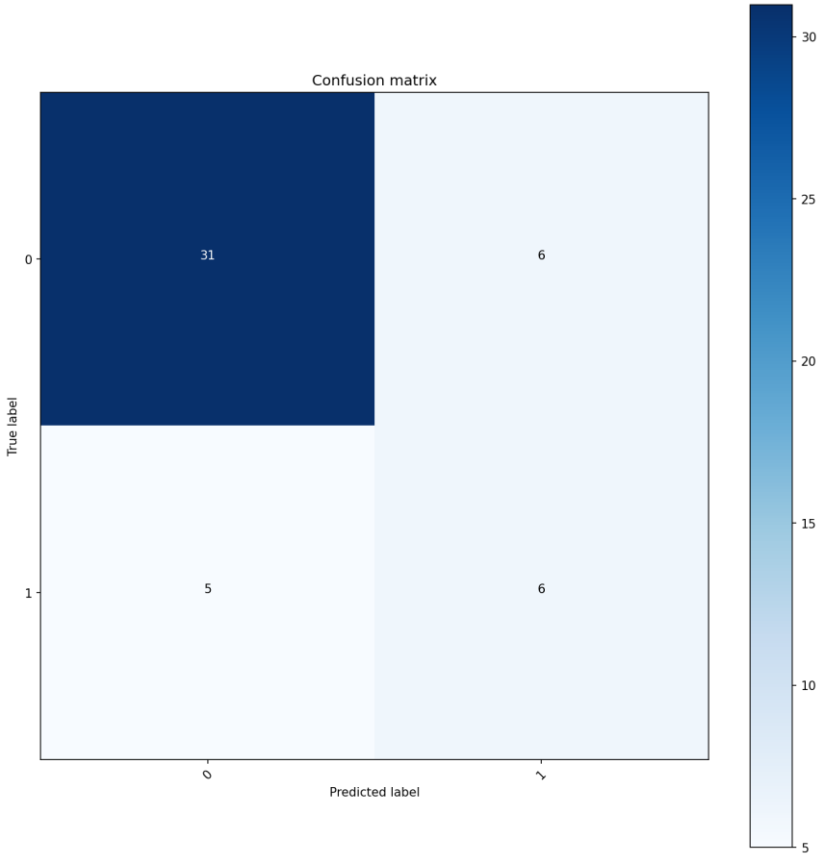


Fig. 58: Confusion matrix of the best-on-validation scoring strategy on the test data.

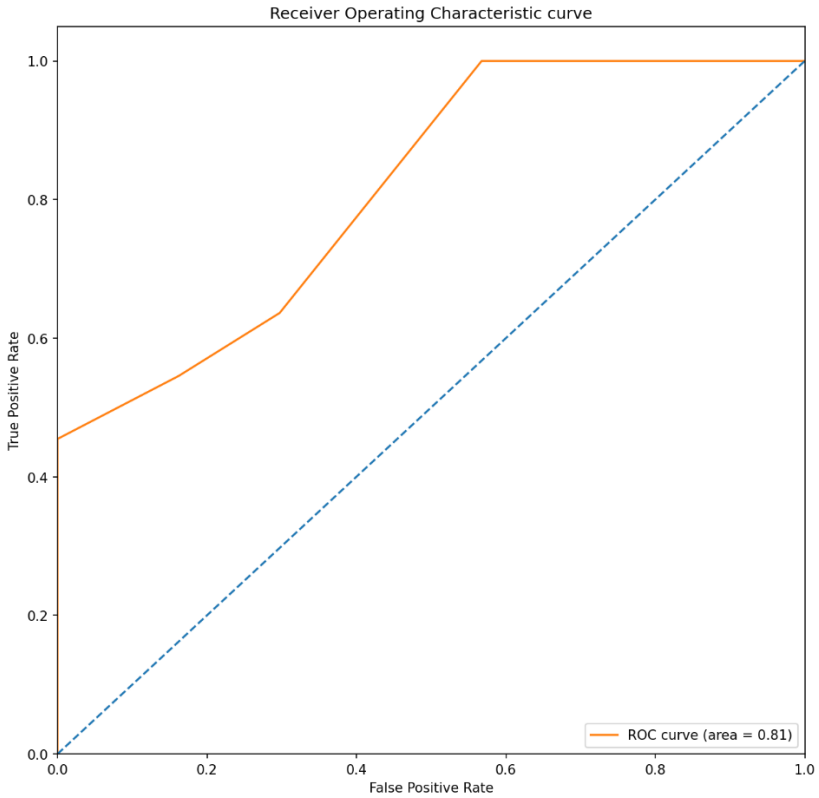


Fig. 59: ROC curve of the best-on-validation scoring strategy on the test data.

4. Discussion

Two goals were defined for this thesis, the identification of protein-ligand interactions which are important for binding, and the development of a novel scoring function based on the interaction frequencies of active and inactive compounds.

The first goal has been addressed by developing a workflow that extracts the interactions and their corresponding frequencies from a set of known PDB structures of a target. The interactions are then ranked based on their frequency in decreasing order. The results showed that interactions that were already known to be of importance for binding also occurred in the selection extracted by PIA for almost all targets – ACHE being the only exception. Moreover, these interactions often were among the top ranked interactions of PIA, implying a relationship between interaction frequency and binding significance. This could be especially of interest for targets where binding interactions are not yet known or fully understood but structural data is already available.

The second goal was built upon the first one and used the extracted interactions and frequencies of active and inactive ligands for scoring. In total four different scoring functions were designed based on the interaction frequencies and they all performed reasonably well on the five evaluated targets. The second of the four scoring functions – which was defined as the difference between the number of positive interactions and the number of negative interactions in a protein-ligand complex – was the most successful one, yielding the best-on-validation accuracy in three out of the five targets. Furthermore, the classification accuracy of the best-on-validation scoring strategy exceeded the baseline accuracy in four of the five targets and resulted in an enrichment in all targets. False positive rates were usually between 0 and 33%. However, it should be noted that performance metrics often fluctuated and were not necessarily consistent across training, validation and test partition. Therefore interaction-based scoring and classification should definitely be seen as a supporting tool to existing VS approaches rather than a standalone solution.

Most recently scoring with PIA has also been applied outside of this thesis' research for evaluation of docking results of vitamin E. Despite returning good results when the docked ligands where compared to an established set of decoys, it also showed a potential weakness of PIA: The discrimination of weak actives from actives or weak actives from inactives is hardly possible using the interaction frequency approach. Although this is to be expected since PIA is purely based on the interacting residues, the interaction types and their frequencies without weighting interactions or accounting for any binding energies, it is a remark that should be especially highlighted.

Weighting of interactions is also an aspect that could be considered for future research building upon this thesis. For example, hydrophobic interactions are comparably weaker than hydrogen bonds yet they contribute equally to the score in PIA. Introducing weighting coefficients for the specific interaction types could potentially further improve results. Another aspect that could be looked upon is how to deal with cofactors. Currently PIA completely ignores any co-

crystallized cofactors since PLIP is not able to detect and characterize cofactor-ligand interactions, however, these interactions could further refine the score. Furthermore, going in the direction of AI and ML could be another way worthwhile of exploring. Development of more sophisticated scoring functions using ML was already considered for this thesis but fell short due to time constraints. Nevertheless, explainable AI approaches could possibly come up with more specialized scoring functions that may also give deeper insight on the importance of specific interactions in protein-ligand complex.

Last but not least the technical implementation of PIA is something that could still be improved in the future. Rewriting molecules from SDF into PDB format is a necessary step because PLIP can only deal with PDB structures. Needless to say that this process is far from optimal since it not only takes a lot of time but also consumes a lot of free disk space. Tighter integration of PLIP into PIA that is not reliant on creating PDB structures and therefore allows to skip this step would be an option to make computation of interaction frequencies a lot faster. Another concern would be the pre-processing of structures and merging of protein and ligand coordinates. This task is currently done using custom self-implemented functions since there are no state-of-the-art solutions available for python, and although the implementation works well for the established workflows, merging of more complex structures – for example a single protein with multiple ligands or small molecules – would possibly pose a problem. Finally, the process of extracting interaction frequencies could also be further optimized: Currently the best pose (if multiple poses are detected) is analysed twice due to the underlying data structures and how the function is designed, however, rewriting the function to re-use the data from the previous analysis instead of re-calculating would definitely be possible. The source code of PIA is publicly available on GitHub via https://github.com/michabirklbauer/protein_docking and anyone is welcome to contribute to the project.

5. Conclusion

Molecular docking is an important tool in virtual screening for the discovery and design of new active agents for drug usage. The docking process is influenced by how well molecules fit in the binding site and which interactions occur between the protein and the ligand. Detection of these interactions can be automated with tools like PLIP. However, identification and assessment of the importance of the different interactions in a protein-ligand complex is still a manual task that requires additional experimental data or domain knowledge about the target. The goals of this thesis were twofold: Firstly, to automatically identify those interactions that have a significant influence on ligand binding, and secondly, to develop a novel scoring function which is able to discriminate active molecules from inactive ones if possible. The underlying data basis were selected targets of the DUD-E and available structures from the PDB. Specifically 11 targets were analysed: HSD11B1, ACHE, FXA, COX1, COX2, DPP4, MAOB, MAPK14, PDE5A, PTP1B and SEH. PLIP was used to extract interactions present in a protein-ligand complex and the respective interaction's frequency was measured across all target structures. Cofactors were excluded from the analysis and hydrophobic interactions were only counted once per residue. Additionally, when analysing docking poses only the pose that had the most interactions contributed to the calculation. Furthermore, four different scoring functions that are based on the differences in frequencies between active and inactive compounds were established and their performance was assessed on an independent test partition containing unseen ligands. The results show that interactions which are known from literature to be important for ligand binding are found for all targets except ACHE, in many cases among the top ranked interactions in terms of frequency. This behaviour implies a relationship between interaction frequency and the interaction's significance in ligand binding. Interaction-frequency-based scoring was tested in five targets and performed above baseline accuracy in four of the five targets. In all targets scoring led to an enrichment of active compounds and false positive rates fluctuated between 0 and 33%. Interaction frequency analysis and interaction-frequency-based scoring could therefore be used as supporting tools in virtual screening to further enhance results.

References (Figures)

Fig. 1 and **Fig. 2** were designed with Zen Flowchart (Zen Flowchart ™, Zen Flowchart Inc., https://www.zenflowchart.com/).

Fig. 3 – **Fig. 59** were created with Matplotlib (Hunter, 2007).

References

Andreini, C., Banci, L., Bertini, I. & Rosato, A. (2006). Zinc through the Three Domains of Life. *Journal of Proteome Research, 5*(11), 3173 – 3178. doi:10.1021/pr0603699

Anslyn, E. V. & Dougherty, D. A. (2006). *Modern Physical Organic Chemistry*. Sausalito, USA: University Science Books.

Arunan, E., Desiraju, G. R., Klein, R. A., Sadlej, J., Scheiner, S., Alkorta, I., Clary, D. C., Crabtree, R. H., Dannenberg, J. J., Hobza, P., Kjaergaard, H. G., Legon, A. C., Mennucci B. & Nesbitt, D. J. (2011). Definition of the hydrogen bond (IUPAC Recommendations 2011). *Pure and Applied Chemistry, 83*(8), 1637 – 1641. doi:10.1351/PAC-REC-10-01-02

Berger, J. P., SinhaRoy, R., Pocai, A., Kelly, T. M., Scapin, G., Gao, Y.-D., Pryor, K. A. D., Wu, J. K., Eiermann, G. J., Xu, S. S., Zhang, X., Tatosian, D. A., Weber, A. E., Thornberry, N. A. & Carr, R. D. (2017). A comparative study of the binding properties, dipeptidyl peptidase-4 (DPP-4) inhibitory activity and glucose-lowering efficacy of the DPP-4 inhibitors alogliptin, linagliptin, saxagliptin, sitagliptin and vildagliptin in mice. *Endocrinology, Diabetes & Metabolism, 1*(1), 1 – 8. doi:10.1002/edm2.2

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235 – 242. doi:10.1093/nar/28.1.235

Biedermann, F. & Schneider, H.-J. (2016). Experimental Binding Energies in Supramolecular Complexes. *Chemical Reviews, 116*(9), 5216 – 5300. doi:10.1021/acs.chemrev.5b0058

Bischoff, E. (2004). Potency, selectivity, and consequences of nonselectivity of PDE inhibition. *International Journal of Impotence Research, 16*(1), 11 – 14. doi:10.1038/sj.ijir.3901208

Bonivento, D., Milczek, E. M., McDonald, G. R., Binda, C., Holt, A., Edmondson, D. E. & Mattevi, A. (2010). Tranylcypromine-inhibited human monoamine oxidase B in complex with 2- (2-benzofuranyl)-2-imidazoline. *The Protein Data Bank, 2XCG*. doi:10.2210/pdb2xcg/pdb

Card, G. L., England, B. P., Suzuki, Y., Fong, D., Powell, B., Lee, B., Luu, C., Tabrizizad, M., Gilette, S., Ibrahim, P. N., Artis, D. R., Bollag, G., Milburn, M. V., Kim, S.-H., Schlessinger, J. & Zhang, K. Y. J. (2004). Structural Basis for the Activity of Drugs that Inhibit Phosphodiesterases. *Structure, 12*(12), 2233 – 2247. doi:10.1016/j.str.2004.10.004

Chen X. (2006). Biochemical Properties of Recombinant Prolyl Dipeptidases DPP-IV and DPP8. In: Lendeckel U., Reinhold D. & Bank U. (eds) *Dipeptidyl Aminopeptidases. Advances in Experimental Medicine and Biology, vol 575*. Boston, USA: Springer. doi:10.1007/0-387-32824-6_3

Cheung, J., Rudolph, M., Burshteyn, F., Cassidy, M., Gary, E., Love, J., Height, J., & Franklin, M. (2012). Crystal Structure of Recombinant Human Acetylcholinesterase in Complex with Donepezil. *The Protein Data Bank, 4EY7*. doi:10.2210/pdb4ey7/pdb

Cheung, J., Rudolph, M. J., Burshteyn, F., Cassidy, M. S., Gary, E. N., Love, J., Franklin, M. C. & Height, J. J. (2012). Structures of Human Acetylcholinesterase in Complex with Pharmacologically Important Ligands. *Journal of Medicinal Chemistry, 55*(22), 10282 – 10286. doi:10.1021/jm300871x

Classen-Houben, D., Schuster, D., Da Cunha, T., Odermatt, A., Wolber, G., Jordis, U. & Kueenburg, B. (2009). Selective inhibition of 11β-hydroxysteroid dehydrogenase 1 by 18α-glycyrrhetinic acid but not 18β-glycyrrhetinic acid. *The Journal of Steroid Biochemistry and Molecular Biology, 113*(3-5), 248 – 252. doi:10.1016/j.jsbmb.2009.01.009

Combs, A. P. (2010). Recent Advances in the Discovery of Competitive Protein Tyrosine Phosphatase 1B Inhibitors for the Treatment of Diabetes, Obesity, and Cancer. *Journal of Medicinal Chemistry, 53*(6), 2333 – 2344. doi:10.1021/jm901090b

Desiraju, G. R., Ho, P. S., Kloo, L., Legon, A. C., Marquardt, R., Metrangolo, P., Politzer, P., Resnati, G. & Rissanen, K. (2013). Definition of the halogen bond (IUPAC Recommendations 2013). *Pure and Applied Chemistry, 85*(8), 1711 – 1713. doi:10.1351/PAC-REC-12-05-10

Dvir, H., Silman, I., Harel, M., Rosenberry, T. L. & Sussman, J. L. (2010). Acetylcholinesterase: From 3D structure to function. *Chemico-Biological Interactions, 187*(1-3), 10 – 22. doi:10.1016/j.cbi.2010.01.042

DrugBank - O76074. (2021). *cGMP-specific 3',5'-cyclic phosphodiesterase*. DrugBank. https://go.drugbank.com/polypeptides/O76074

DrugBank - P00742. (2021). *Coagulation factor X*. DrugBank. https://go.drugbank.com/polypeptides/P00742

Drugbank - P18031. (2021). *Tyrosine-protein phosphatase non-receptor type 1*. DrugBank. https://go.drugbank.com/polypeptides/P18031

DrugBank - P22303. (2021). *Acetylcholinesterase*. DrugBank. https://go.drugbank.com/polypeptides/P22303

DrugBank - P23219. (2021). *Prostaglandin G/H synthase 1*. DrugBank. https://go.drugbank.com/polypeptides/P23219

DrugBank - P27338. (2021). *Amine oxidase [flavin-containing] B*. DrugBank. https://go.drugbank.com/polypeptides/P27338

DrugBank - P27487. (2021). *Dipeptidyl peptidase 4*. DrugBank. https://go.drugbank.com/polypeptides/P27487

DrugBank - P28845. (2021). *Corticosteroid 11-beta-dehydrogenase isozyme 1*. DrugBank. https://go.drugbank.com/polypeptides/P28845

DrugBank - P34913. (2021). *Bifunctional epoxide hydrolase 2*. DrugBank. https://go.drugbank.com/polypeptides/P34913

DrugBank - P35354. (2021). *Prostaglandin G/H synthase 2*. DrugBank. https://go.drugbank.com/polypeptides/P35354

DrugBank - Q16539. (2021). *Mitogen-activated protein kinase 14*. DrugBank. https://go.drugbank.com/polypeptides/Q16539

Edmondson, D. E., Binda, C. & Mattevi, A. (2007). Structural insights into the mechanism of amine oxidation by monoamine oxidases A and B. *Archives of Biochemistry and Biophysics, 464*(2), 269 – 276. doi:10.1016/j.abb.2007.05.006

EMBL-EBI. (2011). *CoFactor - The organic enzyme cofactor database*. EMBL-EBI. http://www.ebi.ac.uk/thornton-srv/databases/CoFactor/

Friedrich, L., Cingolani, G., Ko, Y.-H., Iaselli, M., Miciaccia, M., Perrone, M. G., Neukirch, K., Bobinger, V., Merk, D., Hofstetter, R. K., Werz, O., Koeberle, A., Scilimati, A. & Schneider, G. (2021). Learning from Nature: From a Marine Natural Product to Synthetic Cyclooxygenase-1 Inhibitors by Automated De Novo Design. *Advanced Science, 2100832*(1), 1 – 12. doi:10.1002/advs.202100832

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. doi:10.1021/jm0306430

Goldstein, D. & Gabriel, T. (2005). Pathway to the Clinic: Inhibition of P38 MAP Kinase. A Review of Ten Chemotypes Selected for Development. *Current Topics in Medicinal Chemistry, 5*(10), 1017 – 1029. doi:10.2174/1568026054985939

Güner, O. F. (eds) (2000). *Pharmacophore Perception, Development, and Use in Drug Design*. San Diego, USA: International University Line.

Hartshorn, R. M., Hellwich, K.-H., Yerin, A., Damhus, T. & Hutton, A. T. (2015). Brief guide to the nomenclature of inorganic chemistry. *Pure and Applied Chemistry, 87*(9-10), 1039 – 1049. doi: 10.1515/pac-2014-0718

Havre, P. A., Abe, M., Urasaki, Y., Ohnuma, K., Morimoto, C. & Dang, N. H. (2008). The role of CD26/dipeptidyl peptidase IV in cancer. *Frontiers in Bioscience, 13*(5), 1634 – 1645. doi:10.2741/2787

Huang, N., Shoichet, B. K. & Irwin, J. J. (2006). Benchmarking Sets for Molecular Docking. *Journal of Medicinal Chemistry, 49*(23), 6789 – 6801. doi:10.1021/jm0608356

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90 – 95. doi:10.1109/MCSE.2007.55

IUPAC-IUB Joint Commission on Biochemical Nomenclature (1984). Nomenclature and Symbolism for Amino Acids and Peptides. *European Journal of Biochemistry, 138*(1), 9 – 37. doi:10.1111/j.1432-1033.1984.tb07877.x

Jain, A. N. (2003). Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *Journal of Medicinal Chemistry, 46*(4), 499 – 511. doi:10.1021/jm020406h

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology, 267*(3), 727 – 748. doi:10.1006/jmbi.1996.0897

Karami, L., Saboury, A. A., Rezaee, E. & Tabatabai, S. A. (2016). Investigation of the binding mode of 1, 3, 4-oxadiazole derivatives as amide-based inhibitors for soluble epoxide hydrolase (sEH) by molecular docking and MM-GBSA. *European Biophysics Journal, 46*(5), 445 – 459. doi:10.1007/s00249-016-1188-0

Kohrt, J. T., Bigge, C. F., Bryant, J. W., Casimiro-Garcia, A., Chi, L., Cody, W. L., Dahring, T., Dudley, D. A., Filipski, K. J., Haarer, S., Heemstra, R., Janiczek, N., Narasimhan, L., McClanahan, T., Peterson, J. T., Sahasrabudhe, V., Schaum, R., Van Huis, C. A., Welch, K. M., Zhang, E., Leadley, R. J. & Edmunds, J. J. (2007). The Discovery of (2R,4R)-N-(4-chlorophenyl)-N-(2-fluoro-4-(2-oxopyridin-1(2H)-yl)phenyl)-4-methoxypyrrolidine-1,2-dicarboxamide (PD 0348292), an Orally Efficacious Factor Xa Inhibitor. *Chemical Biology & Drug Design, 70*(2), 100 – 112. doi:10.1111/j.1747-0285.2007.00539.x

Kontoyianni, M. (2017). Docking and Virtual Screening in Drug Discovery. In: Lazar, I., Kontoyianni, M. & Lazar, A. (eds) *Proteomics for Drug Discovery. Methods in Molecular Biology, vol 1647.* New York, USA: Humana Press. doi:10.1007/978-1-4939-7201-2_18

Kramer, J.S., Pogoryelov, D., Hiesinger, K. & Proschak, E. (2018). Soluble epoxide hydrolase in complex with talinolol. *The Protein Data Bank, 6HGV*. doi:10.2210/pdb6hgv/pdb

Lavecchia, A. & Di Giovanni, C. (2013). Virtual Screening Strategies in Drug Discovery: A Critical Review. *Current Medicinal Chemistry, 20*(23), 2839 – 2860. doi:10.2174/09298673113209990001

Lipinski, C. A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies, 1*(4), 337 – 341. doi:10.1016/j.ddtec.2004.11.007

Longenecker, K. L., Fry, E. H., Lake, M. R., Solomon, L. R., Pei, Z. & Li, X. (2006). Crystal structure of human dipeptidyl peptidase IV (DPPIV) complexed with cyanopyrrolidine (C5-pro-pro) inhibitor 21ag. *The Protein Data Bank, 2G5T*. doi:10.2210/pdb2g5t/pdb

Lopes, J. C. D., dos Santos, F. M., Martins-José, A., Augustyns, K. & De Winter, H. (2017). The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability. *Journal of Cheminformatics, 9*(7), 1 – 11. doi:10.1186/s13321-016-0189-4

Martinez, C. R. & Iverson, B. L. (2012). Rethinking the term "pi-stacking." *Chemical Science, 3*(7), 2191 – 2201. doi:10.1039/c2sc20045g

McInnes, C. (2007). Virtual screening strategies in drug discovery. *Current Opinion in Chemical Biology, 11*(5), 494 – 502. doi:10.1016/j.cbpa.2007.08.033

Morisseau, C., Sahdeo, S., Cortopassi, G. & Hammock, B. D. (2013). Development of an HTS assay for EPHX2 phosphatase activity and screening of nontargeted libraries. *Analytical Biochemistry, 434*(1), 105 – 111. doi:10.1016/j.ab.2012.11.017

Muller, P. (1994). Glossary of terms used in physical organic chemistry (IUPAC Recommendations 1994). *Pure and Applied Chemistry, 66*(5), 1077 – 1184. doi:10.1351/pac199466051077

Müller-Jung, J. (2020, December 1). Künstliche Intelligenz macht ernst im Biolabor. *Frankfurter Allgemeine Zeitung*. https://www.faz.net/aktuell/wissen/kuenstliche-intelligenz-macht-ernst-im-biolabor-17079732.html

Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. (2012). Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medical Chemistry, 55*(14), 6582 – 6594. doi:10.1021/jm300687e

Ng, M. C. K., Fong, S. & Siu, S. W. I. (2015). PSOVina: The hybrid particle swarm optimization algorithm for protein-ligand docking. *Journal of Bioinformatics and Computational Biology, 13*(3), 1541007-1 – 1541007-18. doi:10.1142/s0219720015410073

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics, 3*(33), 1 – 14. doi:10.1186/1758-2946-3-33

Österberg, F., Morris, G. M., Sanner, M. F., Olson, A. J. & Goodsell, D. S. (2002). Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins: Structure, Function, and Bioinformatics, 46*(1), 34 – 40. doi:10.1002/prot.10028

Pagadala, N. S., Syed, K. & Tuszynski, J. (2017). Software for molecular docking: a review. *Biophysical Reviews, 9*(2), 91 – 102. doi:10.1007/s12551-016-0247-1

Pham, T. A. & Jain, A. N. (2008). Customizing scoring functions for docking. *Journal of Computer-Aided Molecular Design, 22*(5), 269 – 286. doi:10.1007/s10822-008-9174-y

Pillai, V. B., Sundaresan, N. R., Samant, S. A., Wolfgeher, D., Trivedi, C. M. & Gupta, M. P. (2011). Acetylation of a Conserved Lysine Residue in the ATP Binding Pocket of p38 Augments Its Kinase Activity during Hypertrophy of Cardiomyocytes. *Molecular and Cellular Biology, 31*(11), 2349 – 2363. doi:10.1128/MCB.01205-10

Quinn, D. M. (1987). Acetylcholinesterase: Enzyme Structure, Reaction Dynamics, and Virtual Transition States. *Chemical Reviews, 87*(5), 955 – 979. doi:10.1021/cr00081a005

Rai, R., Sprengeler, P. A., Elrod, K. C. & Young, W. B. (2001). Perspectives on Factor Xa Inhibition. *Current Medicinal Chemistry, 8*(2), 101 – 119. doi:10.2174/0929867013373822

Raschka, S. (2017). BioPandas: Working with molecular structures in pandas DataFrames. *Journal of Open Source Software, 2*(14), 279. doi:10.21105/joss.00279

Ravipati, G., McClung, J. A., Aronow, W. S., Peterson, S. J. & Frishman, W. H. (2007). Type 5 Phosphodiesterase Inhibitors in the Treatment of Erectile Dysfunction and Cardiovascular Disease. *Cardiology in Review, 15*(2), 76 – 86. doi:10.1097/01.crd.0000233904.77128.49

RDKit. (2021). *RDKit: Open-Source Cheminformatics Software*. RDKit. https://rdkit.org/

Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F. & Schroeder, M. (2015). PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Research, 43*(1), 443 – 447. doi: 10.1093/nar/gkv315

Segalés, J., Perdiguero, E. & Muñoz-Cánoves, P. (2016). Regulation of Muscle Stem Cell Functions: A Focus on the p38 MAPK Signaling Pathway. *Frontiers in Cell and Developmental Biology, 4*(91), 1 – 15. doi:10.3389/fcell.2016.00091

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K. & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature, 577*(1), 706 – 710. doi:10.1038/s41586-019-1923-7

Tang, Y.T. & Marshall, G.R. (2011). Virtual Screening for Lead Discovery. In: Satyanarayanajois S. (eds) *Drug Design and Discovery. Methods in Molecular Biology (Methods and Protocols), vol 716*. New York, USA: Humana Press. doi:10.1007/978-1-61779-012-6_1

Thomas, M. P. & Potter, B. V. L. (2011). Crystal structures of 11β-hydroxysteroid dehydrogenase type 1 and their use in drug discovery. *Future Medicinal Chemistry, 3*(3), 367 – 390. doi:10.4155/fmc.10.282

Tripathi, A. & Srivastava, U. C. (2008). Acetylcholinesterase: A Versatile Enzyme of Nervous System. *Annals of Neurosciences, 15*(4), 106 – 111. doi:10.5214/ans.0972.7531.2008.150403

Trott, O. & Olson, A. J. (2009). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry, 31*(1), 455 – 461. doi:10.1002/jcc.21334

Vane, J. R., Bakhle, Y. S. & Botting, R. M. (1998). Cyclooxygenases 1 and 2. *Annual Review of Pharmacology and Toxicology, 38*(1), 97 – 120. doi:10.1146/annurev.pharmtox.38.1.97

Venkatachalam, C. M., Jiang, X., Oldfield, T. & Waldman, M. (2003). LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of Molecular Graphics and Modelling, 21*(4), 289 – 307. doi:10.1016/s1093-3263(02)00164-x

Waltenberger, B., Garscha, U., Temml, V., Liers, J., Werz, O., Schuster, D. & Stuppner, H. (2016). Discovery of Potent Soluble Epoxide Hydrolase (sEH) Inhibitors by Pharmacophore-Based Virtual Screening. *Journal of Chemical Information and Modeling, 56*(4), 747 – 762. doi:10.1021/acs.jcim.5b00592

Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., Tian, S. & Hou, T. (2016). Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics, 18*(18), 12964 – 12975. doi:10.1039/c6cp01555g

Williams, M. A. & Ladbury, J. E. (2003). Hydrogen Bonds in Protein-Ligand Complexes. In: Böhm, H.-J., Schneider, G., Mannhold, R., Kubinyi, H. & Folkers, G. (eds) *Protein-Ligand Interactions: From Molecular Recognition to Drug Design (Methods and Principles in Medicinal Chemistry), vol 19*. Weinheim, DE: WILEY-VCH Verlag GmbH & Co. KGaA. doi:10.1002/3527601813.ch6

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C. & Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research, 46*(1), 1074 – 1082. doi:10.1093/nar/gkx1037

Xu, S., Hermanson, D. J., Banerjee, S., Ghebreselasie, K., Clayton, G. M., Garavito, R. M. & Marnett, L. J. (2014). Crystal Structure of Ovine Cyclooxygenase-1 Complex with Meloxicam. *The Protein Data Bank, 4O1Z.* doi:10.2210/pdb4o1z/pdb

Xu, S., Hermanson, D. J., Banerjee, S., Ghebreselasie, K., Clayton, G. M., Garavito, R. M. & Marnett, L. J. (2014). Oxicams Bind in a Novel Mode to the Cyclooxygenase Active Site via a Two-water-mediated H-bonding Network. *Journal of Biological Chemistry, 289*(19), 6799 – 6808. doi:10.1074/jbc.M113.517987

Xu, Y. & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing, 2*(3), 249 – 262. doi:10.1007/s41664-018-0068-2

Yang, J., Roy, A. & Zhang, Y. (2013). BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research, 41*(1), 1096 – 1103. doi:10.1093/nar/gks966

Zhao, H. & Caflisch, A. (2013). Discovery of ZAP70 inhibitors by high-throughput docking into a conformation of its kinase domain generated by molecular dynamics. *Bioorganic & Medicinal Chemistry Letters, 23*(20), 5721 – 5726. doi:10.1016/j.bmcl.2013.08.009

Appendix

Appendix 1 – PDB structures for HSD11B1:

2BEL, 2ILT, 2IRW, 2RBE, 3BYZ, 3BZU, 3CZR, 3D3E, 3D4N, 3D5Q, 3EY4, 3FCO, 3FRJ, 3H6K, 3HFG, 3OQ1, 3PDJ, 3QQP, 4BB5, 4BB6, 4C7J, 4C7K, 4HFR, 4HX5, 4K1L, 4P38, 4YYZ, 6NJ7;

Appendix 2 – PDB structures for ACHE:

1B41, 1F8U, 2X8B, 3LII, 4BDT, 4EY4, 4EY5, 4EY6, 4EY7, 4EY8, 4M0E, 4M0F, 5FPQ, 5HF5, 5HF6, 5HF8, 5HF9, 5HFA, 5HQ3, 6CQT, 6CQU, 6CQV, 6CQW, 6CQX, 6CQY, 6CQZ, 6F25, 6NEA, 6NTG, 6NTH, 6NTK, 6NTL, 6NTM, 6NTN, 6NTO, 6O4W, 6O4X, 6O50, 6O52, 6O5R, 6O5S, 6O5V, 6O66, 6U34, 6U37, 6U3P, 6WUV, 6WUY, 6WUZ, 6WV1, 6WVC, 6WVP, 6WVQ;

Appendix 3 – PDB structures for FXA:

1EZQ, 1F0R, 1F0S, 1FAX, 1FJS, 1G2L, 1G2M, 1IOE, 1IQE, 1IQF, 1IQG, 1IQH, 1IQI, 1IQJ, 1IQK, 1IQL, 1IQM, 1IQN, 1KSN, 1LPG, 1LPK, 1LPZ, 1LQD, 1MQ5, 1MQ6, 1NFU, 1NFW, 1NFX, 1NFY, 1P0S, 1V3X, 1WU1, 1XKA, 1XKB, 1Z6E, 2BMG, 2BOH, 2BOK, 2BQ6, 2BQ7, 2BQW, 2CJI, 2D1J, 2EI6, 2EI7, 2EI8, 2FZZ, 2G00, 2J2U, 2J34, 2J38, 2J4I, 2J94, 2J95, 2JKH, 2P16, 2P3T, 2P3U, 2P93, 2P94, 2P95, 2PHB, 2PR3, 2Q1J, 2RA0, 2UWL, 2UWO, 2UWP, 2VH0, 2VH6, 2VVC, 2VVU, 2VVV, 2VWL, 2VWM, 2VWN, 2VWO, 2W26, 2W3I, 2W3K, 2WYG, 2WYJ, 2XBV, 2XBW, 2XBX, 2XBY, 2XC0, 2XC4, 2XC5, 2Y5F, 2Y5G, 2Y5H, 2Y7X, 2Y7Z, 2Y80, 2Y81, 2Y82, 3CEN, 3CS7, 3ENS, 3FFG, 3HPT, 3IIT, 3K9X, 3KL6, 3KQB, 3KQC, 3KQD, 3KQE, 3LIW, 3M36, 3M37, 3Q3K, 3SW2, 3TK5, 3TK6, 4A7I, 4BTI, 4BTT, 4BTU, 4Y6D, 4Y71, 4Y76, 4Y79, 4Y7A, 4Y7B, 4ZH8, 4ZHA, 5K0H;

Appendix 4 – PDB structures for sheep COX1:

1CQE, 1DIY, 1EBV, 1EQG, 1EQH, 1FE2, 1HT5, 1HT8, 1IGX, 1IGZ, 1PGE, 1PGF, 1PGG, 1PTH, 1Q4G, 2AYL, 2OYE, 2OYU, 3KK6, 3N8X, 3N8Y, 3N8Z, 4O1Z, 5U6X, 5WBE;

Appendix 5 – PDB structures for human COX2:

5IKR, 5IKQ, 5IKT, 5IKV, 5KIR, 5F1A;

Appendix 6 – PDB structures for mouse COX2:

1CVU, 1CX2, 1DDX, 1PXX, 3HS5, 3HS6, 3HS7, 3KRK, 3LN0, 3LN1, 3MDL, 3MQE, 3NT1, 3NTB, 3NTG, 3OLT, 3OLU, 3PGH, 3Q7D, 3QH0, 3QMO, 3RR3, 3TZI, 4COX, 4E1G, 4FM5, 4M10, 4M11, 4OTJ, 4OTY, 4PH9, 4RRW, 4RRX, 4RRY, 4RRZ, 4RS0, 4RUT, 4Z0L, 5W58, 6BL3, 6BL4, 6COX, 6OFY, 6V3R;

Appendix 7 – PDB structures for DPP4:

1N1M, 1NU8, 1R9N, 1RWQ, 1TKR, 1W1I, 1WCY, 1X70, 2AJL, 2BGN, 2BGR, 2BUB, 2FJP, 2G5P, 2G5T, 2G63, 2HHA, 2I03, 2I78, 2IIT, 2IIV, 2JID, 2OAG, 2OGZ, 2OLE, 2ONC, 2OPH, 2OQI, 2OQV, 2P8S, 2QJR, 2QKY, 2QOE, 2QT9, 2QTB, 2RGU, 2RIP, 3BJM, 3C43, 3C45, 3CCB, 3CCC, 3D4L, 3EIO, 3F8S, 3G0B, 3G0C, 3G0D, 3G0G, 3H0C, 3HAB, 3HAC, 3KWF, 3KWJ, 3NOX, 3O95, 3O9V, 3OC0, 3OPM, 3Q0T, 3Q8W, 3QBJ, 3SWW, 3SX4, 3VJK, 3VJL, 3VJM, 3W2T, 3WQH, 4A5S, 4DSA, 4DSZ, 4DTC, 4G1F, 4J3J, 4JH0, 4KR0, 4L72, 4LKO, 4N8D, 4N8E, 4PNZ, 4PV7, 4QZV, 5I7U, 5ISM, 5J3J, 5KBY, 5T4B, 5T4E, 5T4F, 5T4H, 5Y7H, 5Y7J, 5Y7K, 5ZID, 6B1E, 6B1O;

Appendix 8 – PDB structures for MAOB:

1GOS, 1OJ9, 1OJA, 1OJC, 1OJD, 1S2Q, 1S2Y, 1S3B, 1S3E, 2BK3, 2BK4, 2BK5, 2BYB, 2C64, 2C65, 2C66, 2C67, 2C70, 2C72, 2C73, 2C75, 2C76, 2V5Z, 2V60, 2V61, 2VRL, 2VRM, 2VZ2, 2XCG, 2XFN, 2XFO, 2XFP, 2XFQ, 2XFU, 3PO7, 3ZYX, 4A79, 4A7A, 4CRT, 5MRL, 6FVZ, 6FW0, 6FWC, 6RKB, 6RKP, 6RLE, 6YT2;

Appendix 9 – PDB structures for MAPK14:

1A9U, 1BL6, 1BL7, 1BMK, 1DI9, 1IAN, 1KV1, 1KV2, 1M7Q, 1OUK, 1OUY, 1OVE, 1OZ1, 1W7H, 1W82, 1W83, 1W84, 1WBN, 1WBO, 1WBS, 1WBT, 1WBV, 1WBW, 1YQJ, 1ZYJ, 1ZZ2, 1ZZL, 2BAJ, 2BAK, 2BAL, 2BAQ, 2GFS, 2I0H, 2QD9, 2RG5, 2RG6, 2YIS, 2YIW, 2YIX, 2ZAZ, 2ZB0, 2ZB1, 3BV2, 3BV3, 3BX5, 3CTQ, 3D7Z, 3D83, 3DS6, 3DT1, 3E92, 3E93, 3FC1, 3FI4, 3FKL, 3FKN, 3FKO, 3FL4, 3FLN, 3FLQ, 3FLS, 3FLW, 3FLY, 3FLZ, 3FMH, 3FMJ, 3FMK, 3FML, 3FMM, 3FMN, 3FSF, 3FSK, 3GC7, 3GCP, 3GCQ, 3GCS, 3GCV, 3GFE, 3GI3, 3HA8, 3HEC, 3HEG, 3HL7, 3HLL, 3HP2, 3HP5, 3HRB, 3HUB, 3HUC, 3HV3, 3HV4, 3HV5, 3HV6, 3HV7, 3HVC, 3IPH, 3ITZ, 3IW5, 3IW6, 3IW7, 3IW8, 3K3I, 3K3J, 3KF7, 3KQ7, 3L8S, 3L8X, 3LFA, 3LFB, 3LFC, 3LFD, 3LFE, 3LFF, 3LHJ, 3MPT, 3MVL, 3MVM, 3MW1, 3NEW, 3NNU, 3NNV, 3NNW, 3NNX, 3NWW, 3OCG, 3PG3, 3QUD, 3QUE, 3RIN, 3ROC, 3S3I, 3S4Q, 3U8W, 3UVP, 3UVQ, 3UVR, 3ZS5, 3ZSG, 3ZSH, 3ZSI, 3ZYA, 4A9Y, 4AA0, 4AA4, 4AA5, 4AAC, 4DLI, 4DLJ, 4E6A, 4E6C, 4E8A, 4EH2, 4EH3, 4EH4, 4EH5, 4EH6, 4EH7, 4EH8, 4EH9, 4EHV, 4EWQ, 4F9W, 4F9Y, 4FA2, 4KIN, 4KIP, 4KIQ, 4L8M, 4R3C, 5ML5, 5MTX, 5MTY, 5N63, 5N64, 5N65, 5N66, 5N67, 5N68, 5OMG, 5OMH, 5TBE, 5TCO, 5WJJ, 5XYX, 5XYY, 6ANL, 6HWT, 6HWU, 6HWV, 6M95, 6M9L, 6OHD, 6QDZ, 6QE1, 6SFI, 6SFJ, 6SFK, 6SFO, 6ZWP;

Appendix 10 – PDB structures for PDE5:

1RKP, 1T9R, 1T9S, 1TBF, 1UDT, 1UDU, 1UHO, 1XOZ, 1XP0, 2H42, 2H44, 2XSS, 3B2R, 3BJC, 3SHY, 3SHZ, 3SIE, 3TGE, 3TGG, 4G2W, 4G2Y, 4I9Z, 4IA0, 4MD6, 4OEW, 4OEX, 5JO3, 5ZZ2, 6ACB, 6IWI, 6L6E, 6VBI;

Appendix 11 – PDB structures for PTP1B:

1AAX, 1BZC, 1BZH, 1BZJ, 1C83, 1C84, 1C85, 1C86, 1C87, 1C88, 1ECV, 1EEN, 1EEO, 1G1F, 1G1G, 1G1H, 1G7F, 1G7G, 1GFY, 1JF7, 1KAK, 1KAV, 1L8G, 1LQF, 1NL9, 1NNY, 1NO6, 1NWL, 1ONY, 1ONZ, 1PH0, 1PTT, 1PTU, 1PTV,1 PTY, 1PXH, 1PYN, 1Q1M, 1Q6J, 1Q6M, 1Q6N, 1Q6P, 1Q6S, 1Q6T, 1QXK, 1T48, 1T49, 1T4J, 1WAX, 1XBO, 2AZR, 2B07, 2BGD, 2BGE, 2CM7, 2CM8, 2CMA, 2CMB, 2CMC, 2CNE, 2CNF, 2CNG, 2CNH, 2CNI, 2F6T, 2F6V, 2F6W, 2F6Y, 2F6Z, 2F70, 2F71, 2FJM, 2FJN, 2H4G, 2H4K, 2HB1, 2NT7, 2NTA, 2QBP, 2QBQ, 2QBR, 2QBS, 2VEU, 2VEV, 2VEW, 2VEX, 2VEY, 2ZMM, 2ZN7, 3CWE, 3D9C, 3EAX, 3EB1, 4BJO, 4I8N, 4QAH, 4QAP, 4QBW, 4Y14, 4ZRT, 5K9W, 5T19;

Appendix 12 – PDB structures for SEH:

1VJ5, 1ZD2, 1ZD3, 1ZD4, 1ZD5, 3ANS, 3ANT, 3I1Y, 3I28, 3KOO, 3OTQ, 3PDC, 3WK4, 3WK5, 3WK6, 3WK7, 3WK8, 3WK9, 3WKA, 3WKB, 3WKC, 3WKD, 3WKE, 4C4X, 4C4Y, 4C4Z, 4HAI, 4J03, 4JNC, 4OCZ, 4OD0, 4X6X, 4X6Y, 4Y2J, 4Y2P, 4Y2Q, 4Y2R, 4Y2S, 4Y2T, 4Y2U, 4Y2V, 4Y2X, 4Y2Y, 5AI0, 5AI4, 5AI5, 5AI6, 5AI8, 5AI9, 5AIA, 5AIB, 5AIC, 5AK3, 5AK4, 5AK5, 5AK6, 5AKE, 5AKG, 5AKH, 5AKI, 5AKJ, 5AKK, 5AKL, 5AKX, 5AKY, 5AKZ, 5ALD, 5ALE, 5ALF, 5ALG, 5ALH, 5ALI, 5ALJ, 5ALK, 5ALL, 5ALM, 5ALN, 5ALO, 5ALP, 5ALQ, 5ALR, 5ALS, 5ALT, 5ALU, 5ALV, 5ALW, 5ALX, 5ALY, 5ALZ, 5AM0, 5AM1, 5AM2, 5AM3, 5AM4, 5AM5, 5FP0, 5MWA, 6AUM, 6FR2, 6HGV, 6HGW, 6HGX, 6I5G, 6YL4;