Ein Bild, das Grafiken, Schrift, Grafikdesign, Logo enthält.

KI-generierte Inhalte können fehlerhaft sein.TU_Logo_HKS41.wmf

**Faculty of Computer Science -** Computational Modeling and Simulation

Comparison of experimental verified and predicted protein-ligand interactions using Protenix and PLIP

Masterarbeit

zur Erlangung des Hochschulgrades

Master of Science

im Master-Studiengang

Computational Modeling and Simulation

vorgelegt von: Tränkner, Masine Janet

1. Gutachter: Prof. Dr. Michael Schroeder
2. Gutachter: Dr. Alexander Wurm

Eingereicht am 09.09.2025

Unterschrift

**Bi**bliografischer Nachweis

Tränkner, Masine Janet

Comparison of experimental verified and predicted protein-ligand interactions using Protenix and PLIP

- 2025 - 64 Seiten - 8 Seiten Anlagen

Masterarbeit

Technische Universität Dresden  
Fakultät Computerwissenschaften  
Biotechnologisches Zentrum (BIOTEC)  
Schroeder Gruppe - Bioinformatik

**Abstracts**

Deutsch

Deep-Learning-basierte Strukturvorhersage- und Docking-Methoden wandeln den Bereich der strukturgeführten Wirkstoffforschung. In dieser Arbeit wurden diese Ansätze auf die Phosphopantothenoylcystein-Decarboxylase (PPCDC) angewendet, ein Enzym, das für die Biosynthese von Coenzym A essenziell ist und ein potenzielles therapeutisches Ziel bei akuter myeloischer Leukämie (AML) darstellt.

Mit Protenix erstellte Strukturmodelle bildeten die Grundlage für Docking-Studien mit GNINA. Diese Docking-Ergebnisse wurden durch eine Interaktionsanalyse mit PLIP und eine Bewertung der Ligandenposition und -orientierung ausgewertet, was eine differenzierte Rangfolge potenzieller Inhibitoren ermöglichte. Die Arbeit beleuchtet die Herausforderungen, die sich durch kryptische Bindungsstellen ergeben, und identifiziert mehrere Liganden-Kandidaten mit günstigen Interaktionsprofilen an des katalytischen Zentrums.

Insgesamt zeigt diese Arbeit das Potenzial der Integration von KI-gesteuerter Proteinstrukturvorhersage, Docking und Interaktionsanalyse in einen einheitlichen Arbeitsablauf zur Beschleunigung der Entdeckung neuartiger Therapeutika.

English

Deep learning-based structure prediction and docking methods are transforming the field of structure-guided drug discovery. In this thesis, these approaches were applied to phosphopantothenoylcysteine decarboxylase (PPCDC), an enzyme essential for coenzyme A biosynthesis and a potential therapeutic target in acute myeloid leukemia (AML).

Structural models generated with Protenix provided the basis for docking studies using GNINA. The docking results were evaluated by an interaction analysis with PLIP and ligand pose and orientation evaluation, enabled a refined ranking of potential inhibitors. The study highlights challenges posed by cryptic binding pockets, and identifies several ligand candidates with favorable interaction profiles at the catalytic site.

This work demonstrates the potential of integrating AI-driven protein structure prediction, docking, and interaction analysis into a unified workflow for accelerating the discovery of novel therapeutics.

Table of Contents

[List of Figures III](#_Toc207988547)

[List of Tables IV](#_Toc207988548)

[Formula Index V](#_Toc207988549)

[List of abbreviations VI](#_Toc207988550)

[1. Introduction 1](#_Toc207988551)

[2. Motivation 3](#_Toc207988552)

[2.1 Introduction of Phosphopantothenoylcysteine Decarboxylase 3](#_Toc207988553)

[2.2 Acute Myeloid Leukemia cells’ Dependencies and Coenzyme A Biosynthesis 4](#_Toc207988554)

[3. Theoretical Background 7](#_Toc207988555)

[3.1 Protein-Ligand Complexes 7](#_Toc207988556)

[3.1.1 Protein-Ligand Binding Models 7](#_Toc207988557)

[3.1.2 Types of Molecular Interactions in Protein-Ligand Binding 9](#_Toc207988558)

[3.1.3 Concepts of Inhibition 11](#_Toc207988559)

[3.2 Protein Crystal Structures and Digital Structure Acquisition 14](#_Toc207988560)

[3.3 Molecular Docking 16](#_Toc207988561)

[3.3.1 Defining the Binding Box 17](#_Toc207988562)

[3.3.2 Search Functions 18](#_Toc207988563)

[3.3.3 Scoring Functions 20](#_Toc207988564)

[3.4 Deep Learning-Based 3D Structure Prediction 24](#_Toc207988565)

[4. Materials and Methods 26](#_Toc207988566)

[4.1 System and Environment 26](#_Toc207988567)

[4.2 Data Gathering and Preparation 27](#_Toc207988568)

[4.2.1 Biological and Structural Data 27](#_Toc207988569)

[4.2.2 Ligand Dataset and Preprocessing 28](#_Toc207988570)

[4.3 Docking Experiments 30](#_Toc207988571)

[4.4 Post-Docking Analysis and Interaction Profiling 31](#_Toc207988572)

[5. Results 35](#_Toc207988573)

[5.1 3D Structure for Docking 35](#_Toc207988574)

[5.2 Docking Pipeline 40](#_Toc207988575)

[5.3 Finalized Dataset after Docking 41](#_Toc207988576)

[5.4 PLIP Pattern Results 44](#_Toc207988577)

[5.5 Final Candidate Selection for Experimental Testing 48](#_Toc207988578)

[5.6 Reevaluation with Protenix 50](#_Toc207988579)

[6. Discussion 52](#_Toc207988580)

[7. Outlook 59](#_Toc207988581)

[8. Summary 63](#_Toc207988582)

[Bibliography XVI](#_Toc207988583)

[Erklärung XXIII](#_Toc207988584)

# List of Figures

[Figure 1 - The Coenzyme A pathway [22] 5](#_Toc207988585)

[Figure 2 - Protein-Ligand Binding Models: A) Lock-and-Key model, B) Induced fit model and C) Conformational Selection Model (adapted from Munzar, et al. (2016) [23]) 7](#_Toc207988586)

[Figure 3 - Types of reversible inhibition in comparison to uninhibited binding 12](#_Toc207988587)

[Figure 4 - Composition of ligand search space in degrees of freedom: A) rotational and B) translational degrees of freedom 18](#_Toc207988588)

[Figure 5 - PPCDC holoenzyme (image from the RCSB PDB of 1QZU [12]) 27](#_Toc207988589)

[Figure 6 - PDB Pairwise Structural alignment of human and *At.* PPCDC chain A (created using Mol\* [76]) 35](#_Toc207988590)

[Figure 7 - 1QZU Protenix prediction [9] binding site conformation with and without natural ligands (created using PyMOL [13]) 37](#_Toc207988591)

[Figure 8 - PDB Pairwise Structural alignment of human PPCDC wildtype against 1QZU Protenix prediction with FMN and PPan-Cys and *At.* C175S mutant (created using Mol\* [76]) 38](#_Toc207988592)

[Figure 9 - PCO vs. 1QZU + PPan-Cys Protenix prediction pose comparison 39](#_Toc207988593)

[Figure 10 - Flowchart of the entire analysis, consisting of pre-processing, docking and post-processing 40](#_Toc207988594)

[Figure 11 - Violin plot of docking affinity distribution with applied NeighborSearch filters 42](#_Toc207988595)

[Figure 12 - Distribution of docking affinity against heavy atom count 43](#_Toc207988596)

[Figure 13 - Histogram of MCS-RMSD values across all docked ligands 43](#_Toc207988597)

[Figure 14 - Docking affinity plotted against MCS-RMSD 44](#_Toc207988598)

[Figure 15 - PLIP interaction patterns and overlaps of substrate binding in human PPCDC using Marvin JS [77] 46](#_Toc207988599)

[Figure 16 - comparison of pLDDT scores of the top five candidates after ranking IX](#_Toc207988600)

[Figure 17 - PLIP patterns of the complete dataset (blue) with PCO-1MVN as reference (red) of hydrophobic interactions, hydrogen bonds, halogen bonds, metal complexes XI](#_Toc207988601)

[Figure 18 - PLIP patterns of the complete dataset (blue) with PCO-1MVN as reference (red) of pi stakcs, pi cation interactions, water bridges, salt bridges XII](#_Toc207988602)

[Figure 19 - PLIP patterns of the 4 Å dataset (blue) with PCO-1MVN as reference (red) of hydrophobic interactions, hydrogen bonds, halogen bonds, metal complexes XIV](#_Toc207988603)

[Figure 20 - PLIP patterns of the 4 Å dataset (blue) with PCO-1MVN as reference (red) of pi stakcs, pi cation interactions, water bridges, salt bridges XV](#_Toc207988604)

# List of Tables

[Table 1 - PDB Pairwise Structural alignment results of 1QZU, 1E20, 1MVN [5] 36](#_Toc207988605)

[Table 2 - PDB Pairwise Structural alignment results of 1QZU, 1MVN and the 1QZU protenix prediction with FMN and PPan-Cys [5] 39](#_Toc207988606)

[Table 3 - Overlap and Spearman Correlation Between Rankings Based on Different Criteria Combinations 49](#_Toc207988607)

[Table 4 - Re-evaluation of the five best candidates using Protenix and GNINA 51](#_Toc207988608)

# Formula Index

[(1) Empirical scoring function 21](#_Toc207988609)

[(2) Template modelling score 25](#_Toc207988610)

[(3) Tanimoto coefficient 29](#_Toc207988611)

[(4) Center of Mass 32](#_Toc207988612)

[(5) Center of Mass distance 32](#_Toc207988613)

[(6) Root Mean Square Deviation 32](#_Toc207988614)

[(7) Pearson correlation coefficient 33](#_Toc207988615)

[(8) PLIP score 33](#_Toc207988616)

# List of abbreviations

AML Acute myeloid leukemia

*At. Arabidopsis thaliana*

Cab2 Coenzyme A biosynthesis protein 2

CFCG Chemical Flow Chart Generator

CMCB Center for Molecular and Cellular Bioengineering

CNN Convolutional Neural Network

CoA Coenzyme A

COASY Coenzyme A synthase

COM center of mass

conform conformational

CPU central processing unit

cryo-EM cryo-Electron Microscopy

elec electrostatics

FMN flavin mononucleotide

FRET Förster (or Fluorescence) Resonance Energy Transfer

GA Genetic Algorithm

GPU graphics processing unit

hbond hydrogen bonding

HFCD homo‑oligomeric flavin‑containing Cys decarboxylase

IC₅₀ half maximal inhibitory concentration

IDP Intrinsically Disordered Protein

ipTM interface predicted template modelling

Kd dissociation constant

MCS Maximum Common Substructure

ML Machine Learning

MSA Multiple Sequence Alignment

NMR Nuclear Magnetic Resonance

PANK pantothenate kinase

PCO Pantothenoylaminoethenethiol

PDB Protein Data Bank

pLDDT predicted Local Distance Difference Test

PLIP Protein-Ligand Interaction Profiler

PPanSH 4′-Phosphopantetheine

PPC / PPan-Cys Phosphopantothenoylcysteine

PPCDC Phosphopantothenoylcysteine decarboxylase

PPCS Phosphopantothenoylcysteine synthase

pTM predicted template modelling

RCSB Research Collaboratory for Structural Bioinformatics

RMSD Root Mean Square Deviation

sdf Structures Data File

SMILES Simplified Molecular Input Line Entry System

sol solvation

TM-score Template modeling score

tor torsional

vdW Van der Waals force

XML Extensible Markup Language

# Introduction

In recent years, computational methods have become an indispensable tool in drug discovery, where the complexity of biological systems and the vastness of chemical space make purely experimental approaches both time-consuming and costly ( [1], [2]). Among these computational techniques, molecular docking and structure prediction have gained particular importance, as they allow the in-silico exploration of protein-ligand interactions prior to laboratory validation [3]. Advances in machine learning and deep learning have further extended these possibilities, enabling the prediction of three-dimensional structures and the evaluation of binding affinities with increasing accuracy ( [1], [3]). These developments provide researchers with the opportunity to narrow down potential drug candidates and to better understand the molecular basis of binding events, while significantly reducing experimental overhead. One of the most important applications of this field is the identification of potential inhibitors that can selectively bind to disease-relevant proteins and modulate their function, offering new therapeutic avenues for the treatment of cancer and other diseases ( [1], [4]).

In this context, the enzyme phosphopantothenoylcysteine decarboxylase (PPCDC) represents a promising drug target in the fight against acute myeloid leukemia (AML). PPCDC is involved in the biosynthesis of Coenzyme A, a pathway essential for cell survival, making its inhibition an attractive strategy for disrupting leukemic cell metabolism. The structural basis for investigating PPCDC, as well as for other proteins studied in this work, is provided by the Protein Data Bank (RCSB PDB) [5], which serves as the primary repository for experimentally determined three-dimensional biomolecular structures. By leveraging this structural information, computational pipelines can generate hypotheses about how small molecules interact with PPCDC and how such interactions may be exploited for therapeutic intervention.

The focus of this thesis lies on the identification and evaluation of potential ligands for PPCDC using docking approaches and deep learning-based tools. Molecular docking with GNINA [6] was employed to predict ligand binding poses and affinities, providing an initial ranking of candidates based on their predicted interaction strength. To complement these results, the Protein-Ligand Interaction Profiler (PLIP) [7] was used to analyze and quantify the interaction patterns between ligands and the target binding site, offering an independent metric for comparing predicted complexes.

In parallel, structure prediction methods such as the AlphaFold series [8] and Protenix [9] were applied to investigate the conformational properties of the protein and its binding environment, allowing for a re-evaluation of docking results and revealing potential discrepancies between docking-based predictions and structure-based modeling.

The central aim of this work is therefore twofold: to establish a robust computational pipeline for ligand evaluation and to critically assess the reliability of deep learning models in the context of drug discovery. A significant challenge in targeting PPCDC is the absence of a high-quality, holoenzyme conformation. To address this, deep learning-based structure models (via Protenix / AlphaFold derivatives) are employed as viable substitutes for experimental templates. Within this workflow, GNINA docking is used to explore potential ligand binding poses and affinities, while PLIP analysis provides detailed interaction fingerprints essential for evaluating binding mode plausibility. Together, this setup aims at identifying promising small-molecule inhibitors of PPCDC. This approach ultimately serves to identify candidate compounds for experimental follow-up, with the objective of developing a drug that successfully inhibits PPCDC, while highlighting strengths and limitations of Artificial Intelligence-based structure prediction in drug design.

# Motivation

## Introduction of Phosphopantothenoylcysteine Decarboxylase

The protein Phosphopantothenoylcysteine decarboxylase (PPCDC), also known as 4′-phosphopantothenoylcysteine decarboxylase (EC 4.1.1.36), occupies a critical role in Coenzyme A (CoA) biosynthesis by catalyzing the decarboxylation of phosphopantothenoyl-L-cysteine (PPC or PPan-Cys) to yield 4′-phosphopantetheine (PPanSH) and carbon dioxide, this being the third step in a canonical five-enzyme pathway consuming pantothenic acid (vitamin B₅) as input [10].

Human PPCDC is encoded by PPCDC (UniProt Q96CD2, COAC\_HUMAN) [11] and folds into a homotrimer, where each 206 amino acid long subunit [12] belongs to the homo-oligomeric flavoprotein decarboxylase (HFCD) family [13].

In eukaryotes, PPCDC is monofunctional, existing separately from the adjacent CoA biosynthetic step performed by PPCS (phosphopantothenoyl-cysteine synthase), unlike in most bacteria, where PPCDC and PPCS activities are often fused into a single bifunctional enzyme (CoaBC) [14].

Each PPCDC protomer binds one tightly associated flavin mononucleotide (FMN) cofactor in a non-covalent but high-affinity manner [15]. The enzyme features two functionally linked binding sites: an allosteric FMN-binding site and an adjacent substrate-binding site for PPan-Cys. The close positioning of FMN to the substrate enables the flavin to participate directly in catalysis. Uniquely among human flavoenzymes, PPCDC’s reaction does not involve an overall redox change [14]. Instead, FMN acts as a transient electron sink, stabilizing high-energy reaction intermediates by accepting charge from the substrate thiolate group via the isoalloxazine ring [16]. This electron redistribution lowers the activation barrier for decarboxylation and ensures proper intermediate stabilization.

The loss of FMN binding, either by cofactor removal, point mutations or allosteric inhibition, greatly diminishes or abolishes PPCDC decarboxylase activity, underlining FMN’s architectural and catalytic importance rather than redox role.

Variants such as p.Thr53Pro, affecting a residue critical for FMN binding, and p.Ala95Val, a likely destabilizing substitution, were identified in patient-derived fibroblasts. These cells showed nearly 50% reduced CoA levels and absent PPCDC protein, confirming the role of FMN in enzyme stability and CoA biosynthesis integrity [17].

## Acute Myeloid Leukemia cells’ Dependencies and Coenzyme A Biosynthesis

Phosphopantothenoylcysteine decarboxylase (PPCDC) represents a unique vulnerability in certain cancer cells, particularly in acute myeloid leukemia (AML).

AML is a rare but aggressive blood cancer characterized by the rapid proliferation of immature myeloid cells, also called myeloblasts, in the bone marrow and bloodstream ( [18], [19]). The underlying factors that precipitate AML are not fully understood, yet. It is believed that changes in the DNA of bone marrow cells cause AML [20]. Possible treatments include chemotherapy, stem cell transplants and participation in clinical trials [21]. The main goal to treat AML is to destroy the leukemia cells. Since AML develops rapidly, it is strongly recommended that individuals be diagnosed and treated promptly [21]. Another strategy to destroy AML cells is the targeting of their altered metabolism and dependencies.

In healthy human cells, PPCDC is not essential for Coenzyme A (CoA) biosynthesis (the pathway is seen in Figure 1) because the product of the enzyme, 4′-phosphopantetheine (PPanSH), can either be synthesized through alternative pathways or imported into the cell from extracellular sources. This metabolic redundancy ensures that loss or inhibition of PPCDC does not significantly compromise healthy cellular function.

Ein Bild, das Text, Diagramm, Screenshot, Plan enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 1 - The Coenzyme A pathway [22]

In contrast, AML cells appear to depend on PPCDC activity. Two non-exclusive explanations have been proposed for this selective dependency.

One reason may be an inability to compensate for product loss. Leukemia cells may lack the capacity to obtain PPanSH via alternative biosynthetic routes or uptake mechanisms (the Figure 1 salvage path). As a result, inhibition of PPCDC effectively shuts down the production of this essential metabolite, which impairs CoA-dependent processes that are critical for cellular proliferation and survival. This suggests that the de novo biosynthesis of CoA (see Figure 1) is the only way for AML cells to synthesize CoA.

A second hypothesis for why AML cells depend on PPCDC is that the protein may participate in protein-protein interactions or structural roles that are indispensable in AML cells, beyond its enzymatic activity. These functions could be specific to the proliferative or metabolic state of leukemia cells, making them particularly sensitive to PPCDC inhibition.

Taken together, this selective vulnerability presents a promising therapeutic opportunity. Targeting PPCDC could selectively impair leukemia cell growth while sparing healthy cells, potentially reducing off-target toxicity. The central premise of this study is to clarify the molecular basis of PPCDC dependency in AML and to identify small molecules that effectively inhibit its activity.

# Theoretical Background

In order to comprehend the mechanisms by which small molecules interact with proteins; it is necessary to develop both conceptual models that facilitate the conceptualization of the binding process and a practical accounting of the forces that actually stabilize complexes.

In recent years, machine learning has added another layer: powerful predictors of structure and of likely binding poses. In the following chapter, the classical and modern models of binding are introduced. The non-covalent interactions that stabilize complexes are then reviewed, as are the automated tools that classify them. Finally, deep-learning approaches used to predict structures and poses are discussed, as is the way in which these methods fit into a docking pipeline.

## Protein-Ligand Complexes

### Protein-Ligand Binding Models

Conceptual models of binding are useful because they provide intuition about what “happens” when a ligand meets a protein. Three models dominate historical and contemporary thinking.

Ein Bild, das Kreis enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 2 - Protein-Ligand Binding Models: A) Lock-and-Key model, B) Induced fit model and C) Conformational Selection Model (adapted from Munzar, et al. (2016) [23])

#### Lock-and-key model

The earliest and most intuitive image is Fischer's lock-and-key principle [24], which portrays the protein active site as a preformed cavity that fits a complementary ligand exactly (Figure 2.A). This model is useful when the binding site is rigid and the ligand is geometrically complementary. However, it fails to capture protein flexibility and the frequent conformational rearrangements observed experimentally.

#### Induced Fit Model

Koshland’s induced-fit model [25] explicitly acknowledges flexibility: ligand binding triggers conformational changes in the protein (and possibly the ligand) that optimize contacts and catalysis (see Figure 2.B). The induced-fit model highlights the dynamic interplay between shape complementarity and energetic adaptation. It explains cases in which a ligand binds tightly only after the protein repositions side chains or backbone segments.

#### Conformational Selection Model

More recently, conformational selection has reframed the issue: proteins in solution sample an ensemble of conformations, some of which already resemble the bound state [26]. Rather than inducing a new state, the ligand "selects" and stabilizes one of the pre-existing conformers. These models are not mutually exclusive. Binding can best be described as a combination of conformational selection followed by local induced adjustments. Unlike the Induced Fit Model, the Conformational Selection Model describes conformational change first, followed by binding [27].

A special case of conformational dynamics, cryptic pockets are binding sites that are not visible in the unbound (apo) protein structure but become exposed through structural rearrangement. Typically hidden in static crystal structures, these pockets may open upon ligand binding or during molecular dynamics simulations. Cryptic sites are of great interest in drug design because they offer novel, often more selective, targetable surfaces that are inaccessible to natural substrates.

### Types of Molecular Interactions in Protein-Ligand Binding

Protein-ligand binding is governed by a variety of non-covalent forces that together determine how tightly (affinity) and selectively (specificity) a drug binds its target. In molecular docking and drug design, scoring functions are used to approximate these interactions, which include hydrogen bonds, ionic (electrostatic) contacts, van der Waals (dispersion, polar, and induced interactions), hydrophobic contacts, and π-system interactions [28]. A survey of many protein-ligand complexes found that the most common contacts are hydrophobic contacts, hydrogen bonds and aromatic (π-π) stacking, followed by weaker hydrogen bonds, salt bridges and cation-π interactions [29].

In this chapter each major interaction type (hydrogen bonding, hydrophobic effect, electrostatic interactions including salt bridges, π-π stacking, cation-π interactions, halogen bonds and water bridges) is described, explaining their nature and how they contribute to binding affinity and specificity.

*Hydrogen bonds* occur when a hydrogen atom, which is covalently bound to an electronegative atom (typically N-H or O-H) interacts with a free pair of electrons on an electronegative acceptor atom (often O or N). These contacts are highly directional, favoring linear arrangements (D-H···A angles close to 180°) and typical donor-acceptor distances of approximately 3.0 Å. Due to their directionality and specific geometric requirements, hydrogen bonds strongly contribute to binding specificity [29]. On average, each hydrogen bond contributes between -1.5 and -4.7 kcal/mol to binding free energy. However, the effect of a hydrogen bond on binding can be negligible or even detrimental if the resulting interaction does not outweigh the "desolvation penalty" during ligand binding. Therefore, adding a donor or acceptor does not always improve binding unless it forms a strong, buried hydrogen bond that outcompetes water [29].

*Hydrophobic interactions* arise when nonpolar parts of ligand and protein come into contact, typically involving carbon-rich (aliphatic or aromatic) surfaces. In water, burying hydrophobic surfaces releases ordered water molecules, yielding a favorable entropy change. Hydrophobic contacts are mostly non-directional van der Waals attractions supplemented by the entropy gain from desolvation. These interactions are often considered the “main driving force” for affinity in protein-ligand binding [29].

Ferreira de Freitas et al. discovered that hydrophobic interactions were more prevalent in high-efficiency ligands than hydrogen bonds and ionic interactions [29]. Therefore, hydrophobic interactions significantly enhance binding affinity, particularly in well-buried pockets, though they exhibit reduced specificity, binding with any shape-complementary nonpolar surfaces. In the case of small ligands, polar contacts often prove to be a more effective means of achieving binding efficiency. However, through the process of optimization, it becomes increasingly challenging to satisfy geometric constraints, leading to an increase in the contribution of hydrophobic interactions [29].

*Electrostatic interactions* include attraction between oppositely charged groups (*ionic bonds* or *salt bridges*) and dipole-charge interactions. They are formed between charged amino acids no more than 4 Å away and a ligand atom of opposite charge [30]. These interactions play a key role in determining binding specificity by exhibiting geometrical complementarity and stability in energetics [31].

*π-π stacking* refers to attractive interactions between aromatic rings. These typically occur in two geometries: face-to-face and edge-to-face. These forces arise from a combination of dispersion and electrostatic interactions of the π-electron clouds. This also adds a degree of specificity: the requirement for both partners to have aromatic character and proper orientation means the ligand must complement the receptor’s aromatic residues [29].

Similar toπ-π stacking, a *cation-π interaction* is an electrostatic attraction between a positively charged group and the face of an aromatic ring. The π system’s cloud of electrons creates a negative potential that stabilizes the cation. These interactions are widespread in proteins and are known to play key roles in many recognition events [29]. Also, like π-π stacking, they rely on both a positive charge and an aromatic partner. Thus, they are quite specific: the ligand must present a charged group at just the right distance above an aromatic residue to form the contact.

*Halogen bonds* are directional interactions analogous to hydrogen bonds, but involve a halogen atom (typically Cl, Br or I) as the electrophile and a nucleophilic atom (often O or N) as the acceptor [32]. Halogen bonds exhibit what is known as σ-hole geometry.

For this to occur, the halogen must be covalently bonded in order to carry out a noncovalent interaction with a negatively charged site [33]. They have become useful elements in drug design. Fluorine and chlorine substitutions are often introduced to modulate a drug’s absorption, distribution and metabolic stability, while bromine and iodine are used to target binding selectivity [34].

In some protein-ligand complexes, direct hydrogen bonds between the ligand and protein are not always possible due to distance or geometry. Instead, structured water molecules within the binding site can mediate interactions by bridging hydrogen bonds between the ligand and protein residues. Such a *water bridge* typically involves one water molecule donating or accepting a hydrogen bond to both the ligand and a nearby residue, effectively linking them [30].

### Concepts of Inhibition

Enzyme inhibition refers to the process by which the catalytic activity of an enzyme is reduced or blocked entirely through the interaction with a small molecule. Inhibition is a fundamental mechanism for regulating metabolic pathways and is often exploited in drug development strategies. By preventing or reducing substrate turnover, inhibitors can modulate biological functions with high specificity. Generally, enzyme inhibition can be categorized as either reversible or irreversible, each with distinct biochemical and pharmacological implications.

Reversible inhibitors bind non-covalently to enzymes, allowing for the interaction to be reversed upon dilution or removal of the inhibitor. Three classical types are distinguished:

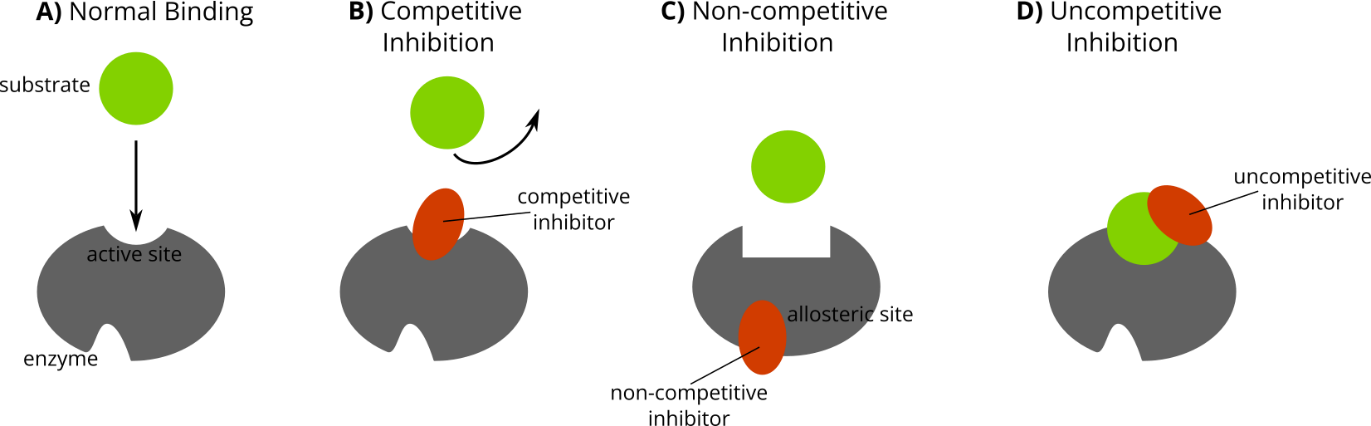


Figure 3 - Types of reversible inhibition in comparison to uninhibited binding

*Competitive inhibition* (Figure 3.B) occurs when an inhibitor binds the enzyme’s active site, directly competing with the substrate. This process is schematically represented as follows:

Ein Bild, das Schwarz, Dunkelheit enthält.

KI-generierte Inhalte können fehlerhaft sein.

Structurally, this utilizes the strategy mimicking the substrate: the inhibitor occupies the catalytic pocket and sterically prevents substrate binding. Kinetically, competitive inhibitors increase the apparent (Michaelis-Menten constant) without changing , the maximum velocity, that is possible for this reaction, in Michaelis-Menten formulations. The Michaelis-Menten constant corresponds to the substrate concentration at which the enzyme operates at half of its maximum possible speed, i.e., half of the active centers are occupied [35]. In other words, a greater amount of substrate is required to maintain the same number of active sites occupied by a substrate molecule.

A special case of competitive inhibition would be the so-called *Feedback Inhibition*. Here, the end product of a metabolic pathway inhibits an upstream enzyme. This is a self-regulating mechanism of fine-tuned regulation that ensures metabolic balance.

*Non-competitive (allosteric) inhibition* (Figure 3.C) arises when an inhibitor binds to a site that is distinct from the active site, altering the enzyme’s catalytic activity regardless of substrate binding. The formation of an Enzyme-Inhibitor-Substrate complex is now possible:

Ein Bild, das Schwarz, Dunkelheit enthält.

KI-generierte Inhalte können fehlerhaft sein.

In this case, decreases while remains unchanged. These interactions are mediated by a type of molecular transition triggered by an "allosteric ligand" [36]. This mechanism is frequently observed in regulatory proteins with allosteric sites.

Nevertheless, it should be noted that “allosteric effects” are generally not always destabilizing and activity-inhibiting. They can also lead to a stabilizing conformational change that activates the active center.

*Uncompetitive inhibition* (Figure 3.D) occurs when an inhibitor binds only to an enzyme-substrate complex, preventing catalysis by blocking the reaction intermediates within the catalytic site:

Ein Bild, das Schwarz, Dunkelheit enthält.

KI-generierte Inhalte können fehlerhaft sein.

This mechanism simultaneously reduces both and simultaneously by the same factor and is often associated with inhibitors that stabilize non-productive enzyme conformations or reaction intermediates [37].

In conclusion, studying enzyme inhibition is essential to understanding both fundamental biochemical pathways and applied drug discovery. Inhibitors can serve as tools to dissect enzyme mechanisms and as therapeutic agents to correct pathological states. Within the framework of this work, inhibition establishes a conceptual link between predicted ligand binding modes and their potential biological effects.

## Protein Crystal Structures and Digital Structure Acquisition

Proteins are inherently dynamic molecules whose three-dimensional conformations determine their biological function. Understanding these structures is essential in fields such as enzymology, drug discovery, and molecular docking studies. Protein crystal structures provide an experimental snapshot of a protein’s three-dimensional fold at atomic resolution. This allows for detailed analysis of active sites, ligand-binding pockets, and conformational integrity.

Protein crystallography with X-ray diffraction remains one of the most widely used methods for obtaining high-resolution structures. The target protein must be isolated in high purity and sufficient concentration. Environmental conditions are optimized to allow proteins to form a regular lattice (a crystal) in which molecules are arranged in repeating patterns. Crystals are then exposed to an X-ray beam. The pattern of diffracted X-rays depends on the electron density within the crystal, encoding the positions of atoms. These can then be used to reconstruct an electron density map through numerical methods [38].

Other techniques, including Nuclear Magnetic Resonance (NMR) spectroscopy and cryo-Electron Microscopy (cryo-EM), complement crystallography by providing structural information for proteins that are difficult to crystallize or for observing conformational dynamics in solution.

Nuclear Magnetic Resonance spectroscopy utilizes magnetic fields to capture proteins in solution, close to the cellular (membranous) environment. NMR measures the magnetic properties of specific nuclei within the protein when placed in a strong magnetic field ( [39], [40]). Through a series of multidimensional NMR experiments, one can obtain distance restraints, bond angle information, and dynamic data about the protein backbone and side chains.

It is particularly beneficial for intrinsically disordered proteins and for probing dynamic processes that are not easily captured by crystallography or cryo-EM.

In cryo-EM, purified protein samples are rapidly frozen in vitreous ice, preserving them in a near-native hydrated state without the need for crystallization. Using an electron microscope, multiple two-dimensional projections of individual protein particles are collected. Computational algorithms then align and reconstruct these projections into a three-dimensional density map [41].

The final product of this workflow is a three-dimensional atomic model of the protein, typically including backbone and side-chain coordinates, water molecules, cofactors, and sometimes other ligands.

Once a protein’s three-dimensional structure is determined experimentally, it can be made available in digital form for computational studies. These digital protein structures are typically stored in standardized formats such as PDB (Protein Data Bank) files, which contain the Cartesian coordinates of all atoms in the protein, along with metadata about the experimental method, resolution, and any bound ligands or cofactors. These structures can be obtained from online databanks, such as the RCSB PDB [5], to be used for computational processing, like removing and adding atoms, aligning structures or visualization.

These digital representations are essential for molecular docking and computational modeling and simulations, as they provide the starting point for investigating the 3D protein structure, possible binding sites and their impact on metabolic pathways and also protein-ligand interactions *in silico*.

## Molecular Docking

Given the absence of known inhibitors for PPCDC, virtual (*in silico*) screening is a logical first step. Molecular docking is a cornerstone technique in structure-based drug design, serving as a computational method to predict how small molecules (ligands) interact with biological targets such as proteins, nucleic acids, or complexes. The fundamental purpose of docking is to replicate the molecular recognition process, with the objective of identifying the most favorable binding pose and estimating the binding affinity between ligand and receptor by optimizing geometrical and energetic compatibility.

This computational strategy addresses a key challenge in drug discovery, like for all search algorithms: exploring a vast chemical space with efficiency. Virtual screening via docking enables the rapid evaluation of large compound libraries, making it possible to filter down to promising candidates before moving to labor- and time-intensive experimental testing.

To achieve this, molecular docking employs two fundamental components:

(1) search algorithms that navigate the high-dimensional conformational space of ligand-protein interactions, answering the question about how and where a ligand binds, and

(2) scoring functions that rank proposed poses according to their predicted binding quality, typically approximating the Gibbs free energy of binding, though their numerical values are heuristic and may not be directly comparable to experimental affinities.

A critical challenge in molecular docking is accurately modeling molecular flexibility, particularly both the ligand's conformational diversity and the dynamic nature of protein receptors. Ligands can exist in multiple conformers due to rotatable bonds, as well as in different isomeric forms, including stereoisomers whose three-dimensional configurations may drastically influence binding outcomes. The common docking workflows attempt to cover ligand flexibility either by performing random translations and rotations of the molecule or by sampling its torsional angles [6].

Equally pressing is protein flexibility. Proteins are inherently dynamic, and their binding pockets can undergo conformational changes upon ligand binding, a process often neglected when the receptor is treated as rigid. This is a common approach, since it keeps computational expenses low. Parallelly, flexibility can be incorporated by diversifying the ligand like mentioned above. Allowing full receptor flexibility is usually computationally prohibitive, especially for high-throughput screening. A practical compromise is side-chain flexibility, which permits select residues to adopt different rotamer conformations in response to ligand binding.

### Defining the Binding Box

The binding box is a crucial element in molecular docking workflows. It establishes the three-dimensional region where the docking algorithm explores possible ligand binding orientations. Choosing an appropriate binding box is essential for the reliability and efficiency of docking results.

In instances where a co-crystallized ligand is present, the binding box is often centered on said ligand and expanded by a small margin. Otherwise, in instances where the crystallized structure does not contain a bound ligand, the utilization of an annotated active site can facilitate the approximation of a binding box, while allowing for a margin that accommodates ligand spacing.

When ligand-bound data are absent or binding sites are unknown, a so called blind docking case, tools such as CB-Dock detect promising cavities and estimate the box center and dimensions based on pocket geometry and ligand size [42].

GNINA [6], on the same note, provides an intuitive “autoboxing” feature that allows users to automatically define the docking search space based on input coordinates. This functionality enables a versatile docking setup, ranging from focused regions around known binding sites to more exploratory searches “all around” the receptor, giving ligands freedom to "choose" their binding location. Importantly, GNINA ensures that each side of the automatically generated box is at least as long as the maximum interatomic distance within the reference ligand.

If any dimension is shorter, that side is extended to match the ligand’s longest span, thereby ensuring the ligand can rotate freely within the box without suffering an “out-of-box” penalty, which is applied to poses found outside the defined docking volume [6].

### Search Functions

The core challenge in molecular docking is navigating the high-dimensional conformational space defined by the ligand’s six rigid-body degrees of freedom (three translations and three rotations) as demonstrated in Figure 4, along with numerous internal torsion angles, complicated further if receptor flexibility is considered.

Ein Bild, das Grün, Licht enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 4 - Composition of ligand search space in degrees of freedom: A) rotational and B) translational degrees of freedom

To manage this high-dimensional search space efficiently, docking engines employ a variety of strategic sampling approaches:

#### Systematic Search

The systematic search, also called exhaustive, grid or brute-force search, is a deterministic, grid-based approach, which discretizes rotations, translations, and torsions within practical limits, exhausting every possible conformation. They are deterministic but scale poorly with ligand flexibility. The more flexibility one wants to allow, the higher the computational cost that needs to be accounted for.

#### Stochastic Sampling

Approaches like Monte Carlo sampling and simulated annealing introduce randomness into conformational sampling. A initial pose is sampled at random and then accepted or rejected by energy criteria, in case of simulated annealing a temperature-based acceptance criterion. This enables efficient exploration of conformational space while avoiding local minima.

GNINA, for instance, uses multiple Monte Carlo chain runs in parallel. Each chain independently explores the ligand's conformational space. With each step a pose is randomly selected, either a translation, rotation, or torsion-angle change. The conformation is then mutated with torsional angles with a higher probability than the other poses. Subsequently, the minimal energy of the ligand is estimated [6].

#### Genetic Algorithms (GA)

The idea behind Genetic Algorithms is based on nature’s genetics and biological evolution. It represents ligands conformations (arrangement of translation, orientation, and conformation of the ligand) as state variables or, in the context of GAs, genes [43]. The application of noise serves to modify the genetic material, a process that can be interpreted as a form of "mutation". The fitness of an individual is assessed through the evaluation of the scoring function, selecting poses based on their interaction energy of the ligand with the protein. This process can be interpreted as natural selection. Better suited poses survive and reproduce, while poorly fitted ones die [43]. However, the system must also be subjected to adjustments and refinements. For instance, it is necessary to define the criteria for defining generations, establish clear guidelines for the removal of individuals, and specify the metric that will be used for the calculation of said fitness.

#### Incremental or fragment-based construction

Incremental or fragment-based construction is predicated on a discrete model and employs a tree-search technique for the incremental placement of the ligand into the active site [44].

The ligand is constructed within the pocket, with the selection of the base fragment being the first stage. The base fragment should contain as many potential interaction groups as possible. The next stage is placement and expansion; whereby further elements are added step by step. The process is modelled as a tree structure, with each node representing a possible placement of the ligand [44]. The algorithm takes into account the flexibility of the ligand through discrete conformation models, but the selection and placement of the base fragment is essential and can cause problems if, for example, the binding site is not precisely known or cannot be determined.

### Scoring Functions

After each searching step, a scoring function is applied to qualify the result. It is called for a metric that indicates if the positioning and sampling of poses suits the input or not. A scoring function maps a protein-ligand pose to a numerical value that should correlate with binding free energy. These scores fulfill three key roles in molecular docking [45]:

1. They determine the binding mode and site of a ligand on a protein.
2. They predict the absolute binding affinity between protein and ligand, allowing optimization, referring to improvement of tight binding for low-affinity hits or lead compounds that have been identified. This, in turn, increases the efficiency of optimization.
3. They can be applied in structure-based drug design, in which potential drug hits/leads for a given protein target are identified by searching large ligand databases.

Over time, several different approaches have been developed to implement scoring functions. These are typically grouped into four predominant classifications:

#### Force-field based Scoring

Also referred to as physics-based scoring, these rely on physics-based terms, mostly representing electrostatic interactions [46]. They are based on physical atomic interactions, including van der Waals (VDW) interactions, electrostatic interactions, and bond stretching, bending and torsional forces. Their parameters are derived from both experimental data and “ab initio quantum mechanical calculations” [45].

These functions are conceptually closest to ΔG but are computationally expensive and highly sensitive to the quality of the underlying force field [45].

#### Empirical scoring functions

These are fitted to experimental data. Empirical scoring functions may include counts of specific features as well as physics-inspired pairwise potentials [46]. They estimate the binding affinity of a complex based on a weighted sum of energy terms. Typically they are represented as a linear combination of physically motivated terms (hydrogen bonds, hydrophobic surface area, rotatable bond penalties, etc.), with coefficients fitted to experimental affinities [45].

An example of an empirical scoring function is Equation (1) ( [43], [45]), where each free energy term (Δ*Gi*) represents contributions such as van der Waals energy (vdw), electrostatics (elec), hydrogen bonds (hbond), conformational strain (conform), torsional restriction (tor), and (de-) solvation effects (sol). The coefficients (*Wi*)​ are determined by fitting binding affinity data of known protein-ligand complexes) ( [43], [45]).

|  |  |
| --- | --- |
|  | (1) |

These functions are much faster than physics-based scoring but are limited in generalizability since they are fitted to specific datasets [45]. The application of these function might also be dependent on the training set due to their ability to match the binding affinities of a small data set [45].

In GNINA, when CNN scoring is disabled (--cnn\_scoring=none), docking relies on an empirical scoring function optimized through fast, grid-based approximations [6]. The empirical function itself derives from the AutoDock Vina scoring function, which is a weighted sum of interaction terms, much like Equation (1) ( [43], [45]), consisting of a symmetric set of interaction functions of interatomic distances. The final score can be viewed as the sum of intermolecular and intramolecular contributions [47]. Accepted poses undergo local refinements using the exact scoring function and the final affinity values and scores for the pose are calculated [6].

#### Knowledge-based functions

Also known as statistical potentials, these derive interaction energies from observed frequencies and distances of atom-pair contacts across experimentally determined structures ( [45], [48]). More frequent contacts imply favorable interactions, converted into energy terms via inverse Boltzmann relations ( [45], [48]).

While computationally efficient, these functions can be biased by dataset composition [46]. The function also utilizes a so called reference state where interatomic interactions are zero. The challenge of deriving a knowledge-based function lies in deriving this reference state, which is according to Thomas and Dill [49] not accurately possible [45]. Therefor, dealing with knowledge-based functions also means dealing with approximating the reference state, which could be inaccurate [45]. Finding a good approximation is computationally costly and time consuming.

#### Machine-Learning based Scoring

The latest category, ML-based scoring functions, use algorithms such as random forests, support vector regression and convolutional neural networks (CNNs). These models learn complex, non-linear relationships from rich data representations without predefined functional forms. When well trained and validated, ML and deep learning scoring can outperform classical methods in binding affinity prediction and virtual screening [50].

If datasets and feature selection allow it, “family-specific” scoring functions could be tailored to particular protein families. ​Synthetic data, for example docking poses, can enhance training datasets, especially for targets with limited experimental data [50].

A notable example of modern deep-learning based scoring is GNINA, which integrates 3D convolutional neural networks (CNNs) to evaluate docking poses ( [6], [46]). GNINAs workflow allows user to choose which parts of the docking pipeline are assessed by CNN scoring.

If the CNN scoring is disabled, the used scoring function is essentially the same as the function utilized by AutoDock Vina [47], a empirical scoring function (see “Empirical scoring functions” above). When CNN scoring is enabled, GNINA can be applied at different stages of the docking pipeline to enhance accuracy. The “refinement” option uses the CNN after the Monte Carlo sampling process: once ligand poses are generated and selected by Monte Carlo chains, the CNN is employed to refine these poses toward local minima, after which the refined conformations are sorted according to their CNN-derived scores [6]. In contrast, the “all” option applies CNN scoring throughout the entire docking pipeline, including minimization within the Monte Carlo chains, refinement after sampling, and ranking of the final output conformations, thereby ensuring CNN-guided optimization at every stage of docking [6].

Additionally, GNINA offers a “hybrid refinement mode” (mix\_emp\_force), in which ligand poses are optimized using a linear combination of CNN gradient information and conventional non-CNN scoring force fields. This hybrid approach balances the strengths of physics-based and data-driven scoring, aiming to capture both physically motivated interactions and complex learned patterns in protein-ligand binding [6].

## Deep Learning-Based 3D Structure Prediction

Recent advances in machine learning now impact both ends of the structure-based pipeline: prediction of accurate protein folds and prediction/scoring of ligand poses. In the following section, three representative tools are examined: AlphaFold, Protenix and GNINA, highlighting their modes of operation and roles in a docking workflow.

Modern deep-learning methods can predict a protein’s 3D shape directly from its amino acid sequence. For example, the AlphaFold series ( [51], [52], [8]) of algorithms has revolutionized the field of protein structure predictions. These systems take a protein sequence as input, together with alignments of related sequences from databases (MSAs) and structural information about ligands and covalent bonds, and use very large neural networks to infer how the sequence folds into a stable 3D structure [8]. The output is an all-atom model of the protein. In practice, AlphaFold and similar systems (like the open-source tool Protenix) rely on vast training sets of known structures to learn typical folding patterns. For instance, AlphaFold 3 extends this approach to complexes: it can predict assemblies of proteins with bound ligands and nucleic acids in a single deep-learning framework [8].

Protenix is explicitly built as a PyTorch reproduction of this next-generation AlphaFold (so-called “AlphaFold3”), aimed at making a high-accuracy, trainable folding model openly available [9].

Once a structure is predicted, these programs also report confidence metrics to help users interpret the results. The predicted Local Distance Difference Test (pLDDT) score per-residue confidence score ranges from 0 to 100 and reflects how accurately local features of the model, such as backbone and side-chain positions, agree with what would be expected from an experimentally determined structure. These scores often vary along the chain, highlighting which regions are well-determined and which might be flexible or uncertain [53]. Beyond pLDDT, AlphaFold‑derived tools provide global metrics for whole complexes.

The predicted Template Modelling score (pTM) reflects the accuracy of the overall fold of the protein relative to an unknown true structure, analogous to a TM-score but derived from the prediction model rather than alignment to a known structure. A high pTM value (close to 1.0) indicates a reliable global fold prediction.

In contrast, the interface predicted Template Modelling score (ipTM) specifically measures the reliability of predicted interfaces between chains or between protein and ligands within the structure. ipTM values close to 1.0 therefore suggest a robust confidence in how the model positions interact with domains or binding partners [54].

Both metrics are based on the Template Modelling (TM) score. The TM-score is a widely used metric to quantify the overall distance between two three-dimensional structures. The Equation (2) incorporates the length of the native structure . Over the course of the length of the aligned residues () to the tempalte, i.e. residues that both structures have in common, the distances of each pair of aligned residues () is normalized over a distance scale [55].

|  |  |
| --- | --- |
|  | (2) |

The TM-score provides a normalized measure of structural similarity, ranging between (0, 1]. Scores above 0.5 generally indicate the same fold, while values approaching 1.0 suggest near-identical structures [55].

Predicting protein structure is only half the story for drug design; we also need to know how small molecules (ligands) fit into those structures. GNINA extends conventional docking by combining sampling algorithms (as in AutoDock Vina [47]) with convolutional neural networks that score 3D voxelised representations of the protein-ligand interface [6]. The CNN component is trained on experimental complexes and is intended to improve pose ranking and affinity prediction compared with classical scoring functions [6].

In practice, GNINA is utilized for the following purposes: The first part involves generating poses via standard search procedures. The second part entails re-scoring poses with learned potentials. The third part optionally performs ensemble docking against multiple receptor conformations. Therefore, GNINA occupies the intersection of pose generation and machine-learned affinity estimation [6].

# Materials and Methods

## System and Environment

All docking experiments were carried out on the CMCB Biocluster (Technische Universität Dresden).

The docking pipeline was executed with the preinstalled GNINA package (gnina master:fc5d7bb, built Oct 4 2024).

*Hardware resources used for the docking jobs:*

* 1 node with 128 CPU cores
* 256 GB RAM
* Maximum runtime per job: 2 hours

*Software environment:*

* GNINA [6]: for molecular docking.
* Protenix Server ( [9], [56]): deep learning-based structure prediction for protein-ligand complexes.
* PLIP (18 Dec 2024 build) [7]: interaction detection and interaction XML export.
* Python 3.10 with standard scientific libraries:
  + numpy, pandas
  + Biopython for the NeighborSearch algorithm
  + Matplotlib, seaborn: statistical visualization
  + Lxml for further utilization of XML files
  + RDKit [57]: molecular manipulation, fingerprint generation, and conformer sampling, MCS and RMSD calculation and Butin Clustering.
* PyMOL (Version 3.1.4.1) [13]: visualization, inspection, and separation of protein-ligand PDB files.
* OpenBabel (Version 3.1.1) [58]: conversion between molecular file formats, including PDB to PDBQT.

## Data Gathering and Preparation

### Biological and Structural Data

Human PPCDC is represented in the Protein Data Bank [59] only by a single crystal structure, called by the ID 1QZU ( [12], [15]), which forms the foundation of this study.

Ein Bild, das Zeichnung, Entwurf, Kunst, Strichzeichnung enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 5 - PPCDC holoenzyme (image from the RCSB PDB of 1QZU [12])

The crystal structure of 1QZU reveals a homo‑trimeric configuration, consisting of three identical subunits (see Figure 5) each roughly 204 amino acids in length, being part of the family of lyases, specifically the homo‑oligomeric flavin‑containing Cys decarboxylase (HFCD) superfamily (Uniprot Q96CD2 [11]). The structure consists of a flavin mononucleotide (FMN)-binding site and a substrate cavity, critical for enzymatic function.

In addition to being the only available human PPCDC structure, 1QZU poses important limitations. It lacks a bound substrate and the active site is not fully resolved, leaving uncertainties regarding substrate orientation and the geometry of the catalytic site. In order to prevail over this challenge, the study incorporates structural information from homologous proteins, computational predictions, and ligand analogues.

Several homologous structures provide complementary insights. The proteins 1E20 ( [60], [61]), 1MVN ( [62], [16]) and 1MVL ( [63], [16]) all derive from *Arabidopsis thaliana* AtHal3, the plant PPCDC homolog, which have been shown to share both sequence and structural similarity with human PPCDC. The structure of 1E20 is consistent with the wildtype sequence of AtHal3. In contrast, the other two structures correspond to mutations C175S, where the cysteine of the annealed binding site is substituted for a serine. Among these, 1MVN is of particular significance due to the presence of a bound ligand that exhibits a high degree of similarity to the natural substrate, pantothenoyl aminomethionethiol (PCO), an Ene-thiol reaction intermediate [16]. This allows it to serve as a proxy for defining the substrate binding site in human PPCDC.

The 6AIM ( [64], [65]) structure of the Cab2 mutant H337A, which represents the Phosphopantothenate-cysteine ligase (PPCS) CAB2 (step 2 of the Coenzyme A pathway) is, despite the absence of both functional and structural similarity, is the sole available PPCDC-related structure in the PDB that features a bound PPan-Cys. Consequently, it offers an additional reference point for substrate interaction patterns, since PPan-Cys is its natural product.

Combining the information derived from homologous structures with computational predictions of the original human sequence of PPCDC (PDB: 1QZU [12]) from Protenix [9] and structure-based alignments allows to define a “ground truth” active site model for human PPCDC.

This consensus model is based on the substrate binding confirmations given by the PCO-bound active site of 1MVN and refined using predicted human PPCDC conformations with the biological substrate PPan-Cys, enabling the identification of both substrate and FMN-binding geometries.

### Ligand Dataset and Preprocessing

The initial ligand dataset was compiled from the RCSB PDB [5] Ligands Chemical Component Dictionary, yielding 47,043 unique ligands with associated SMILES strings (OpenEye with stereo) and PDB chemical identifiers [66] [67]. This dataset encompasses a broad chemical space, including small molecules, cofactors, and larger complexes.

To improve computational tractability and avoid artifacts during docking, several filtering steps were applied.

First, ligands with SMILES strings shorter than 5 or longer than 70 characters were excluded to remove overly small fragments or atoms and excessively large complexes. Such extremes tend to produce unreliable docking scores, either due to insufficient interaction potential or steric overfitting [68].

Next, chemical redundancy was addressed by calculating pairwise Tanimoto similarities between ligands using Morgan fingerprints (radius 2 and 1024 bits) for each molecule. Those fingerprints are a computational representation of a molecule’s structure, encoded as a fixed-length binary vector.

Ligands with near-identical fingerprints (Tanimoto similarity ≥ 0.95) were removed. The Tanimoto coefficient (Equation (3)) is a metric used to calculate chemical similarity upon using Morgen fingerprints. It is defined as the ratio of the number of features common to two fingerprints A and B to the total amount of features [69], calculating all-to-all similarities for the entire dataset.

|  |  |
| --- | --- |
|  | (3) |

To further reduce dataset complexity, the Butina clustering algorithm [70] was applied with a cluster-cutoff of 0.65, grouping ligands with moderate chemical similarity. Butina clustering was accessed through RDkit [57]. Within each cluster, the ligand exhibiting the highest mean inner-cluster similarity was selected as the cluster representative.

For the mean inner-cluster similarity the mean Tanimoto similarity of each member to all other members was computed. The representative for each cluster is the molecule with the highest mean similarity, the cluster center.

## Docking Experiments

For the docking experiments, GNINA [6], a molecular docking software that extends the AutoDock Vina [71] framework with convolutional neural network (CNN)-based scoring functions, was employed.

As a basis for the docking process, the receptor was provided as a version of the 1QZU Protenix prediction that had been cleaned, with all ligands removed (--receptor cleaned\_protein.pdb). The ligands were supplied in SDF format (--ligand docked\_ligand.sdf). The Structure Data File (SDF) format is a widely used cheminformatics format that stores three-dimensional molecular structures along with atom connectivity and other chemical information. In the context of docking, the 3D coordinates contained in the SDF define how each ligand is initially positioned in space, providing GNINA with the necessary structural information to generate binding poses.

A central methodological challenge in this study was the cryptic nature of the substrate-binding pocket within PPCDC. Since the active site is not readily detected by conventional binding pocket prediction tools, the use of a predefined docking box was not feasible. Instead, the process relied on GNINA’s autoboxing feature, which defines a docking search space around a reference ligand or protein coordinates. The default parameter “--autobox\_add 4” was kept to ensure adequate sampling space around the pocket.

The option “--cnn\_scoring=none” disables GNINA’s CNN scoring functions, ensuring that only the classical Vina-like scoring was applied [6].

The autobox reference in this study was the entire protein structure including FMN and the predicted substrate (--autobox\_ligand protein\_with\_ligands.pdb). By using this complex, the docking box was automatically centered on the biologically relevant substrate-binding region, capturing the correct pocket geometry and steric environment.

Another important consideration was the presence of the FMN cofactor. All docking runs were performed with FMN retained in the structure, since this cofactor is essential for the *in vivo* conformational state of PPCDC. The inclusion of FMN ensured that steric hindrance effects and its role in shaping the binding cavity were faithfully represented during docking.

To establish a reference point for binding affinity, a separate docking experiment was carried out with the natural substrate (PPan-Cys). In this case, only a single docking run was performed, since the goal was not to explore alternative poses but to obtain a baseline affinity value against which docking results of other ligands could be compared.

For affinity evaluation, a separate single scoring method was applied to docked poses. This was achieved using GNINA’s “--score\_only” mode, which calculates a binding score for a given receptor-ligand complex without redocking. Initially, the PDB file containing both protein and ligand was separated into two files, a protein-only and, respectively, a ligand-only PDB file using the software program PyMOL [13]. The conversion of PDB files into pdbqt files was carried out utilizing the Open Babel software [58].

This procedure allows to calculate consistent affinity scores for downstream comparison. In particular, affinity scoring was only applied to the reference structures 1MVN and 1QZU Protenix prediction with PPan-Cys, which serve as structural benchmarks for the docking study.

## Post-Docking Analysis and Interaction Profiling

Following the docking simulations, the resulting ligand poses were subjected to a series of analyses to quantify binding quality and identify molecules most likely to inhibit PPCDC activity.

Additional filtering was performed to ensure biological relevance and also to minimize the number of relevant ligands.

Firstly, ligands were filtered using NeighborSearch of the *biopython* package [72] to exclude molecules that do not contact the target binding site, differentiating between the complete dataset and subsets such as 9 Å or 4 Å cutoff groups. Secondly, the affinity value from a separate docking experiment with only the substrate was used as a reference. Ligands exhibiting stronger predicted binding than the substrate were prioritized, while those with weaker binding were excluded.

To evaluate which ligands are more likely to be a good match for inhibiting the active site, the docking poses and interaction types need to be taken into account.

A useful and computationally simple way to evaluate docking results is through center of mass (COM) analysis. The COM provides a compact descriptor of a ligand’s position in the binding pocket. The COM represents the average position of all ligand atoms ( in Equation (4)), weighted by their atomic masses ( in Equation (4)) [73].

|  |  |
| --- | --- |
|  | (4) |

Heavy atoms such as oxygen, nitrogen, or sulfur shift the COM more strongly than lighter atoms like hydrogen. By calculating the distance between the COM of a docked ligand (*lig*) and that of a reference ligand (*ref*: PPan-Cys) (see Equation (5)), one obtains a straightforward measure of positional similarity in 3D space.

|  |  |
| --- | --- |
|  | (5) |

The center of mass analysis can be used to determine whether the docked ligands are actually located in the immediate proximity of the reference or substrate. This analysis can also provide indirect evidence regarding the location of the ligand in the binding site. To ascertain the congruence of the orientation with the reference, the data set was subjected to an MCS-RMSD analysis.

First, docking poses were evaluated using root-mean-square deviation (RMSD) calculations (Equation (6)). Since ligands in the dataset vary structurally, RMSD was computed using a maximum common substructure (MCS) approach. This method identifies the largest shared chemical scaffold between a ligand and a reference compound, and the RMSD is calculated based on the alignment of these shared atoms.

|  |  |
| --- | --- |
|  | (6) |

To quantify potential linear relationships between docking-derived metrics, including docking affinity, maximum common substructure (MCS) size, and MCS-RMSD values, Pearson correlation coefficients were calculated. The Pearson correlation coefficient (Equation (7)) measures the strength and direction of a linear relationship between two variables and is defined as [74]:

|  |  |
| --- | --- |
|  | (7) |

where ​ and ​ are individual observations of variables and , and and denote their respective means. Values of range from -1 to 1, with 1 indicating a perfect positive linear relationship, -1 a perfect negative linear relationship, and 0 indicating no linear correlation. These coefficients were used to assess the interrelationships among docking affinity, structural similarity (MCS), and geometric deviation (MCS-RMSD) across the ligand dataset.

To capture the functional relevance of ligand binding, docking poses were analyzed using the Protein-Ligand Interaction Profiler (PLIP). The tool automatically detects and classifies a range of interactions, including hydrogen bonds, hydrophobic contacts, salt bridges, π-stacking, and water bridges. For each ligand, an interaction fingerprint was derived (format “interaction type: residue number”), representing the set of contacts it forms with the protein.

These fingerprints were compared to reference interaction patterns obtained from homologous structures (PCO from the 1MVN mutant) and Protenix-predicted poses of the natural substrate (PPan-Cys from the 1QZU Protenix prediction). Similarity scores were computed according to:

|  |  |
| --- | --- |
|  | (8) |

where *L* is the set of interactions for the docked ligand and *R* is the reference set. Scores range from 0 to 1: a value of 1 indicates that the ligand reproduces all interactions observed in the reference, while 0 indicates no overlap. These scores provided a quantitative basis for prioritizing ligands that mimic the substrate’s binding mode.

Additionally, histograms of interaction types were generated to assess the distribution of contacts across the ligand set. This should give insight not only into interaction that are represented in the reference, but might also shed light on new types of interaction which may show a consensus within the docked ligand dataset.

To assess the practical availability of promising ligands, all candidate molecules were compared against the Enamine Hit Locator Diversity Library (460,160 small molecules readily available for experimental testing) [75]. For this purpose, SMILES strings of the docked ligands were converted into Morgan chemical fingerprints (512-bit length, radius = 2). Pairwise Tanimoto similarity was calculated between candidate ligands and the Enamine set.

# Results

The complete set of all input data, analysis scripts, and result files was made publicly accessible in a GitHub repository (see Tränkner, 2025 [76]).

## 3D Structure for Docking

Before performing molecular docking, it was essential to verify and compare the available three-dimensional structures of the target protein. Structural alignment allows assessment of the similarity between different experimental PDB entries as well as predicted models. For this purpose, the PDB Pairwise Structure Alignment tool [43] was used.

In the original human PPCDC structure (PDB ID: 1QZU, ochre-colored in Figure 6), a major portion of the binding site is absent, limiting its suitability for direct docking analysis. To compensate for this, the structure was compared with its *Arabidopsis thaliana* homologs: the wildtype (1E20, green in Figure 6) and the C175S mutant (1MVN, blue in Figure 6). Both homologs retain intact binding site regions, yet they differ significantly in their local topology. The *A. thaliana* wildtype binding site adopts a flexible, poorly defined loop conformation, whereas the mutant binding site displays a more rigid double-strand topology.

Ein Bild, das Screenshot, Grafikdesign, Grafiken, Cartoon enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 6 - PDB Pairwise Structural alignment of human and *At.* PPCDC chain A (created using Mol\* [77])

Table 1 summarizes the structural alignment between the target protein 1QZU and the two related entries (1E20 and 1MVN). The reported values include the root mean square deviation (RMSD), template modeling score (TM-score), sequence identity, number of aligned residues, and overall sequence length.

Table 1 - PDB Pairwise Structural alignment results of 1QZU, 1E20, 1MVN [5]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Entry | Chain | RMSD | TM-score | Identity | Aligned Residues | Sequence Length |
| [1QZU](https://www.rcsb.org/structure/1QZU) | A | - | - | - | - | 206 |
| [1E20](https://www.rcsb.org/structure/1E20) | A | 0.98 Å | 0.96 | 51 % | 158 | 209 |
| [1MVN](https://www.rcsb.org/structure/1MVN) | A | 0.86 Å | 0.95 | 51 % | 158 | 209 |

The high TM-scores (> 0.95) and low RMSD values (< 1.0 Å) in Table 1 confirm that 1QZU, 1E20, and 1MVN adopt very similar structural conformations despite moderate sequence identity (51 %).

One might expect a higher TM-score, contemplating that these proteins are homologs and have the exact same function in eukaryotes. Also, there is a notable difference between the RMSD score of 1E20, the wildtype structure, and 1MVN, the mutant structure. These differences are likely due to the conformational differences within the binding site, already noticeable in Figure 6.

To further probe conformational variability and construct a robust structural template for docking, different structural states were compared. Figure 7 shows conformations that were modeled using Protenix predictions [9] and experimental PDB structures, with and without cofactors FMN and PPan-Cys.

Ein Bild, das Grün, Cartoon, Kunst, Screenshot enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 7 - 1QZU Protenix prediction [9] binding site conformation with and without natural ligands (created using PyMOL [13])

|  |
| --- |
| ***(A)*** *pure 1QZU single chain prediction only based on the protein sequence,* ***(B)*** *native single chain protein conformation with natural bound FMN (blue) based on the 1QZU sequence, without a bound substrate or alternative,* ***(C)*** *1QZU single chain prediction based on the 1QZU sequence with PPan-Cys (orange), but without the stabilizing FMN,* ***(D)*** *a full single chain model of the 1QZU sequence prediction with bound FMN (blue) and natural substrate PPan-Cys (orange).* |

The comparison highlights the structural stabilization provided by FMN and the additional conformational shifts induced by substrate binding. These states form the basis for assessing ligand placement during docking.

In addition to experimentally resolved structures, Protenix was used to generate structural predictions of 1QZU in different ligand-bound states. The quality of these predictions was evaluated using several confidence metrics provided by the model. For the 1QZU Protenix prediction with FMN and PPan-Cys, the resulting values were pTM = 0.97 and ipTM = 0.99, both indicating very high structural confidence.

The 1QZU Protenix prediction model received an overall pLDDT score of 96.0 and with that takes part of the “high prediction confidence” category which are typically associated with very accurate models. When considered as a whole, these scores yield a precise prediction, both globally and locally.

To validate the quality of the Protenix prediction, structural comparisons were carried out using pairwise PDB alignments against experimentally determined human PPCDC (1QZU) and its *A. thaliana* homolog (C175S mutant, 1MVN). The results are summarized in Table 2 and visualized in Figure 8.

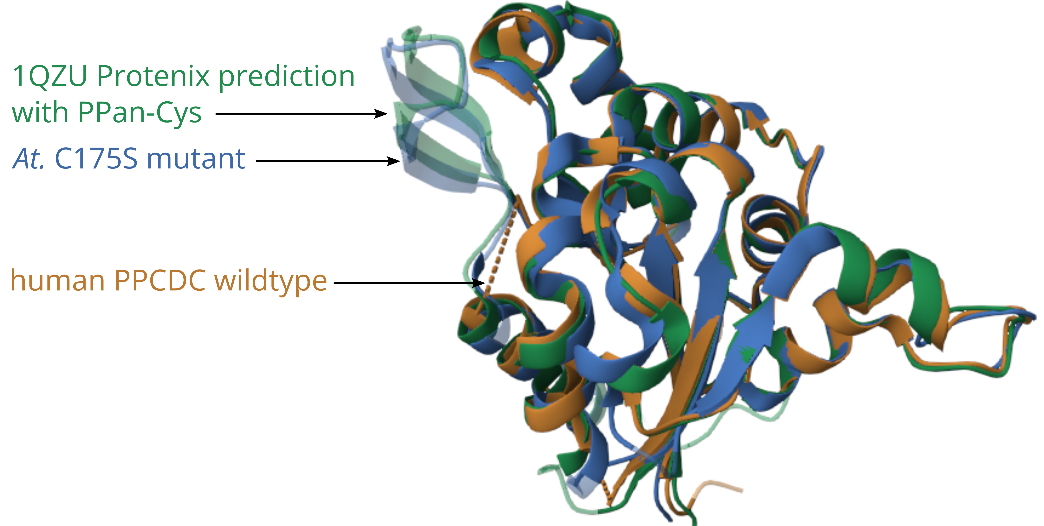


Figure 8 - PDB Pairwise Structural alignment of human PPCDC wildtype against 1QZU Protenix prediction with FMN and PPan-Cys and *At.* C175S mutant (created using Mol\* [77])

The alignment of the Protenix-predicted 1QZU structure with FMN and PPan-Cys shows a high similarity to the experimental model, with an RMSD of 0.84 Å and a TM-score of 0.97, indicating nearly identical global folding. The sequence identity is reported as 99%, with the only discrepancy arising from the first two residues of 1QZU, which are artifacts and not modeled in the PDB structure (as annotated in the original 1QZU entry [12]). In contrast, alignment with the *A. thaliana* mutant (1MVN) results in a lower sequence identity (51%) despite similarly good fold agreement (RMSD 0.86 Å, TM-score 0.96), reflecting evolutionary divergence.

Table 2 - PDB Pairwise Structural alignment results of 1QZU, 1MVN and the 1QZU protenix prediction with FMN and PPan-Cys [5]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Entry | Chain | RMSD | TM-score | Identity | Aligned Residues | Sequence Length |
| [1QZU](https://www.rcsb.org/structure/1QZU) | A | - | - | - | - | 206 |
| [1MVN](https://www.rcsb.org/structure/1MVN) | A | 0.86 Å | 0.96 | 51 % | 158 | 209 |
| Protenix pred. with FMN + PPan-Cys | A | 0.84 Å | 0.97 | 99 % | 158 | 206 |

Another approach for qualifying the Protenix prediction beyond global structural integrity is the evaluation of ligand pose similarity. In this context, the experimentally observed ligand PCO from the *A. thaliana* mutant structure (1MVN) was compared to the predicted PPan-Cys pose within the human PPCDC Protenix model.

The comparison reveals that the PCO pose aligns closely with the Protenix-predicted binding site, both in terms of spatial positioning and overall orientation. Quantitatively, The MCS-RMSD (see Equation (6)) analysis resulted in an RMSD of 0.625 Å, with 42 atoms of a total of 55 overlapping (maximum common substructure).

Ein Bild, das Screenshot, Grün, Grafiken, Grafikdesign enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 9 - PCO vs. 1QZU + PPan-Cys Protenix prediction pose comparison

As shown in Figure 9, the positions and orientations of the two ligands are in near-complete agreement. With the exception of the phosphate and terminal carboxyl groups, all other matching atoms are superimposed and share the same interacting amino acids within the binding site.

These findings strongly support the reliability of the Protenix-predicted binding mode, confirming that it captures the essential molecular recognition features observed experimentally and thus, it can be used as a template for the following docking process.

## Docking Pipeline

The multi-step pipeline, combining pre-selection of ligands, high throughput docking, and post-processing or ranking of ligands is visualized in Figure 10.

Ein Bild, das Text, Screenshot, Schrift, Handschrift enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 10 - Flowchart of the entire analysis, consisting of pre-processing, docking and post-processing

The computational workflow for ligand screening was designed to efficiently process a large chemical library while simultaneously maintaining a high level of chemical diversity and keeping computational expenses low.

Initially, the ligand dataset was pre-processed to reduce redundancy and focus on chemically diverse compounds. Chemical similarity was assessed using Tanimoto coefficients derived from molecular fingerprints, and compounds were grouped into clusters. From each cluster, representative molecules were selected through mean inner-cluster similarity, resulting in a smaller, more manageable subset of ligands for docking.

For docking, the high-confidence Protenix prediction of human PPCDC served as the receptor structure, and the filtered ligand dataset was used as input. Docking was performed with GNINA using the Autobox feature. The search for optimal ligand poses employed Monte Carlo sampling across the defined search space, while the empirical scoring function evaluated each pose to estimate binding affinity.

Post-processing of the docking results involved multiple complementary metrics to assess the quality and plausibility of the predicted binding interactions. Only ligands with predicted affinities better than the natural substrate were considered for further analysis. The center of mass (COM) of each ligand was calculated relative to the reference substrate to evaluate spatial proximity within the binding pocket. To quantify the chemical and geometric similarity of ligand poses, maximum common substructure (MCS)-based RMSD calculations were performed. Additionally, protein-ligand interactions were analyzed using PLIP, and a scoring scheme was applied to summarize interaction patterns across the dataset. Finally, Enamine's availability of ligands for experimental validation was determined to prioritize candidates for subsequent wet-lab screening.

## Finalized Dataset after Docking

Following the execution of the docking pipeline, a refined dataset was generated that consolidates all relevant structural and interaction-based information for downstream analysis. The finalized dataset is derived from the filtered ligand library, which had already been pre-processed by chemical similarity clustering to reduce redundancy and increase chemical diversity.

After completion of the docking pipeline, a total of 8,231 ligands were obtained with valid docking poses against the Protenix-predicted PPCDC structure. To refine the dataset toward biologically more plausible candidates, the ligands were filtered based on docking affinity relative to the substrate. Substrate affinity was determined by two independent strategies:

1) A Scoring-only GNINA run was executed to measure the binding affinity of the already existing binding results of the computational prediction via Protenix and the experimentally supported positioning.

The predicted substrate pose in the Protenix model received a docking score of -4.34 kcal/mol, while the crystallographic substrate pose in the *A. thaliana* mutant (1MVN) structure scored -3.82 kcal/mol.

2) An independent docking experiment was conducted, where the re-docking of the substrate yielded an affinity value of -6.94 kcal/mol. This value was selected as the more stringent threshold to filter the dataset.

Applying the latter criterion reduced the dataset to 3,764 ligands with docking affinities higher than the experimentally inspired cutoff.

Ein Bild, das Diagramm, Origami enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 11 - Violin plot of docking affinity distribution with applied NeighborSearch filters

To assess whether high-affinity ligands preferentially cluster near the binding site (residue 175 of the sequence), a NeighborSearch filter was applied. The violin plot (Figure 11) illustrates the distribution of docking scores before and after applying this spatial filter.

Ein Bild, das Text, Schrift, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 12 - Distribution of docking affinity against heavy atom count

Additionally, docking affinity was plotted against ligand heavy atom count (Figure 12), showing a broad distribution where larger ligands tend to achieve stronger affinities, though with considerable variance.

Ein Bild, das Diagramm, Screenshot, Text, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 13 - Histogram of MCS-RMSD values across all docked ligands

To evaluate structural similarity of the docked ligands to the reference substrate, MCS-RMSD (Maximum Common Substructure Root Mean Square Deviation) values were calculated. As demonstrated in Figure 13, the histogram shows that a substantial proportion of ligands is found within a cluster at a RMSD of less than 4 Å, while a secondary cluster is observed between 4-10 Å.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 14 - Docking affinity plotted against MCS-RMSD

When plotting affinity against MCS-RMSD (Figure 14), the same tendencies as in Figure 13 can be seen. Two clusters are observed: one corresponding to near-native pose (0-4 Å) with a wide affinity spread, and a second cluster with larger deviations (4-10 Å), generally still distributed between and affinity of -8 and -4 kcal/mol.

Pearson correlation analysis revealed no significant linear relationships between docking affinity, MCS size, and MCS-RMSD, with coefficients near zero (affinity vs. MCS: 0.015; affinity vs. RMSD: -0.065; MCS vs. RMSD: 0.096).

## PLIP Pattern Results

To further evaluate the docking results, protein-ligand interactions were analyzed using PLIP (Protein-Ligand Interaction Profiler), a tool that automatically detects and categorizes different interaction types, such as hydrogen bonds, hydrophobic contacts, salt bridges, π-stacking, and water bridges, between a ligand and its receptor. PLIP provides detailed information on interaction distances, geometry, and the specific residues involved.

For this study, interaction patterns were established by aggregating the types of interactions and amino acids in contact across all docked ligands. These patterns reveal the regions of the protein that are the most frequently involved in binding and allow assessment of whether ligand interactions are consistent with those of the natural substrate. Quantification with the established PLIP score enabled the ranking of ligands by providing an additional layer of evaluation that considers binding position and, in a way, orientation by mimicking the interaction types of the in-nature template, the substrate.

As a basis for PLIP interaction patterns, the predicted human PPCDC structure (1QZU) with substrate PPan-Cys (Figure 15.A) was compared to the experimentally determined ligand binding pose of the *At.* mutant protein (1MVN) with the reaction intermediate PCO (Figure 15.B). In this instance, some positions are identical, as evidenced by the alignment of LYS34 and MET183. PLIP indicates that for the Protenix prediction, LYS34 addresses the phosphate group, which is absent in PCO, thereby assuming the role of Val30 in terms of interaction. Moreover, a theoretically identical position exists (Val (PPan-Cys) or ALA (PCO) for 174). Due to the low sequence identity of 1QZU and 1MVN (see Table 1), the amino acids specifically addressed differ. However, given that both amino acids are non-polar and hydrophobic, it can be assumed that they serve analogous functions. Furthermore, there are residues that are not identical in both ligands but are similar, such as VAL30 and ALA31, which are located in approximately the same region.

Overall, the regions in question seem to fit for both interaction profiles, but only two of those residues are an exact match. Interestingly, most of the amino acids interacted with, are either non-polar and hydrophobic or positively charged and alkaline.

|  |  |  |
| --- | --- | --- |
| Ein Bild, das Text, Diagramm, Reihe, Screenshot enthält.  KI-generierte Inhalte können fehlerhaft sein. | Ein Bild, das Text, Diagramm, Reihe, Screenshot enthält.  KI-generierte Inhalte können fehlerhaft sein. |  |

Figure 15 - PLIP interaction patterns and overlaps of substrate binding in human PPCDC using Marvin JS [78]

|  |
| --- |
| ***(A)*** *Protenix prediction of PPCDC (1QZU) with PPan-Cys (9Z3)* [79]*,* ***(B)*** *Experimental PPCDC structure (1MVN) with PCO* [80]*,* ***(C)*** *Overlap of PLIP interaction patterns between PPan-Cys in 1QZU and 6AIM* [64]*.* |

As indicated in section 4.2.1, another starting point for modeling the binding site of 1QZU was to compare it with an existing binding site of another protein, namely 6AIM. The protein has no functional or structural similarities with PPCDC, apart from the both cathalytical reactions consisting of PPan-Cys. Due to the more centrally located binding site, significantly more interactions are recorded than in the 1QZU prediction. In addition, the interaction types and types of amino acids also differ.

Figure 15.C shows the interactions that both proteins (PPCDC and Cab2 mutant H337A) have with the molecule. The amino acids that have the same interaction types but differ in certain properties are highlighted in light blue. In the case of valine (PPCDC) and leucine (Cab2), the two amino acids are sterically different, but both are non-polar and uncharged and thus have the same physicochemical properties. In the case of lysine (PPCDC) and arginine (Cab2), the physicochemical properties of positive charge and alkaline characteristics are also retained.

As a final remark, it can be stated that the parallels between the previously binding interaction of 1QZU PPCDC as substrate and the experimentally determined interactions of PPan-Cys with PPCS as product are a positive indication of correct orientation within the binding pocket. However, the number of observed interactions is inadequate for drawing definitive conclusions regarding the binding pocket and the Protenix prediction.

PLIP was also applied to the complete docking dataset to investigate protein-ligand interactions across all 3,764 ligands with affinities higher than the substrate.

The analysis revealed that many previously known interaction clusters were preserved and expanded due to the presence of larger ligands with extended chemical groups. In comparison to PCO and PPan-Cys, these ligands can interact with additional residues surrounding the canonical binding sites, effectively enlarging the interaction footprint. For example, the regions identified in the Protenix prediction and 1MVN experimental structures (residue positions 30-34, 171-174, and 183) were reinforced, while new interaction sites emerged between residues of position 17-19, 50, 70, 72, 77, and 92, as well as clusters spanning positions 126-129 and 142-144 (Figure 17and Figure 18).

The later case is interesting, as it reveals novel regions of the protein accessed by ligands, particularly in areas not occupied by the substrate due to size constraints. Some of these interactions could be explained to the Autobox search employed by GNINA, which distributes ligands across the entire receptor surface, allowing exploration beyond the canonical binding pocket. To distinguish which interactions originated from ligands in proximity to the binding site versus those arising from more distant contacts, the PLIP interaction pipeline was applied to the NeighborSearch subgroup with a radius of 4 Å (see subchapter 5.3, Figure 19 and Figure 20). This filtering confirmed that known clusters (residues 29-37, 171-177, 181-183) remained, while additional residues in the intermediate regions (61 and 108) and a recurring cluster at 140-142 were also identified.

The quantitative scoring of PLIP interaction patterns (see Equation (8)) is an important bridge between qualitative interaction pattern comparison and quantitative filtering of candidates. It reflects how closely a candidate ligand reproduces the interaction network of the biological substrate. For this purpose, the PLIP score was calculated for each molecule of the PLIP analysis. Here, a distinction was made between similarity to the natural substrate (PPan-Cys), but only obtained through Protenix prediction, and experimentally determined interactions with a similarity to PCO, which is not part of the Coenzyme A pathway.

## Final Candidate Selection for Experimental Testing

After the computational screening and evaluation of ligand docking results, a final set of candidate molecules was selected for further experimental testing. The selection process was based on a combination of criteria to ensure that the chosen ligands not only demonstrated favorable binding properties but also maintained a level of biological relevance. Specifically, three major aspects were taken into account:

The tightness of the overall binding is given by the binding affinity, as estimated by GNINA docking scoring. These scores serve as an initial measure of the predicted ligand-protein interaction strength. It is important to note that the smaller the value, the more desirable it is.

Secondly, the binding site position is evaluated on a global scale. This evaluation is based on a short center-of-mass distance between the docked ligand and the substrate. It has been demonstrated that the efficacy of the measurement is directly proportional to the distance from the reference point.

Thirdly, pose stability and similarity are evaluated through MCS and MCS-RMSD calculations relative to the reference substrate pose. This is done to assess whether ligands adopted binding modes consistent with the known active site orientation. An elevated MCS and a decreased MCS-RMSD are indicative of enhanced performance.

The fourth point pertains to the interaction pattern similarity, which is determined by comparing PLIP-derived interaction profiles with those of the biological substrate (PPan-Cys in the Protenix prediction (PLIP\_1QZU) and PCO in the experimental reference structure (PLIP\_1MVN)). This approach ensured that the candidate ligands not only exhibited strong binding properties and reproduced the essential molecular interactions that are essential for catalytic functionality. The closer the score is to 1, the larger the overlay of interaction types with the reference.

The ranking was compiled from the individual criteria: COM, affinity and PLIP scores. Consequently, the entity occupying the uppermost position in the ranking consistently attained the leading placement across all subcategory rankings. The MCS and MCS-RMSD factors were not decisive factors in the ranking.

The application of these filters led to the refinement of the initial ligand library, resulting in a more manageable subset of final candidates. These molecules balance predicted binding strength with structural and functional plausibility, making them the most promising compounds for experimental validation.

To evaluate how robust the ranking is when using different subsets of criteria, pairwise comparisons were performed between alternative ranking schemes. The overlap of the top 100-ranked ligands and the correlation of their ordering between the new ranking and the all-criteria-ranking were used as indicators of robustness (see Table 3).

Table 3 - Overlap and Spearman Correlation Between Rankings Based on Different Criteria Combinations

|  |  |  |
| --- | --- | --- |
| Criteria | Overlap (%) | Spearman’s ρ |
| Affinity | 28 | -0.23 |
| COM | 22 | 0.20 |
| PLIP\_1MVN | 7 | -0.61 |
| PLIP\_1QZU | 15 | 0.31 |
| Affinity + COM | 54 | 0.50 |
| Affinity + PLIP\_1MVN | 42 | 0.18 |
| Affinity + PLIP\_1QZU | 54 | 0.19 |
| COM + PLIP\_1MVN | 35 | 0.24 |
| COM + PLIP\_1QZU | 37 | 0.39 |
| PLIP\_1MVN + PLIP\_1QZU | 29 | -0.03 |
| Affinity + COM + PLIP\_1MVN | 75 | 0.86 |
| Affinity + COM + PLIP\_1QZU | 63 | 0.60 |
| Affinity + PLIP\_1MVN + PLIP\_1QZU | 60 | 0.41 |
| COM + PLIP\_1MVN + PLIP\_1QZU | 44 | 0.34 |

Rankings based on only two criteria generally showed limited stability. For example, combining affinity and center-of-mass distance preserved about half of the top ligands (54 % overlap), with a moderate correlation in their order (ρ ≈ 0.50). By contrast, replacing affinity with PLIP interaction scores resulted in much weaker agreement: overlaps dropped to around 40-50 %, and correlations were close to zero. The least stable combinations were those relying solely on interaction-based scores, which shared less than one-third of the top ligands and showed almost no agreement in their ordering.

Including three criteria improved robustness considerably. In particular, the combination of affinity, COM distance, and PLIP scores from the 1MVN reference structure preserved 75 % of the top ligands and maintained a strong correlation (ρ ≈ 0.86). Other three-criteria sets, such as affinity + COM + PLIP(1QZU), also showed reasonably high robustness (63 % overlap, ρ ≈ 0.60). In contrast, subsets that excluded affinity or relied heavily on interaction scores were less stable.

The fifth criterion relates to the availability of wet-lab screening for further consideration. The top 200 matches were filtered against the Enamine Hit Locator Diversity Library. Similarity screening against the Enamine diversity library revealed that a subset of 147 molecules of the computationally prioritized ligands had close analogues (chemical similarity larger than 0.65) that are commercially available and could be considered for clinical screening.

## Reevaluation with Protenix

To complement the initial docking-based screening, a re-evaluation of the five top candidate ligands was performed using Protenix structure prediction in combination with GNINA rescoring.

Only a subset of top five candidates was chosen for this analysis (Table 4), since performing the procedure for the entire ligand library would have been computationally prohibitive.

Protenix predictions yielded highly confident structural models, with pTM values ranging from 0.88 to 0.93 and iPTM values between 0.81 and 0.96, indicating reliable folding and complex modeling (see Table 4). The average per-residue confidence (here with a pLDDT of approximately 90) confirmed that the binding regions were predicted with high structural accuracy.

When rescoring the Protenix-generated complexes using GNINA in affinity-only mode, notable differences were observed compared to the original docking scores. While the original GNINA docking produced favorable binding energies (-8.6 to -9.6 kcal/mol), the affinity-only rescoring yielded weaker or in some cases even unfavorable estimates (for example V2O with +0.61 kcal/mol).

Table 4 - Re-evaluation of the five best candidates using Protenix and GNINA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PDB-ID | WDB [81] | 87B [82] | A1LVI [83] | V2O [84] | YSH [85] |
| Original GNINA score (kcal/mol) | -9.07 | -8.68 | -9.6 | -9.19 | -8.6 |
| pTM | 0.88 | 0.90 | 0.88 | 0.88 | 0.93 |
| iPTM | 0.81 | 0.86 | 0.82 | 0.81 | 0.96 |
| pLDDT | 89.57 | 89.74 | 90.22 | 89.97 | 91.88 |
| GNINA scoring only affinity (kcal/mol) | -1.79 | -4.95 | -2.68 | 0.61 | -5.82 |

Notably, YSH achieved the best results at all levels, even though its original affinity was the lowest of all. A examination of the binding poses of the selected ligands reveals that YSH is the sole ligand exhibiting a pose analogous to that observed during docking. In all other instances, the ligand separated further from the binding site. In addition, this molecule is the only one in this subset for which the Protenix prediction shows a correctly formed binding loop. In all other cases, the general double-strand topology is present, but does not curve around the molecule (see Figure 16).

# Discussion

The present study combined structure-based virtual screening with interaction pattern analysis and structure prediction to identify promising candidate ligands for experimental validation. Using GNINA as the primary docking framework allowed efficient sampling of ligand poses and initial scoring, while complementary evaluation with PLIP and Protenix ensured a more robust assessment of binding modes and protein-ligand interactions.

The discussion should begin by reiterating that the proposed docking pipeline of this study, utilizing GNINA, was a de novo approach rather than a simple redocking. While the software was originally developed for re-docking or for providing complementary insights when no crystal structure is available [6], in this project it served a very different purpose: to generate entirely new data starting from scratch. In this case no experimental human PPCDC-ligand structure was available, so a way to overcome this issue was with a Protenix/AlphaFold predicted model of the PPCDC structure.

Moreover, it can be hypothesized that PPCDC’s active site is cryptic. It presumably only stabilizes when the cofactor FMN is already bound and the final change in conformation is induced by the substrate. The Protenix predictions demonstrated a transition in the binding site from a loose loop shape to a defined double-stranded conformation. It is noteworthy that in the model incorporating a substrate, the binding loop is fully formed. However, only in models with the substrate at the canonical binding site is the substrate surrounded by the double strand, suggesting additional interactions between the protein and the substrate.

With the knowledge of substrate binding interactions and FMN-protein interactions, two potential inhibition sites could be derived.

Firstly, the FMN binding site. Since FMN is essential for catalysis and protein stability, small molecules that mimic FMN’s isoalloxazine or ribityl phosphate moieties could competitively block cofactor binding or distort its orientation [16]. FMN-site inhibitors could thus trigger both catalytic arrest and conformational destabilization [17].

The second aspect to consider is the catalytic site. Compounds mimicking PPan-Cys (substrate analogs) or its transition state could outcompete the natural substrate, preventing decarboxylation.

The reason for not involving the FMN binding site in this study's experiments is that FMN is ubiquitous as a substance in the human body. As established by Lienhart, et al. (2013) [14], the human flavoprotein consists of 90 genes encoding for flavin-dependent proteins. More than 16 % of those flavoproteins utilize FMN as a coenzyme. It was not considered as reasonable to outperform FMN *in vivo*, thus the focus was actively set exclusively on the catalytic (active) site for inhibition. By concentrating on the active site, the chance of finding inhibitors that block the enzyme’s known activity was maximized. In other words, the goal was to find substrate‐competitive or transition-state‐mimicking hits, so the obvious target was the only site known to perform catalysis. This choice necessarily limited the search space, but it was scientifically justified by prior knowledge that PPCDC activity hinges on the catalytic core.

One potential complication is that the reference ligand used for pose evaluation, phosphopantothenoylcysteine (PCO) from the PDB entry 1MVN [62], comes from a structure where the catalytic cysteine (Cys175) is mutated to serine (C175S). Cysteine-to-serine is a conservative change: both have a polar sidechain (SH versus OH), are neutral at physiological pH, and have a similar steric size. The mutant serine cannot form disulfide bridges or covalent bond. However, structurally the pocket volume and polarity are little changed. Thus, it can be argued that docking poses generated against the C175S model would still approximate the true binding mode of PCO-like ligands. In practice the absence of the cysteine’s extra reactivity and disulfide potential means the model may slightly underestimate interactions that depend on a thiol group, but it should not introduce gross steric clashes or charge changes. In addition, PLIP patterns (Figure 15. A and B) show that the residue in question, position 175, is not directly targeted by either PCO or PPan-Cys.

In summary, the C175S mutation likely had minimal effect on pocket geometry or size, so using 1MVN-derived PCO as a reference remains a reasonable proxy for the wild‐type binding site.

In light of this observation, it should be noted that the template structure, in its final version, is subject to certain constraints. Protenix’s prediction appears to align well with known PPCDC homologs, but it is still a single, rigid computational model. However, the presence of a cryptic pocket requires the exact opposite: certain chains and loops must be shifted in order to reveal the binding site. Prior work on cryptic sites shows that molecular dynamics simulations often reveal pocket openings that are absent in the static “apo” structure [86]. Furthermore, ligand-induced pockets may not be apparent in single structures [86]. However, since the existing experimentally determined structures 1E20 and 1MVN exhibit precisely such dynamics (Figure 6) and could also be reproduced by Protenix prediction in various variations with and without cofactor and substrate (Figure 7), it can be assumed that the template used shows a snapshot of the true conformation.

On a related note, PPCDC’s active site is formed at the interface of its subunits. The literature is clear that Hal3 or PPCDC are homotrimers ( [60], [12]) and that the catalytic pocket lies between protomers [87]. In other words, residues from two adjacent monomers jointly form the binding site. This study’s docking, however, treated a single monomer at a time.

Consequently, two potential issues emerge: First, any residues contributed by the neighboring subunit would be missing - though none were registered by the PLIP patterns (Figure 15.B) -possibly leaving an artificially enlarged pocket or omitting key contacts. Second, steric constraints from the adjacent chain are absent. A hit ligand might sterically clash with the actual holoenzyme in reality, even if it fit the monomeric model.

In summary, docking to a monomer lacks quaternary context, so there may be slightly different pocket shape and volume than in the full enzyme. So, the single-chain docking within this study is necessarily an approximation of that multi-chain pocket.

Moving on to other decisions made during the preprocessing. In clustering and filtering the initial dataset, the Tanimoto fingerprint similarity was used as a chemical metric. The Tanimoto coefficient is a widely used measure of two-dimensional structural similarity that quantifies the overlap of binary fingerprints [88]. It is well established as a top-performing metric for chemical similarity (along with Dice and Cosine) [88].

In practice, Tanimoto is easy to compute and interprets shared substructure bits, so it is suitable for grouping compounds into clusters of related chemotypes and for eliminating (near-) duplicates. Tanimoto has some disadvantages, such as being biased by molecular size. However, it remains the “coefficient of choice” for fingerprint comparisons in virtual screening, apart from molecule size [89]. For this work, regarding filtering and Tanimoto-based clustering, this approach was beneficial for filtering out similar compounds.

An initial challenge was also to identify and target the binding pocket. Therefore, it was not possible to manually set a binding box for docking, even though the central residue was known. Instead, GNINA’s auto-box feature was used to let the software “find” the pocket region from the protein coordinates. This automated boxing avoids bias and helps deal with a partially hidden (cryptic) catalytic cleft that might not be obvious in the static model. To this end, autoboxing created a search grid large enough to capture any transiently accessible pocket around the predicted catalytic residues and also all over the protein to filter out ligands that favor different parts of the protein.

Another methodological point in docking is that GNINA’s CNN-based scoring was disabled, and the scoring process relied on the empirical (Vina-style) scoring functions [47]. GNINA’s convolutional neural networks are powerful for pose classification and affinity prediction when trained on known protein-ligand complexes, but they carry risks of overfitting or bias to the training data. In the de novo context of this study, an unusual target with a cryptic pocket, there is no guarantee that the CNN would be applicable. Since this conformational environment is poorly represented in available docking training sets, or structural data in general, CNN models may incorrectly prioritize poses that are geometrically valid but biologically implausible. By contrast, the empirical scoring (built on physics-inspired terms) is more transparent in including criteria that are actually known. This approach keeps the results interpretable: high-scoring compounds are those predicted to form favorable interactions (hydrogen bonds, hydrophobic contacts, etc.) rather than being favored by a black-box model.

Despite the utilization of filtering and clustering, a substantial number of molecules, amounting to nearly 4,000, remained to be evaluated post-docking.

The center-of-mass (COM) distance between the docked ligand and a reference molecule (the substrate bound to the Protenix prediction) was considered. A small COM distance indicates that the ligand sits in roughly the same part of the pocket as the reference molecule. Alone, COM distance does not capture orientation, only the coordinates of the ligand as a point mass. A ligand could have the right location but be flipped or rotated. For pose comparison an MCS-RMSD (“maximum common substructure RMSD”) was used rather than an all-atom RMSD. Conventional RMSD requires atom-to-atom correspondence across the whole ligand, which only makes sense when comparing identical or very closely related molecules. In the used library the docked compounds could be quite dissimilar, so aligning all atoms would pair unrelated atoms and give meaningless values.

Instead, MCS-RMSD first finds the largest matching substructure between two molecules, aligns that common core, and computes RMSD on those atoms. If chemically dissimilar ligands share a common substructure, MCS-RMSD is ametric to tell how similarly those cores are positioned. It therefore captures pose alignment of the shared structural motif while ignoring other parts of the ligand.

Both COM distance and MCS-RMSD alone do not ensure the ligand is in the correct pocket location. Combining them might give a more comprehensive view. MCS-RMSD shows how well the core scaffold of the hit overlaps the reference pose, while COM distance gives insight into how close the ligand’s centroid is to the reference centroid. Matches with both low MCS-RMSD and low COM distance thus share both orientation and spatial placement with the reference.

Ultimately, COM distance was used to filter out ligands far away from the desired position, MCS-RMSD can be regarded as providing supplementary information, particularly in cases where close matches to natural structures are of interest.

Importantly, MCS-RMSD was not included as a primary ranking criterium. The reason for that is similar to the goal of “scaffold hopping”, an approach for discovering structurally novel compounds by modifying a known active compound [90]. The goal was to find novel chemotypes that inhibit PPCDC, not just more substrate analogs. If the matches were ranked by a close common core structure, the final list would be biased towards known scaffolds, that are close to the natural substrate.

Finally, to partially address the uncertainty within the docking results, the top matches were re-evaluated as a separate Protenix prediction. The aim of this step was to 1) verify whether Protenix correctly reproduced the binding site in the expected region of the protein and to assess how the binding loop was formed, and 2) evaluate how well GNINA’s scoring function generalized when applied to these independently predicted protein-ligand complexes.

Interestingly, this re-evaluation highlighted a specific case: the ligand YSH [85] achieved slightly better results (in terms of both scoring and pose) in Protenix compared to other ligands that ranked above it in the GNINA scoring. At present, the exact reason for this discrepancy is unclear.

On the one hand, Protenix predicted a highly favorable binding mode in which YSH interacts not only with the protein side chains of the binding pocket but also aligns closely with the binding loop, thereby creating a tight and potentially stabilizing fit. On the other hand, it is equally conceivable that other ligands, despite being scored lower by Protenix, could still exhibit stronger or more favorable binding in an *in vitro* or *in vivo* context, particularly if their interactions extend more broadly onto or into the protein rather than relying on the relatively small and flexible binding loop. This experiment clearly shows that results obtained with Protenix prediction may differ from GNINA docking results. The examples YSH and 87B were fortunate examples that obtained relatively similar results with both GNINA and Protenix (+ GNINA rescoring). For the other examples from Chapter 5.6, the ligand determined by Protenix strayed further away from the binding box.

At this point, the question of which result to trust arises. In considering the five most successful matches, it is evident that all exhibited superior performance across all evaluation metrics. A preliminary assessment indicates that the Protenix predictions are satisfactory with respect to ipTM, pTM, and pLDDT. Nonetheless, with regard to the docking pose and position, the outcomes are suboptimal in certain instances.

Artificial intelligence-driven predictors like AlphaFold and Protenix have revolutionized structural biology by generating high-quality 3D models and serving as valuable starting hypotheses [91]. Notably, AlphaFold-style predictions are static snapshots that omit cellular context. These models do not take the presence of ligands (in some cases), water molecules, ions or covalent modifications into account, thus providing and incomplete structural model ( [91], [92]). In the context of docking experiments, Scardino, et al*.* found that AlphaFold-based receptor models performed consistently worse in virtual screening and docking tests than actual crystal structures, concluding that “this [accuracy] might not be enough to guarantee that AF models can be reliably used” for high-throughput docking [92].

In practice, this means we should treat any single AI‐predicted model as a testable hypothesis rather than a definitive answer. Confidence scores (like AlphaFold’s pLDDT) and ensemble approaches can help identify more reliable regions, but should not be stand-alone results.

Equally, one should also be critical of the docking affinity results. Docking programs output a predicted binding energy (scoring function) for each ligand pose. In AutoDock, this score is a pseudo-free-energy [47]: more negative values imply stronger predicted binding. However, one must treat these values with caution. Docking scoring functions are highly simplified (usually empirical or force-field based), so their numerical “affinities” are only rough estimates. In fact, it is well documented that docking scores often do not correlate quantitatively with experimental binding energies ( [93], [94]). Similarly, experimental affinities (like Kd, the dissociation constant, and IC₅₀, the half maximal inhibitory concentration) have their own uncertainties. Assay conditions, protein conformation, and measurement error are reported causes for reported affinities to vary. A recent analysis showed that even state-of-the-art free-energy calculations (±1 kcal/mol error) must contend with experimental variance; the combination of computational and experimental error can lead to “unexpected outcomes” when comparing affinities [95].

Therefore, docking scores are used to rank candidates rather than to predict the exact experimental parameters. Any promising hits from docking must ultimately be subjected to a final validation via biochemical assays or molecular dynamics.

# Outlook

The present study has illustrated how computational docking and deep-learning-based structure prediction can provide a rational starting point for the identification of potential inhibitors against PPCDC, a protein of medical relevance in the context of acute myeloid leukemia (AML). However, computational predictions alone cannot confirm functional inhibition or therapeutic relevance. Success can only be determined by experimental validation. Docking predictions, even when refined with advanced methods such as GNINA rescoring or Protenix structure prediction, remain hypotheses about possible interactions. They can provide valuable hints about which ligands might bind to a receptor, but they do not constitute evidence for such interactions. Therefore, the next essential step will be wet-lab validation through biochemical binding assays and high-throughput screening experiments. Only these assays can confirm whether the computationally identified ligands truly interact with PPCDC under physiological conditions. Beyond this, *in vivo* studies will be necessary to establish whether PPCDC inhibition has therapeutic relevance, or whether non-canonical functions of the enzyme dominate its role in disease. Such experiments will also clarify if direct inhibition of the active site is a viable approach or if alternative mechanisms should be targeted. In this sense, the present study provides a systematic basis for focusing laboratory resources, enabling a more informed starting point than random screening of unrelated compounds. In order to ensure that this transition is strategic rather than arbitrary, priority will be given to ligands with strong docking and interaction scores that also have commercially available analogs. In the domain of drug discovery, a similarity cutoff of 0.65 is widely acknowledged as a critical balance point, given its ability to maintain substantial structural and functional similarity to established reference compounds [96] and should be considered when selecting matches for further screening.

At the same time, the computational pipeline itself can be further improved and refined for future studies. One central point concerns the filtering of candidate molecules. In this work, clustering was applied with a very strict cutoff of 65 % dissimilarity, resulting in clusters of chemically diverse molecules. A promising alternative would be to focus on substances that share structural features with the natural substrate, an approach closer to scaffold hopping. Such compounds may exhibit comparable binding affinity but could block enzymatic conversion by occupying the binding site without being processed. A direct way to achieve this is to screen all compounds of the original dataset without filtering.

Additionally, future pipelines should account more explicitly for the structural characteristics of PPCDC. In this study, the binding box was defined broadly, including the entire protein with its cofactors FMN and PPan-Cys, due to the cryptic nature of the binding loop and to hopefully filter out ligands that might bind to other interfaces of the protein in *in vivo* experiments. However, the addition of 4 Å using the “autobox\_add” parameter may have created an excessively large search space, leading to matches far from the intended site. Reducing the search space should have timely advantages and should also result in tighter fits. Similarly, the GNINA “flex” option, which allows limited side-chain flexibility while keeping the protein backbone fixed, may offer a better way to model conformational changes associated with cryptic pockets and deserves further testing [6].

Further refinement of the ranking of docking results can also be achieved through improvements in the scoring procedure. The PLIP score developed in this study, which incorporates not only interaction counts but also the overlap with reference interaction patterns, proved useful in distinguishing promising candidates. An adjusted PLIP score has the potential to differentiate interactions based on their similarity to a reference pattern, as opposed to identity. At present, the score only covers interactions of the docked ligand with the same residue as the reference. To expand the scope, the score could be adjusted to encompass interactions with residues in the immediate vicinity of the reference residue.

Finally, while convolutional neural network (CNN) scoring was intentionally disabled during the main experiments in this thesis to avoid premature filtering, the reduced and refined ligand dataset obtained here may now be suitable for re-evaluation with GNINA’s deep learning scoring. Combining these refinements with targeted experimental screening would help validate both the computational predictions and the docking pipeline itself.

When considered as a whole, these steps describe a strategy in which computational and experimental methods progressively inform one another, ultimately determining the feasibility of using PPCDC inhibition as a therapeutic approach for AML.

Beyond the specific example of PPCDC in the context of AML therapeutics, several broader questions emerged during the investigation concerning the functionality of PPCDC’s two binding sites.

The binding loop of PPCDC appears disordered in the absence of ligand and becomes ordered upon ligand binding. This observation was made purely *in silico* through AlphaFold predictions and needs to be experimentally validated, for example through FRET (Förster (or Fluorescence) Resonance Energy Transfer) experiments [97]. Though, it motivates a more general hypothesis that small molecules or cofactors can convert intrinsically disordered or marginally ordered segments into more ordered, catalytically competent conformations.

This phenomenon has been observed in a number of examples. For example, zinc finger peptides require their cofactors (heme and zinc, respectively) to fold properly. Without these cofactors, the polypeptides remain unstructured. Here, cofactor binding not only stabilizes but also actively drives their folding toward the native conformation [98]. Additionally, studies on SH3 domains of phosphatidylinositol 3-kinase demonstrate how peptides induce folding under denaturing conditions, showing ligand concentration–dependent shifts between conformational selection and induced-fit mechanisms [99]. An important open question is whether this represents a general principle that extends beyond PPCDC. To test this, several approaches can be envisioned. First, one could compare ligand-annotated binding sites across the PDB with the corresponding regions in AlphaFold Database [100] models that lack bound ligands. Specifically, residues surrounding established binding pockets could be evaluated for low pLDDT scores, which would indicate structural uncertainty consistent with disorder in the ligand-free state. Second, AlphaFold 3 could be used for co-folding experiments, modeling the protein with and without the ligand. If the presence of the ligand increases local confidence metrics and produces a more ordered pocket, this would support the hypothesis that ligands induce order. Finally, this analysis could be extended across multiple proteins and cofactors to assess how widespread this phenomenon is, and whether certain sequence or structural features predict ligand-induced ordering.

Besides the active site, alternative targeting strategies should also be considered. PPCDC functions as a homotrimer, and the protomer-protomer interface may present an additional, less conventional target for inhibition. Targeting such interfaces could interfere with protein assembly or stability, offering a complementary therapeutic avenue to active-site inhibition. Since no trimeric structure of PPCDC is available in the PDB, AlphaFold 3 modeling could be used to reconstruct the oligomer. Furthermore, the presence or absence of FMN is likely to play a major role in PPCDC stability, as observed in the binding-site analysis. A critical question is whether the trimer can assemble at all in the absence of FMN, or whether FMN binding is essential for oligomer formation and stability.

A further interesting observation concerns ligand placement in the absence of FMN. In docking experiments, when FMN was not included, the predicted ligand occasionally bound to the FMN binding site rather than to the active site (Figure 7.C). This suggests a “second best is best” principle: in the absence of the cofactor, the model tends to favor the next most complementary pocket, even if it is normally reserved for a cofactor.

To validate this principle, one could select a PDB structure containing both FMN and a small-molecule ligand and perform AlphaFold 3 predictions with and without FMN. If, in the FMN-free condition, AlphaFold 3 misplaces the ligand into the FMN site, whereas adding FMN restores correct placement in the active site, this would support the hypothesis. More generally, proteins that bind cofactors such as FAD, NAD(P) or heme could be surveyed to test whether analogous behavior occurs when cofactors are omitted.

The availability of cofactors in cells must also be taken into account from a physiological perspective. Ligand competition for specific cofactors may not be significant *in vivo* if these sites are normally present at near-saturating levels. Nevertheless, the modeling bias remains relevant for *in silico* workflows, emphasizing on the need to include appropriate cofactors during structure prediction and docking to prevent misplacement and false positives.

# Summary

The increasing availability of deep learning approaches in structure prediction methods and advanced docking algorithms has significantly accelerated the process of *in silico* drug discovery. In this thesis, these tools were applied to the study of phosphopantothenoylcysteine decarboxylase (PPCDC), a protein with probable medical relevance in the context of acute myeloid leukemia (AML). Since PPCDC plays a key role in coenzyme A biosynthesis, and its inhibition may selectively interfere with metabolic pathways critical for leukemic cells, it represents a promising therapeutic target.

Two practical obstacles complicate structure-based inhibitor discovery for human PPCDC. The Protein Data Bank contains only a single human PPCDC crystal structure (PDB ID 1QZU), and that entry does not present a fully resolved substrate-bound active site. Parts of the catalytic loop are missing and the substrate pocket exhibits cryptic behavior, becoming fully formed only upon ligand binding or in particular oligomeric states.

The AlphaFold derivative Protenix was used to generate high-confidence protein structures, while docking experiments were performed with GNINA to explore possible ligand-protein interactions. Interaction patterns were analyzed with PLIP, and a scoring scheme was developed as an efficient way of ranking candidate ligands.

The results of the computational screening highlight several ligands that display promising interaction profiles with the PPCDC active site. These findings suggest multiple opportunities for inhibitor design, ranging from substrate-mimicking ligands that block catalysis to compounds that possibly interfere with protein assembly. At the same time, the study demonstrates the value of incorporating flexibility into docking protocols and ways to accommodate cryptic pockets.

By combining experimental screening with improved computational strategies, including refined ligand filtering, optimized docking protocols, PLIP interaction scoring, ligand pose and orientation evaluation, the identification of potent and specific PPCDC inhibitors can be pursued more effectively in the future.

Ultimately, while computational approaches can significantly narrow the chemical search space, experimental validation remains indispensable. The ligands identified in this thesis now form a prioritized set of candidates for *in vitro* binding assays and, subsequently, *in vivo* studies. In this way, the work presented here provides not only insights into PPCDC as a potential therapeutic target in AML but also a blueprint for integrating deep learning-based structure prediction, docking, and interaction analysis into a coherent pipeline for structure-guided drug discovery.

**List of appendices**

[Appendix 1: pLDDT results for the top five matches IX](#_Toc207988617)

[Appendix 2: PLIP results - complete dataset X](#_Toc207988618)

[Appendix 3: PLIP results - 4 Å dataset XIII](#_Toc207988619)

Appendix

Appendix 1: pLDDT results for the top five matches

Figure 16 summarizes the pLDDT results of the re-docking experiment of the top five matches after Candidate ranking (WDB [81], 87B [82], A1LVI [83], V2O [84], YSH [85]) done with Protenix.

Ein Bild, das Text, Screenshot, Karte enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 16 - comparison of pLDDT scores of the top five candidates after ranking

Appendix

Appendix 2: PLIP results - complete dataset

PLIP patterns of the complete dataset (without NeighborSearch) with reference of 1MVN interactions

Appendix

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 17 - PLIP patterns of the complete dataset (blue) with PCO-1MVN as reference (red) of hydrophobic interactions, hydrogen bonds, halogen bonds, metal complexes

Appendix

Ein Bild, das Text, Diagramm, Screenshot, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 18 - PLIP patterns of the complete dataset (blue) with PCO-1MVN as reference (red) of pi stakcs, pi cation interactions, water bridges, salt bridges

Appendix

Appendix 3: PLIP results - 4 Å dataset

PLIP patterns of the 4 Å dataset (NeighborSearch radius = 4) with reference of 1MVN interactions

Appendix

Ein Bild, das Text, Diagramm, Screenshot, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 19 - PLIP patterns of the 4 Å dataset (blue) with PCO-1MVN as reference (red) of hydrophobic interactions, hydrogen bonds, halogen bonds, metal complexes

Appendix

Ein Bild, das Text, Screenshot, Diagramm, Rechteck enthält.

KI-generierte Inhalte können fehlerhaft sein.

Figure 20 - PLIP patterns of the 4 Å dataset (blue) with PCO-1MVN as reference (red) of pi stakcs, pi cation interactions, water bridges, salt bridges

# Bibliography

|  |  |
| --- | --- |
| [1] | P. Garg, G. Singhal, P. Kulkarni, D. Horne, R. Salgia and S. S. Singhal, "Artificial Intelligence–Driven Computational Approaches in the Development of Anticancer Drugs," *cancers,* vol. 16, no. 22, p. 3884, 2024. |
| [2] | B. Shaker, S. Ahmad, J. Lee, C. Jung and D. Na, "In silico methods and tools for drug discovery," *Computers in Biology and Medicine,* vol. 137, p. 104851, 2021. |
| [3] | A. V. Sadybekov and V. Katritch, "Computational approaches streamlining drug discovery," *Nature,* vol. 616, p. 673–685, 2023. |
| [4] | D. B. Kitchen, H. Decornez, J. R. Furr and J. Bajorath, "Docking and scoring in virtual screening for drug discovery: methods and applications," *Nature Reviews Drug Discovery,* vol. 3, p. 935–949, 2004. |
| [5] | S. Bittrich, J. Segura, J. M. Duarte, S. K. Burley and Y. Rose, "RCSB protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments," *Bioinformatics,* vol. 40, no. 6, p. btae370, 2024. |
| [6] | A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri and D. R. Koes, "GNINA 1.0: molecular docking with deep learning," *Journal of Cheminformatics,* vol. 13, no. 43, 2021. |
| [7] | P. Schake, S. N. Bolz, K. Linnemann and M. Schroeder, "PLIP 2025: introducing protein–protein interactions to the protein–ligand interaction profiler," vol. 53, no. W1, 2025. |
| [8] | J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O’Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature,* vol. 630, p. 493–500, 2024. |
| [9] | ByteDance AML AI4Science Team, "Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction," *bioRxiv,* 2025. |
| [10] | M. Daugherty, B. Polanuyer, M. Farrell, A. Lykidis, V. de Crécy-Lagard and A. Osterman, "Complete Reconstitution of the Human Coenzyme A Biosynthetic Pathway via Comparative Genomics," *Journal of Biological Chemistry,* vol. 277, no. 24, pp. 21431 - 21439, 2002. |
| [11] | The UniProt Consortium, "UniProt: the Universal Protein Knowledgebase in 2025," *Nucleic Acids Res.,* vol. 53, p. D609–D617, 2025. |
| [12] | N. Manoj and S. Ealick, "crystal structure of human phosphopantothenoylcysteine decarboxylase," 2004. [Online]. Available: https://doi.org/10.2210/pdb1QZU/pdb. [Accessed 14 08 2025]. |
| [13] | Schrödinger, LLC, *The PyMOL Molecular Graphics System, Version~3.1.4.1,* 2015. |
| [14] | W.-D. Lienhart, V. Gudipati and P. Macheroux, "The human flavoproteome," *Archives of Biochemistry and Biophysics,* vol. 535, no. 2, p. 150–162, 2013. |
| [15] | N. Manoj and S. Ealick, "Unusual space-group pseudosymmetry in crystals of human phosphopantothenoylcysteine decarboxylase," *Acta Crystallographica Section D, Structural Biology,* vol. D59, pp. 1762-1766, 2003. |
| [16] | S. Steinbacher, P. Hernández-Acosta, B. Bieseler, M. Blaesse, R. Huber, F. A. Culiáñez-Macià and T. Kupke, "Crystal Structure of the Plant PPC Decarboxylase AtHAL3a Complexed with an Ene-thiol Reaction Intermediate," *Journal of Molecular Biology,* vol. 327, no. 1, pp. 193-202, 2003. |
| [17] | I. Bravo-Alonso, M. Morin, L. Arribas-Carreira, M. Álvarez, C. Pedrón-Giner, L. Soletto, C. Santolaria, S. Ramón-Maiques, M. Ugarte, P. Rodríguez-Pombo, J. Ariño, M. Á. Moreno-Pelayo and B. Pérez, "Pathogenic variants of the coenzyme A biosynthesis-associated enzyme phosphopantothenoylcysteine decarboxylase cause autosomal-recessive dilated cardiomyopathy," *Journal of inherited hetabolic desease,* vol. 46, no. 2, pp. 261-272, 2022. |
| [18] | A. Ambinder and A. Gerds, "Acute Myeloid Leukemia Treatment (PDQ®)–Health Professional Version," National Cancer Institute, 14 03 2025. [Online]. Available: https://www.cancer.gov/types/leukemia/hp/adult-aml-treatment-pdq. [Accessed 19 08 2025]. |
| [19] | A. Vakiti, S. B. Reynolds and P. Mewawalla, "Acute Myeloid Leukemia," in *StatPearls [Internet]*, Treasure Island (FL), StatPearls Publishing, 2025. |
| [20] | National Library of Medicine, "Acute Myeloid Leukemia," MedlinePlus [Internet], 27 11 2023. [Online]. Available: https://medlineplus.gov/acutemyeloidleukemia.html. [Accessed 19 08 2025]. |
| [21] | Leukemia & Lymphoma Society, "Acute Myeloid Leukemia - Treatment," [Online]. Available: https://www.lls.org/leukemia/acute-myeloid-leukemia/treatment. [Accessed 19 08 2025]. |
| [22] | C. Cavestro, D. Diodato, V. Tiranti and I. Di Meo, "Inherited Disorders of Coenzyme A Biosynthesis: Models, Mechanisms, and Treatments," *International Journal of Molecular Sciences,* vol. 24, no. 6, p. 5951, 2023. |
| [23] | J. D. Munzar, A. Ng, M. Corrado and D. Juncker, "Complementary Oligonucleotides Regulate Induced Fit Ligand Binding in Duplexed Aptamers," *Chemical Science,* vol. 8, no. 3, 2016. |
| [24] | E. Fischer, "Einfluss der Configuration auf die Wirkung der Enzyme," *Berichte der deutschen chemischen Gesellschaft,* vol. 27, no. 3, pp. 2985-2993, 1894. |
| [25] | D. E. Koshland, Jr., "Application of a Theory of Enzyme Specificity to Protein Synthesis," *PNAS,* vol. 44, no. 2, pp. 98-104, 1958. |
| [26] | F. Paul and T. R. Weikl , "How to Distinguish Conformational Selection and Induced Fit Based on Chemical Relaxation Rates," *PLoS Computational Biology,* vol. 12, no. 9, p. e1005067, 2026. |
| [27] | G. G. Hammes, Y.-C. Chang and T. G. Oas, "Conformational selection or induced fit: A flux description of reaction mechanism," *PNAS,* vol. 106, no. 33, pp. 13737-13741, 2009. |
| [28] | T. Pantsar and A. Poso, "Binding Affinity via Docking: Fact and Fiction," *Molecules,* vol. 23, no. 8, p. 1899, 2018. |
| [29] | R. Ferreira de Freitas and M. Schapira, "A systematic analysis of atomic protein–ligand interactions in the PDB," *Medicinal Chemistry Communications,* vol. 8, pp. 1970-1981, 2017. |
| [30] | S. Salentin, V. J. Haupt, S. Daminelli and M. Schroeder, "Polypharmacology rescored: Protein–ligand interaction profiles for remote binding site similarity assessment," *Progress in Biophysics and Molecular Biology,* vol. 116, no. 2-3, pp. 174-186, 2014. |
| [31] | D. Xu, C.-J. Tsai and R. Nussinov, "Hydrogen bonds and salt bridges across protein-protein interfaces," *Protein Engineering,* vol. 10, no. 9, pp. 999-1012, 1997. |
| [32] | C. Bissantz, B. Kuhn and M. Stahl, "A Medicinal Chemist’s Guide to Molecular Interactions," *Journal of Medicinal Chemistry - Perspective,* vol. 53, p. 5061–5084, 2010. |
| [33] | P. Politzer, J. S. Murray and T. Clark, "Halogen bonding and other σ-hole interactions: a perspective," *Physical Chemistry Chemical Physics,* vol. 15, pp. 11178-11189, 2014. |
| [34] | N. K. Shinada, A. G. de Brevern and P. Schmidtke, "Halogens in Protein–Ligand Binding Mechanism: A Structural Perspective," *Journal of Medicinal Chemistry,* vol. 62, no. 21, pp. 9341-9356, 2019. |
| [35] | G. Job and R. Rüffler, Physikalische Chemie - Eine Einführung nach neuem Konzept mit zahlreichen Experimenten (2. Auflage), Wiesbaden, Germany: Springer Spektrum, 2021. |
| [36] | J. Monod, J. Wyman and J.-P. Changeux, "On the Nature of Allosteric Transitions: A plausible model," *Journal of olecular Biology,* vol. 12, no. 1, pp. 88-118, 1965. |
| [37] | G. N. Stephanopoulos, A. A. Aristidou and J. Nielsen, "CHAPTER 5 - Regulation of Metabolic Pathways," in *Metabolic Engineering - Principles and Methodologies*, Academic Press, 1998, pp. 147-202. |
| [38] | A. McPherson and J. A. Gavira , "Introduction to protein crystallization," *Acta Crystallographica F,* vol. 70, no. Pt 1, p. 2–20, 2013. |
| [39] | Y. Hu, K. Cheng, L. He, X. Zhang, B. Jiang, L. Jiang, C. Li, G. Wang, Y. Yang and M. Liu, "NMR-Based Methods for Protein Analysis," *Analytical Chemistry,* vol. 93, no. 4, pp. 1866-1879, 2021. |
| [40] | F. W. Muskett, "Sample Preparation, Data Collection and Processing," in *Protein NMR Spectroscopy: Practical Techniquesand Applications*, West Sussex, United Kingdom, John Wiley & Sons, Ltd, 2011, p. 5. |
| [41] | X.-c. Bai, G. McMullan and S. H. W. Scheres, "How cryo-EM is revolutionizing structural biology," *Trends in Biochemical Sciences,* vol. 40, no. 1, pp. 49-57, 2015. |
| [42] | Y. Liu, M. Grimm, W.-t. Dai, M.-c. Hou, Z.-X. Xiao and Y. Cao, "CB-Dock: a web server for cavity detection-guided protein–ligand blind docking," *Acta Pharmacologica Sinica,* vol. 41, no. 1, p. 138–144, 2019. |
| [43] | G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, "Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function," *Journal of Computational Chemistry,* vol. 19, no. 14, pp. 1639-1662, 1998. |
| [44] | M. Rarey, B. Kramer, T. Lengauer and G. Klebe, "A Fast Flexible Docking Method using an Incremental Construction Algorithm," *Journal of Molecular Biology,* vol. 261, no. 3, pp. 470-489, 1996. |
| [45] | S.-Y. Huang, S. Z. Grinter and X. Zou, "Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions," *Physical Chemistry Chemical Physics,* vol. 12, no. 40, pp. 12899-12908, 2024. |
| [46] | J. Sunseri and D. R. Koes, "Virtual Screening with Gnina 1.0," *Molecules,* vol. 26, no. 23, p. 7369, 2021. |
| [47] | O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *Journal of Computational Chemistry,* vol. 31, no. 2, pp. 249-461, 2010. |
| [48] | X.-Y. Meng, H.-X. Zhang, M. Mezei and M. Cui, "Molecular Docking: A powerful approach for structure-based drug discovery," *Current Computer-Aided Drug Design,* vol. 7, no. 2, p. 146–157, 2012. |
| [49] | P. D. Thomas and K. A. Dill, "Statistical Potentials Extracted From Protein Structures: How Accurate Are They?," *Journal of Molecular Biology,* vol. 257, no. 2, pp. 457-469, 1996. |
| [50] | Q. U. Ain, A. Aleksandrova, F. D. Roessler and P. J. Ballester, "Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening," *WIREs Computational Molecular Science,* vol. 5, no. 6, pp. 405-459, 2015. |
| [51] | J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature,* vol. 596, pp. 583-589, 2021. |
| [52] | J. Fleming, P. Magana, S. Nair, M. Tsenkov, D. Bertoni, I. Pidruchna, M. Q. Lima Afonso, A. Midlik, U. Paramval, A. Žídek, A. Laydon, O. Kovalevskiy, J. Pan, J. Cheng, Ž. Avsec, C. Bycroft, L. H. Wong, M. Last, M. Mirdita, M. Steinegger, P. Kohli, M. Váradi and S. Velankar, "AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities," *Journal of Molecular Biology,* vol. 437, no. 15, 2025. |
| [53] | EMBL-EBI, "pLDDT: Understanding local confidence," EMBL-EBI, 06 2025. [Online]. Available: https://www.ebi.ac.uk/training/online/courses/alphafold/inputs-and-outputs/evaluating-alphafolds-predicted-structures-using-confidence-scores/plddt-understanding-local-confidence. [Accessed 21 08 2025]. |
| [54] | EMBL-EBI, "Confidence scores in AlphaFold-Multimer," EMBL-EBI, 06 2025. [Online]. Available: https://www.ebi.ac.uk/training/online/courses/alphafold/inputs-and-outputs/evaluating-alphafolds-predicted-structures-using-confidence-scores/confidence-scores-in-alphafold-multimer. [Accessed 21 08 2025]. |
| [55] | Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins,* vol. 57, pp. 702-710, 2004. |
| [56] | Bytedance Inc., "Protenix Server," 26 12 2024. [Online]. Available: https://protenix-server.com. [Accessed 27 08 2025]. |
| [57] | G. Landrum, "RDKit," Open-source cheminformatics, 03 05 2025. [Online]. Available: https://www.rdkit.org. [Accessed 18 08 2025]. |
| [58] | N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, "Open babel: An open chemical toolbox," *Journal of Cheminformatics,* vol. 3, no. 33, 2011. |
| [59] | H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research,* vol. 28, pp. 235-242, 2000. |
| [60] | A. Albert, M. Martinez-Ripoll, A. Espinosa-Ruiz, L. Yenush, F. Culianez-Macia and R. Serrano, "The FMN binding protein AtHal3," 2000. [Online]. Available: https://www.rcsb.org/structure/1E20. [Accessed 14 08 2025]. |
| [61] | A. Albert, M. Martinez-Ripoll, A. Espinosa-Ruiz, L. Yenush, F. Culianez-Macia and R. Serrano, "The X-Ray Structure of the Fmn-Binding Protein Athal3 Provides the Structural Basis for the Activity of a Regulatory Subunit Involved in Signal Transduction," *Structure,* vol. 8, no. 9, pp. 961 - 969, 2000. |
| [62] | S. Steinbacher, P. Hernandez-Acosta, B. Bieseler, M. Blaesse, R. Huber, F. Culianez-Macia and T. Kupke, "PPC decarboxylase mutant C175S complexed with pantothenoylaminoethenethiol," 2003. [Online]. Available: https://www.rcsb.org/structure/1MVN. [Accessed 14 08 2025]. |
| [63] | S. Steinbacher, P. Hernández-Acosta, B. Bieseler, M. Blaesse, R. Huber, F. A. Culiáñez-Macià and T. Kupke, "PPC decarboxylase mutant C175S," 2003. [Online]. Available: https://www.rcsb.org/structure/1MVL. [Accessed 14 08 2025]. |
| [64] | P. Zheng and Z. Zhu, "Cab2 mutant H337A complex with phosphopantothenate-cysteine," 2019. [Online]. Available: https://www.rcsb.org/structure/6AIM. [Accessed 14 08 2025]. |
| [65] | P. Zheng, Z. Zhu, M. Khan, H. Liu, Y. Jin, J. Yue, Y. Gao, M. Teng, Z. Zhu and L. Niu, "Crystallographic Analysis of the Catalytic Mechanism of Phosphopantothenoylcysteine Synthetase from Saccharomyces cerevisiae," *Journal of Molecular Biology,* vol. 431, no. 4, pp. 764-776, 2019. |
| [66] | wwpdb.org, "Chemical Component Dictionary," RCSB PDB, [Online]. Available: https://www.wwpdb.org/data/ccd. [Accessed 15 08 2025]. |
| [67] | J. D. Westbrook, C. Shao, Z. Feng, M. Zhuravleva, S. Velankar and J. Young, "The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank Bioinformatics," *Bioinformatics,* vol. 31, no. 8, p. 1274–1278, 2015. |
| [68] | A. Wilantho, S. Tongsima and E. Jenwitheesuk, "Pre-docking filter for protein and ligand 3D structures," *Bioinformation,* vol. 3, no. 5, p. 189–193, 2008. |
| [69] | D. J. Rogers and T. T. Tanimoto, "A Computer Program for Classifying Plants," *Science,* vol. 132, no. 3434, pp. 1115-1118, 1960. |
| [70] | D. Butina, "Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets," *Journal of Chemical Information and Computer Sciences,* vol. 39, no. 4, p. 747–750, 1999. |
| [71] | J. Eberhardt, D. Santos-Martins, A. F. Tillack and S. Forli, "AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings," *Computational Chemistry,* vol. 61, no. 8, pp. 3891-3898, 2021. |
| [72] | P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox , A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. L. de Hoon, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics,* vol. 25, no. 11, p. 1422–1423, 2009. |
| [73] | D. Stojanov, "On the in silico application of the center-of-mass," *Gene & Protein in Disease,* vol. 3, no. 1, p. 2657, 2024. |
| [74] | M. M. Mukaka, "A guide to appropriate use of Correlation coefficient in medical research," *Malawi Medical Journal,* vol. 24, no. 3, p. 69–71, 2012. |
| [75] | Enamine Ltd., "Diversity Libraries," [Online]. Available: https://enamine.net/compound-libraries/diversity-libraries. [Accessed 02 09 2025]. |
| [76] | M. Tränkner, "ppcdc-docking-pipeline," 2025. [Online]. Available: https://github.com/m-traenkn/ppcdc-docking-pipeline. [Accessed 03 09 2025]. |
| [77] | D. Sehnal, S. Bittrich, M. Deshpande, R. Svobodová, K. Berka, V. Bazgier, S. Velankar, S. Burley, J. Koča and A. Rose, "Mol\* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures," *Nucleic Acids Research,* vol. 49, pp. W431-W437, 2021. |
| [78] | Chemaxon, "Marvin JS," [Online]. Available: https://docs.chemaxon.com/display/docs/marvin-js\_index.md. [Accessed 18 08 2025]. |
| [79] | RCSB PDB, "9Z3 - N-[(2R)-2-hydroxy-3,3-dimethyl-4-(phosphonooxy)butanoyl]-beta-alanyl-L-cysteine," [Online]. Available: https://www.rcsb.org/ligand/9Z3. [Accessed 18 08 2025]. |
| [80] | RCSB PDB, "PCO - 2,4-DIHYDROXY-N-[2-(2-MERCAPTO-VINYLCARBAMOYL)-ETHYL]-3,3-DIMETHYL-BUTYRAMIDE," [Online]. Available: https://www.rcsb.org/ligand/PCO. [Accessed 18 08 2025]. |
| [81] | RCSB PDB, "WDB - (~{E})-3-[2-[(3-chlorophenyl)methyl]-4-oxidanylidene-3~{H}-quinazolin-7-yl]prop-2-enoic acid," [Online]. Available: https://www.rcsb.org/ligand/WDB. [Accessed 26 08 2025]. |
| [82] | RCSB PDB, "87B - N-(2-cyclobutyl-1H-1,3-benzodiazol-5-yl)benzenesulfonamide," [Online]. Available: https://www.rcsb.org/ligand/87B. [Accessed 26 08 2025]. |
| [83] | RCSB PDB, "A1LVI - 3-[3-[[4-(pyridin-2-yloxymethyl)phenyl]methyl]-1,2-oxazol-5-yl]pyridin-2-amine," [Online]. Available: https://www.rcsb.org/ligand/A1LVI. [Accessed 26 08 2025]. |
| [84] | RCSB PDB, "V2O - 5-[1-fluoro-3-hydroxy-7-(3-hydroxy-3-methylbutoxy)naphthalen-2-yl]-1lambda~6~,2,5-thiadiazolidine-1,1,3-trione," 02 08 2023. [Online]. Available: https://www.rcsb.org/ligand/V2O. [Accessed 26 08 2025]. |
| [85] | RCSB PDB, "YSH - 1-[3-CYANO-4-(NEOPENTYLOXY)PHENYL]-1H-PYRAZOLE-4-CARBOXYLIC ACID," 05 06 2020. [Online]. Available: https://www.rcsb.org/ligand/YSH. [Accessed 26 08 2025]. |
| [86] | A. Meller, M. Ward, J. Borowsky, M. Kshirsagar, J. M. Lotthammer, F. Oviedo, J. Lavista Ferres and G. R. Bowman, "Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network," *Nature Communications,* vol. 14, no. 1177, 2023. |
| [87] | A. Casamayor and J. Ariño, "Fungal Hal3 (and Its Close Relative Cab3) as Moonlighting Proteins," *Journal of Fungi,* vol. 8, no. 10, p. 1066, 2022. |
| [88] | D. Bajusz, A. Rácz and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *Journal of Cheminformatics,* vol. 7, no. 20, 2015. |
| [89] | P. Willett, "Similarity-based virtual screening using 2D fingerprints," *Drug Discovery Today,* vol. 11, no. 23-24, pp. 1046-1053, 2006. |
| [90] | H. Sun, G. Tawa and A. Wallqvist, "Classification of Scaffold Hopping Approaches," *Drug Discovery Today,* vol. 17, no. 7-8, pp. 310-324, 2012. |
| [91] | T. C. Terwilliger, D. Liebschner, T. I. Croll, C. J. Williams, A. J. McCoy, B. K. Poon, P. V. Afonine, R. D. Oeffner, J. S. Richardson, R. J. Read and P. D. Adams, "AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination," *Nature Methods,* vol. 21, pp. 110-116, 2024. |
| [92] | V. Scardino, J. I. Di Filippo and C. N. Cavasotto, "How good are AlphaFold models for docking-based virtual screening?," *iScience,* vol. 26, no. 1, p. 105920, 2023. |
| [93] | D. Ramírez and J. Caballero, "Is It Reliable to Use Common Molecular Docking Methods for Comparing the Binding Affinities of Enantiomer Pairs for Their Protein Target?," *Internatoinal Journal of Molecular Sciences,* vol. 17, no. 4, p. 525, 2016. |
| [94] | J. M. Paggi, A. Pandit and R. O. Dror, "The Art and Science of Molecular Docking," *Annual Review of Biochemistry,* vol. 93, pp. 389-410, 2024. |
| [95] | G. Tresadern, K. Tatikola, J. Cabrera, L. Wang, R. Abel, H. van Vlijmen and H. Geys, "The Impact of Experimental and Calculated Error on the Performance of Affinity Predictions," *Journal of Chemical Information and Modeling,* vol. 62, no. 3, pp. 703-717, 2022. |
| [96] | K. Szilágyi, B. Flachner, I. Hajdú, M. Szaszkó, K. Dobi, Z. Lőrincz, S. Cseh and G. Dormán, "Rapid Identification of Potential Drug Candidates from Multi-Million Compounds’ Repositories. Combination of 2D Similarity Search with 3D Ligand/Structure Based Methods and In Vitro Screening," *Molecules,* vol. 26, no. 18, p. 5593, 2021. |
| [97] | M. Dimura, T.-O. Peulen, H. Sanabria, D. Rodnin, K. Hemmen, C. A. Hanke, C. A. M. Seidel and H. Gohlke, "Automated and optimally FRET-assisted structural," *Nature Communications,* vol. 11, no. 5394, 2020. |
| [98] | J. Tang, S.-G. Kang, J. G. Saven and F. Gai, "Characterization of Cofactor-Induced Folding Mechanism of a Zinc Binding Peptide Using Computationally Designed Mutants," *Journal of Molecular Biology,* vol. 389, no. 1, pp. 90-102, 2009. |
| [99] | S. Sen and J. B. Udgaonkar, "Binding-induced folding under unfolding conditions: Switching between induced fit and conformational selection mechanisms," *Journal of Biological Chemistry,* vol. 294, no. 45, pp. 16942-16952, 2019. |
| [100] | M. Varadi, D. Bertoni, P. Magana, U. Paramval, I. Pidruchna, M. Radhakrishnan, M. Tsenkov, S. Nair, M. Mirdita, J. Yeo, O. Kovalevskiy, K. Tunyasuvunakool, A. Laydon, A. Žídek, H. Tomlinson, D. Hariharan, J. Abrahamson, T. Green, J. Jumper, E. Birney, M. Steinegger, D. Hassabis and S. Velankar, "AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences," *Nucleic Acids Research,* vol. 52, no. D1, p. D368–D375, 2024. |

Erklärung

Hiermit erkläre ich, dass die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt wurde.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Görlitz,

Unterschrift, Masine Tränkner