Charles University

Faculty of Science

Study programme: Bionformatics

Branch of study: N-BINF

Bc. Miloš Halda

Physics-based protonation of protein-ligand complexes

Fyzikálně opodstatněná protonace komplexů proteinů s ligandy

MASTER THESIS

Supervisor: RNDr. Martin Lepšík, Ph.D.

Consultant: doc. RNDr. Jan Řezáč, Ph. D.

Prague 2025

I declare that I carried out this master thesis on my own, and only with the cited sources, literature and other professional sources. I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

First of all, I would like to thank my supervisor, Martin Lepšík, for his guidance throughout the entire process and for always making time – whether to listen to my successes or to help with the problems at the edge of my abilities. And for all the insightful suggestions!

I would also like to thank my consultant, Jan Řezáč, who helped me a lot with the framing of this project and was able to quickly distinguish the essential from the unimportant.

I would also like to thank Jindřich Fanfrlík for consultations and for providing me with calculated energies for water species, and Adam Pecina for listening to my super-long presentation and giving me valuable comments on my work. My thanks go to all the authors of SQM2.20, which is an essential part of my project.

And my thanks also go to the other team members at IOCB, especially Vilhelmiina, who gave me important academic "black market" tips, Martin for the many silly and interesting fun facts he is always ready to share, and Tomáš for the evenings in the office before our deadlines.

Thanks also to Johana. You were my best cheerleader and critic at the same time. You helped me keep a reasonable pace even in the most stressful moments. I promise to take you out for an ice cream!

Title: Physics-based protonation of protein-ligand complexes

Author: Bc. Miloš Halda

Supervisor: RNDr. Martin Lepšík, Ph.D.

Consultant: doc. RNDr. Jan Řezáč, Ph. D.

Abstract: Structure-based computer-aided drug design relies on 3D structures of protein–ligand (P–L) complexes for predictions of the binding affinities in the process of scoring. One of the crucial steps is the modeling of hydrogen atom positions which are absent from crystallographic structures. Few tools for automatic protonation of P–L complexes are available but their reliability is questionable. The aim of this thesis was to test whether SQM2.20, the recently developed semiempirical quantum mechanical (SQM) scoring function can recognize the correct protonation variant(s) in P-L complexes. Experimental neutron diffraction structures of P–L complexes have thus been collected from the PDB as a reference for the location of hydrogen atoms. Using the automatic pipeline that was developed, multiple protonation variants of four selected P–L complexes were evaluated by SQM2.20. The protocol unequivocally identified the experimentally determined protonation variant, which was not the case for the standard tools. In conclusion, the presented results pave the way toward a chemically general and fully automated process of computationally defining the correct protonation state of P–L complex based on quantum-mechanical methods for further scoring.

Keywords: protonation, protein–ligand binding, scoring function, semiempirical quantum mechanics, neutron diffraction

Název práce: Fyzikálně opodstatněná protonace komplexů proteinů s ligandy

Autor: Bc. Miloš Halda

Vedoucí práce: RNDr. Martin Lepšík, Ph.D.

Konzultant: doc. RNDr. Jan Řezáč, Ph. D.

Abstrakt: Výpočetní návrh léčiv založený na trojrozměrné stuktuře cíle terapeutického zásahu odhaduje afinitu mezi proteinem a ligandem pomocí tzv. skórování. Klíčovým krokem v tomto procesu je modelování poloh atomů vodíku, které nebývají zjistitelné pomocí rentgenostrukturní analýzy. Existující nástroje pro jejich automatické přidávání mají jistá omezení. Cílem této práce je otestovat schopnost nedávno vyvinuté semiempirická kvantově mechanické skórovací funkce SQM2.20 rozpoznat správné varianty protonace v komplexech proteinů s ligandy. V rámci tohoto projektu byla tato schopnost otestována na několika strukturách komplexů zjištěných pomocí neutronové difrakce, které sloužily jako reference pro umístění vodíků. Pomocí automatické metody vyvinuté v rámci této diplomové práce bylo s využitím SQM2.20 vygenerováno a ohodnoceno několik variant protonace čtyř referenčních struktur. Tento postup bezchybně určil experimentální variantu protonace pro všechny referenční struktury, narozdíl od dostupných metod. Závěrem lze říci, že prezentované výsledky přispěly k vytvoření chemicky obecného a plně automatického procesu výpočetního určení správné protonace komplexů proteinů s ligandy založeném na kvantově-chemických metodách.

Klíčová slova: protonace, interakce protein–ligand, skórovací funkce, semiempirická kvantová mechanika, neutronová difrakce

Contents

1 Introduction

1.1 Drug Design and Development

Ever since, the history of humanity has been troubled by various diseases caused either by pathogens or malfunctions of common processes in the human body. They can be treated by medicinal substances called drugs. Historically, the drugs for disease treatment were found mostly by chance. Some of them were natural products, such as in the case of quinine for the treatment of malaria from *Cinchona* tree bark or salicylic acid as anti-inflammatory drug from willow bark. Others were prepared by chemical synthesis, such as Salvarsan discovered by Paul Ehrlich, which, however, showed high toxicity. [1, 2]

Modern science introduced a deeper understanding of the biological and chemical systems, which allowed the introduction of rational rules for drug design and development. One of the central assumptions is called the lock-and-key model and dates to the end of the 19^{th} century. It states that the drug or substrate (key) binds to a protein (lock) if they have complementary geometries. This assumption proved to be generally correct. The current model, however, considers the dynamics of both binding partners in the process called "conformational selection" and the binding is mainly caused by non-covalent interactions, hydrogen bonds, dispersion, etc. If the function of the protein is involved in the mechanism of the disease, the small ligand may be able to deactivate it and halt the disease's development.

Since the 1950s, drug development has been significantly enhanced by the advent of molecular biology, enabling the direct search for drugs that target specific proteins through a process known as rational drug design. In this approach, the first ligand to demonstrate suitable binding attributes to a target protein is termed a *hit*. This hit molecule is then subjected to further investigation and structural optimization to obtain a *lead* molecule, which has improved affinity for the target, a drug property of utmost importance.

The 1970s saw the advancement of structural methods, particularly X-ray diffraction, which provided 3D models of biological macromolecules. This break-

through was essential for the development of structure-based drug design, where the lock-and-key model is directly applied. In addition, the rapid uprise of computational technologies has further advanced drug design through computer–aided methods. Since then, small molecule drug candidates have been computationally modeled (e.g., *docked*) into the target protein's binding site to estimate their binding affinity.

Currently, drug design is a thriving field of science with high expectations from society. Many steps in drug development are demanding in terms of time and cost. Computer-aided drug design (CADD) aims to reduce the time and expenses required to identify suitable drug candidates, hits, and to optimize these molecules so they effectively fulfill their intended purpose, a process known as lead optimization. This is achieved through the use of computational approaches rather than experiments. [3]

There are two basic perspectives of drug design – an already mentioned *structure-based* and *ligand-based*. Structure–based drug design (SBDD) uses knowledge of the protein involved in the process of disease development, also called the *target*, and ligand-based drug design (LBDD) proceeds from the perspective of the relatively small therapeutic molecules in the search for the *lead* molecule. Of course, there is some overlap between the two, but the terms can be useful for focusing on the specifics of protein and ligand based research. SBDD requires the determination, preparation (correction) and 3D representation of the protein (using molecular formats like PDB, mmCIF and MOL). The LBDD is interested in small molecules and their properties. In many cases, these small molecules do not even need to be stored in 3D, allowing the use of one-dimensional formats such as SMILES (Simplified Molecular Input Line Entry System) or InChI for a very simple and memory efficient representation of such molecules. [1, 2]

1.2 Structure-Based Drug Design

The structure–based approach of CADD (SBDD) is dependent on the 3D model of the target, mostly in its ligand–bound state, also called protein–ligand (P–L) complex. [1] These structures can be obtained experimentally (X-ray diffraction

– see Chapter 1.2.1, nuclear magnetic resonance, cryogenic electron microscopy or neutron diffraction (ND) – see Chapter 1.2.2) or predicted using computational methods (homology modeling or artificial intelligence models such as Alpha Fold [4]).

A process called *docking* is required to computationally generate a model of a P–L complex from a potential drug candidate (ligand) and the structure of a target protein. This process involves finding the optimal position and orientation of a ligand within the binding site of a protein. Since most drug candidate molecules can adopt different conformations, as well as tautomer and protonation variants, tools are needed to generate the most probable and chemically valid options. To decide between alternative poses and ligand variants and to estimate binding affinity of the ligand towards the target protein, a process of *scoring* is performed. The approaches for scoring range from simple scoring functions (SFs) to advanced physics–based methods as described in Chapter 1.2.5.

This thesis does not discuss cryogenic electron microscopy (too low resolution for efficient SBDD) and nuclear magnetic resonance (lack of ligands[1]) methods for protein structure determination. It also does not cover computational methods for protein structure resolution, because they may contain artifacts or have otherwise inappropriate properties. For instance, AlphaFold predicted structures do not contain hydrogens. However, the most common method for protein structure determination, X-ray diffraction, and ND have properties that need to be discussed because they are critical to the aims of this project.

1.2.1 X-ray Diffraction

The well-established method of X-ray diffraction employs X-rays to diffract off the atoms of a crystallized protein, thereby reconstructing its 3D structure. However, this method is incapable of determining the positions of hydrogen atoms due to their single electron and higher disorder compared to heavier atoms, which results in minimal contribution to the total scattering. Even the most ordered hydrogen atoms may only appear at atomic resolutions of 1.2 Å or better,

[1] In a very recent study, the inclusion of ligands in the structure solved by nuclear magnetic resonance is discussed and applied. [5]

a condition that is relatively uncommon. [6]

Hydrogen atoms may be added to a structure in places where they are assumed to be present. For example, for carbon atoms, the hydrogen positions follow the well-known geometry of valences. However, for hydroxyl groups, such an assumption is not possible. The hydrogen atom may be absent, depending on the pK_a value of the specific hydroxyl group (see Chapter 1.4.1), or its position may rotate. [7] In such cases, the rotation of the hydrogen must be resolved, for example, using hydrogen annealing.

1.2.2 Neutron Diffraction

The neutron diffraction method can be used to observe hydrogen positions. The method uses atom nuclei for scattering instead of electron density. Hydrogens mostly need to be exchanged for deuterium ^{2}H atoms) and appear at resolutions below 2.5 Å.[6]

The hydrogen isotope ^{1}H (also called protium) consists of one proton and has small and negative neutron scattering length [8, 9] (-0.374×10^{12}cm) compared to other atoms (0.665×10^{12}cm for C, 0.936×10^{12}cm for N and 0.581×10^{12}cm for O). That makes the ^{1}H hardly observable. An important part of protein crystal preparation for the neutron diffraction is therefore an exchange of hydrogen atoms for deuterium, (^{2}H or D) with scattering length of 0.667×10^{12}cm. This method makes the hydrogen positions in protein structure visible in most cases. [6]

In a manner analogous to the nomenclature employed for X-ray structures, which are commonly referred to as *X-ray structures*, the structures solved by neutron diffraction are referred to as *neutron structures*. This convention is also employed in this thesis.

1.2.3 Joint X-ray and Neutron Diffraction

It is a common practice to solve protein structure both by neutron and X-ray diffraction. [10, 11] The joint X-ray/neutron diffraction structures allow high precision positioning of heavy atoms (such as carbon, nitrogen or oxygen) using X-ray diffraction addition of hydrogen positions from the neutron diffraction experiment. As you can see in Figure 1.1, the density maps prepared by different

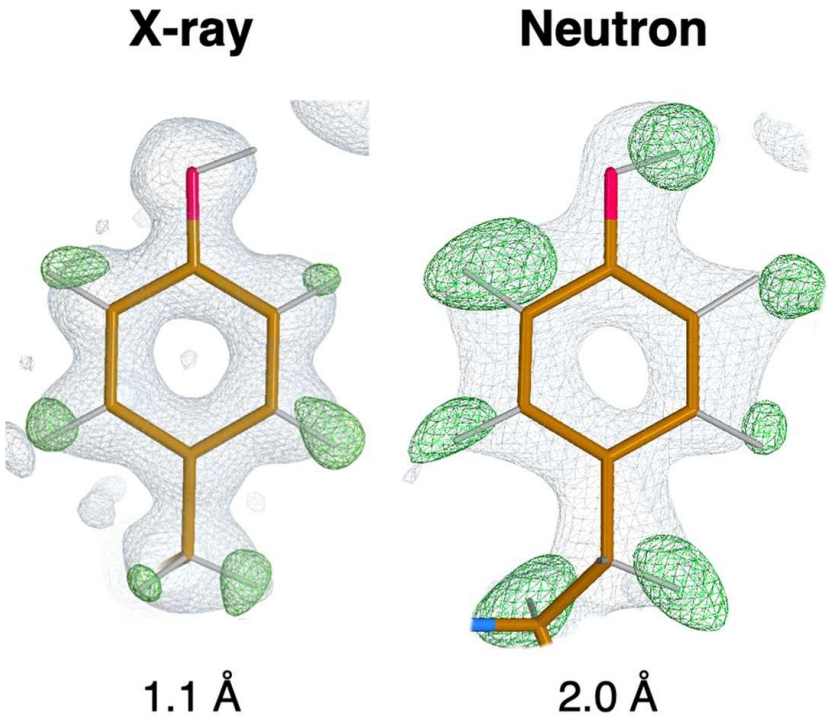


Figure 1.1 Left image shows electron density maps for Tyr12 in PDB (Protein Database) entry 3KYU, at 1.1 Å resolution. The hydroxyl atom is not visible. Right image shows neutron scattering length density maps for Tyr146 in PDB entry 1CQ2 solved at 2.0 Å resolution. All hydrogens are visible including hydroxyl. The figures were obtained from [6].

methods contain different valuable information. For the purposes of this project, structures solved using joint X-ray/neutron diffraction are referred to simply as neutron structures, as they offer comparable accuracy in determining hydrogen atom positions. Furthermore, any mention of neutron diffraction (ND) experiments should be understood to include joint X-ray/neutron diffraction unless stated otherwise.

1.2.4 Issues in Neutron Diffraction Structures

The neutron structures may have several problems, which require careful treatment. The first is, that a simple form hydrogen - deuterium exchange of a P–L complex works for only about a quarter of exchangeable hydrogens in a system. The rest are not going to be exchanged for deuterium, if the P–L complex prepared by standard means is simply introduced into D_2O solvent. In

order to have all H atoms exchanged for deuteriums. For better structures, the proteins and ligands need to be synthesized from the deuterium. [12]

Second problem is, that even if the hydrogens are exchanged for deuteriums, the atoms may still have higher disorder and therefore may not be reflected in the resulting structure. [13]

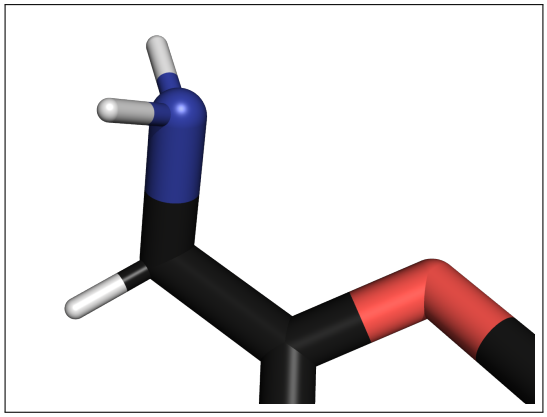
The source of neutrons for neutron diffraction is quite weak compared to the source of X-ray which leads to generally lower resolution of the structure. [10] We can expect the neutron protein structures to miss the parts of the structure, which are not in a highly stable position. Similarly to X-ray method, loops of the protein are often disordered and therefore their scattering reflection is not useful for the structure reconstruction.

In this thesis project, one particularly problematic part was the unreliable protonation of amines in the neutron structures. As shown in Figure 1.2a, the primary amine of the SIS ligand is not protonated correctly. Amines generally become protonated and form ammonium ions at relatively high (basic) pH. In this specific example, the MolGpKa tool predicts a micro-pK_a value for the primary amine of 9.8.[2] Therefore, the amine group is expected to be protonated as displayed in Figure 1.2b. The third hydronium ion may partially exchange with the solvent and is therefore unobservable in the neutron density maps. Similarly, another problem was identified in structure 7TUR as displayed in Figure 1.2c with the ligand PLA, where a hydrogen atom is placed in an unexpected position, close to both the secondary amine and the hydroxyl group. In conclusion, amines present a problematic aspect of neutron structures and therefore require special treatment.

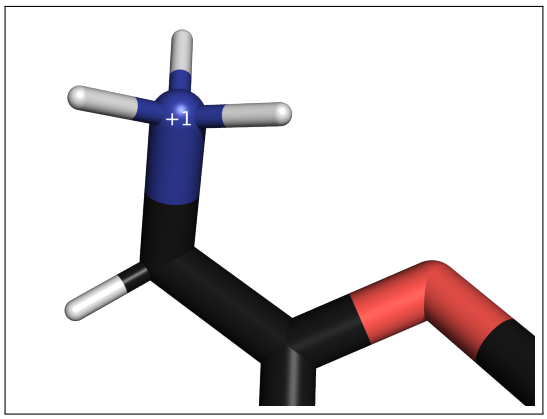
1.2.5 Binding Affinity Predictions

Methods for P–L binding affinity predictions are crucial for efficient and reliable CADD. There are different approaches based on the purpose they are

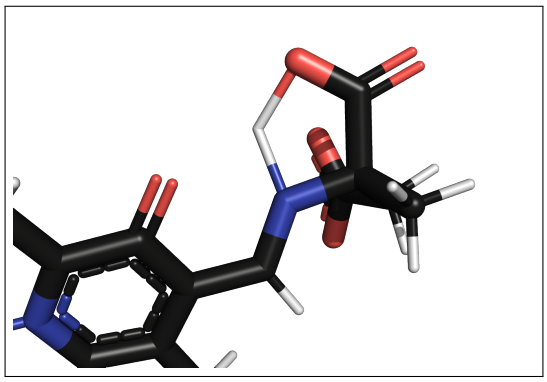
[2]The MolGpKa tool [14] was used in this case for demonstration purposes. It was accessed via its website (`https://xundrug.cn/molgpka`), and the job was submitted using the SMILES string of the SIS ligand obtained from its RCSB (Research Collaboratory for Structure Bioinformatics) PDB entry (`https://www.rcsb.org/ligand/SIS`). Both tasks were performed on 25th April 2025.



(a) Part of the ligand SIS from the neutron structure 6BBZ. The displayed primary amine group has two hydrogens and neutral charge.



(b) Part of the ligand SIS from the neutron structure 6BBZ. The displayed primary amine was corrected in PyMOL for demonstration purposes to match the expected protonation state.



(c) Part of the ligand PLA from the neutron structure 7TUR. The hydrogen is incorrectly assigned to both the secondary amine and to the carboxyl group.

Figure 1.2 Carbon atoms are displayed as black, hydrogen atoms are white, nitrogen atom is naval blue and the oxygen atom is salmon red. The image was prepared using PyMOL.

used for. Approximate SFs have traditionally used knowledge from databases in a statistical manner to set up an empirical master equation to predict affinity based on structural features. Due to their speed, they can be used for molecular docking, to score multiple P–L complex conformations in seconds. [15]

The current state-of-the-art methods to calculate P–L binding free energies are based on the classical approximation in molecular mechanics – the atoms are treated as spheres, bonds are approximated as springs and the effect of electrons are subsumed into the parameters, overall called force field. The solution of Newton's equations of motion gives rise to the general method of molecular dynamics. The methods for binding affinity predictions fall into three categories: endpoint, alchemical or pathway methods. [16–18] They are mostly slow (hours to days of computational time), require powerful hardware (GPUs) and are liable to force field limitations, such as the inability to describe polarization or charge transfer effects.

Such limitations are absent from quantum-mechanical (QM) treatment where the electrons are considered explicitly. Their even larger computational cost can be reduced via numerous fragmentation schemes. [19] Another non-exclusive option is to use semiempirical QM (SQM) methods, described in detail in Chapter 1.5.

1.2.6 Protein–Ligand Datasets

The development of reliable, robust and efficient methods for CADD requires a resolved structures of P–L complexes. The more precise the method aims to be, the more accurate the structures are needed. Average structures in the PDB [20] have flaws, which may lead to erroneous assumptions in the development of the new methods and also does not allow reliable testing of the methods. That creates a demand for curated datasets of structures with well-defined properties, such as provided by the CASF-2016 update benchmarking study[21] or the recently published PL-REX (Protein–Ligand Refined Experiment) [22] dataset.

PL-REX dataset consists of ten diverse protein targets with ten to thirty ligands each. Every P–L structure in the dataset was manually checked for non-trivial issues regards the molecule geometry and protonation. As we can see in Figure 1.3, the PL-REX preparation workflow follows parts of the process of

structure-based drug design to produce high-quality P–L structures.

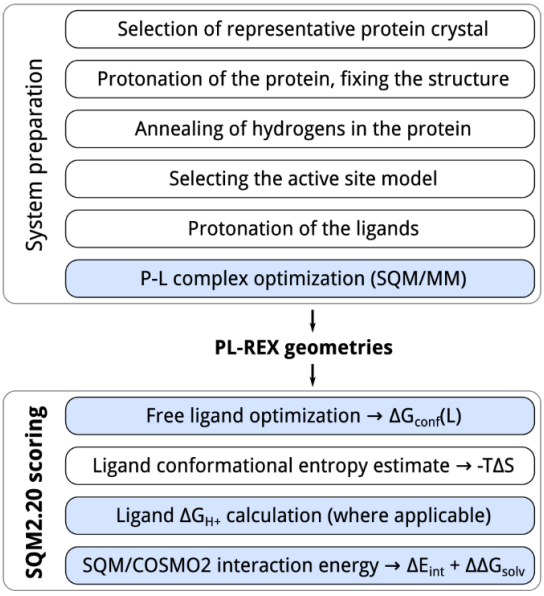


Figure 1.3 Workflow of both the PL-REX dataset preparation process and SQM2.20 scoring. The blue-filled steps make use of SQM methods. The figure was obtained from [22].

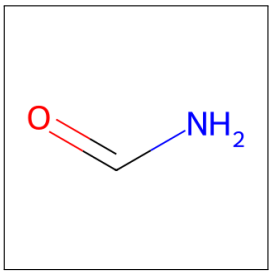
1.3 Ligand–Based Drug Design

Ligand-based drug design (LBDD) focuses on the analysis of ligands and their properties to identify potential drug candidates. The primary objective is to discover small molecules that exhibit the desired biological activity within a specific chemical context. To evaluate the suitability of a ligand as a drug candidate, an ADMET analysis is conducted. ADMET stands for Absorption, Distribution, Metabolism, Excretion, and Toxicity – key pharmacokinetic and pharmacodynamic properties that influence a compound's efficacy and safety profile. [23] An important concept of LBDD is pharmacophore defined as the spatial arrangement of chemical features essential for molecular recognition and biological activity. [24] Numerous computational tools have been developed for such predictions, including ChemAxon's Marvin suite [25], LigandScout [26], or Phase module of the Schrodinger suite [27]. These platforms can be used for ligand analysis and interactions within P–L complexes, they are, however, commercially licensed. For most of the individual cheminformatic tasks there are, however, tools

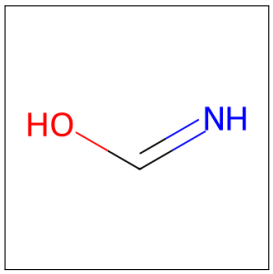
which are developed under academic or even open access policies.

1.3.1 Tautomerization

A common phenomenon that adds further complexity to the chemistry of certain ligand functional groups is *tautomerization* — a switch of a double bond coupled with a hydrogen transfer within a molecule. Tautomerization can be predicted, and tables of known probabilities indicate the likelihood of a molecule adopting a particular tautomeric form. One example, which also needed to be addressed in this project, is the amide–imide tautomerism. The simplest amide (formamide) has the formula $O = CH - NH_2$, while its imidic acid form is $OH - CH = NH$. Their 2D structures are illustrated in Figure 1.4. Epikx [28] and LigPrep [29] are able to generate ligand tautomers, which enhances the understanding of the tautomeric states that may occur within the studied P–L complex.



(a) 2D drawing of formamide.



(b) 2D drawing of imidic acid.

Figure 1.4 2D drawings prepared using RDKit.

1.3.2 Lipinski Rules

Lipinski Rule of Five is a set of rules derived from observations of attributes of orally applicable drugs. The rules are used to estimate the compound's druglikeness. The rules state that a promising drug candidate should have less than 5 H-bond donors, less than 10 H-bond acceptors, molecular weight less than 500 Da and Log P smaller than 5. [30]

These rules had a rich follow up discussion and many other rules of thumb for drug-like molecules were invented with Lipinski Rule of Three among them.

Lipinski Rule of Three was prepared for screening of small molecule fragment libraries in search for hits which could then be joined. [31] Such fragments are defined as by molecular weight, which needs to be less than 300 Da. The three rules are then LogP less than 3, number of H-bond donors and acceptors less than 3 and less than 3 rotatable bonds. [32]

1.3.3 Cheminformatic Tasks

In LBDD, important tasks involving small molecules include, for example,

- the search for maximal common substructures,
- simple reconstruction of three-dimensional conformations,
- storage of key molecular attributes,
- conversion of molecule file formats without loosing valuable information,
- unambiguous atom ordering,
- drawing of molecules for a better inside and more.

Such tasks can be efficiently performed using RDKit, a free and open-source toolkit that provides a comprehensive Python application programming interface (API), allowing easy integration into custom computational pipelines. [33]

1.4 Common Concepts of SBDD and LBDD

1.4.1 pKa Estimation

Acid dissociation constant, $\mathrm{p}K_\mathrm{a}$, is a dimensionless number, which describes the equilibrium of concentration of acid and its dissociation. [34] For reaction equation of acid AH, dissociated acid A^- and of hydronium ion H^+ written as:

$$\mathrm{AH} \rightleftharpoons \mathrm{A}^- + \mathrm{H}^+$$

and the concentration of the acid and products of its dissociation denoted by square brackets, the $\mathrm{p}K_\mathrm{a}$ value is defined as:

$$\mathrm{p}K_a = \log_{10} \frac{[\mathrm{HA}]}{[\mathrm{A}^-][\mathrm{H}^+]} \tag{1.1}$$

For pK_a estimation of ligands, there are multiple free tools, such as MolGpKa [14] which employs convolutional neural networks or QupKake [35], which is based on both machine learning and quantum chemistry. PROPKA 3 [36] is an open source tool, which is capable of pK_a estimation for whole proteins or P–L complexes. The commercial Schrodinger suite includes tools for both small molecules and P–L complexes with various approaches - empirical (Epik Classic), machine learning (Epik 7) or quantum chemical (Jaguar pK_a). [37]

1.4.2 Protonation of P–L Complexes

Correct assignment of hydrogen atoms to both small molecules and biomacromolecules is a crucial prerequisite for successful physics-based scoring. [38, 39] It is an uneasy task, which can be performed using either an experimental or a computational (empirical, machine-learning, or physics-based) approach. The protonation state of a given group is described by the Henderson-Hasselbalch equation. [40] Protonation state is tightly connected to the pK_a value of a given chemical group by Henderson-Hasselbalch equation, which relates

- pH of the solvent,
- pK_a of the chemical group and
- fraction of protonated and deprotonated groups $\frac{[\mathrm{Base}]}{[\mathrm{Acid}]}$

as follows:

$$\mathrm{pH} = \mathrm{p}K_a + \log_{10} \frac{[\mathrm{Base}]}{[\mathrm{Acid}]} \tag{1.2}$$

The equation assumes, that there is an equilibrium between protonated and deprotonated variants, based on the pK_a value of such a group at a given pH. Protons exchange between the solvent and the group. Protonation depends not only on the chemical identity of the group but also on its immediate surroundings, which is quite typical for P–L complexes. [41, 42] For instance, only one out of two catalytic aspartates of HIV-1 protease is protonated based on the interaction with a ligand (see Figure 1.5). [43]

There are tools available to predict protonation of both, small molecules and P–L complexes. The simplest approach is implemented for instance in Open

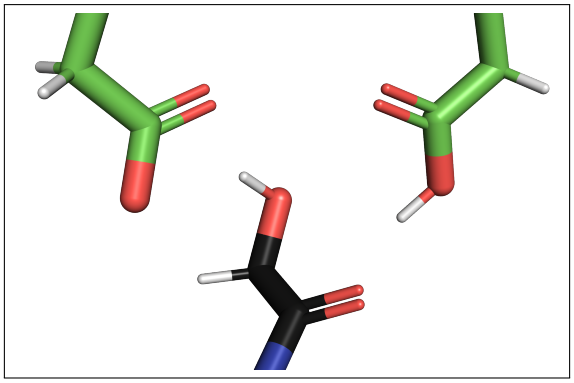


Figure 1.5 Detail of binding site of HIV-1 (Asp25 and Asp125) protease with bound inhibitor KNI. The Carbon atoms of ligand are displayed as black, carbon atoms of the protein are green, hydrogen atoms are white, nitrogen atoms are blue and the oxygen atoms are red. The image was prepared using PyMOL.

Babel [44], which applies a table of transformations [40] of groups in the form of SMARTS (SMILES ARbitrary Target Specification) [45] strings with pK_a values and includes the more probable form (for a given pH) in the output structure. Similarly, the tool Dimorphite [46] uses experimentally obtained pK_a values to predict protonation states of molecules at a given pH.

For P–L complexes, heuristic tool ProToss is available for free. [7] Tools from Schrodinger suite (namely LigPrep [29] and Epik Classic [47]/Epik 7 [28]) are also capable of protonation of both small and large molecules, however, the license is commercial. [27]

Another plausible approach is to use pK_a prediction tools to estimate the microstates of a given structure and determine its most probable protonation state accordingly, e.g. using Open Babel with edited configuration.[3]

1.4.3 Molecular File Formats and Conversion

The file formats utilized in both SBDD and LBDD must possess the necessary attributes to fulfill their specific roles. All file formats inherently involve a certain level of approximation. The formats employed in this thesis are detailed in the following paragraphs.

[3]Open Babel allows to do so by specifying the protonatable groups and corresponding pK_a values in its "phmodel.txt" configuration file. [40]

The SMILES (Simplified Molecular Input Line Entry System) format, with the .smi file extension, is used for representing small molecules. It encodes a molecule as a single-line string composed of atomic symbols, with optional stereochemistry and charge annotations. This representation allows for the reconstruction of hydrogen atoms, which are often omitted to simplify the notation. [48]

SMARTS (SMILES Arbitrary Target Specification) is an extension of the SMILES format designed for specifying molecular substructures. It enables the definition of patterns to match specific molecular fragments, functioning similarly to regular expressions in text processing. [45]

The SDF (Structure-Data File) format, with the .sdf extension, lies at the opposite end of the approximation spectrum relative to line-notation formats. An SDF explicitly encodes every atom in a molecule–including its element type, 3D coordinates, bond orders, and formal charges. In addition, a single SDF may contain multiple molecule variants (i.e., different conformers or tautomeric forms) and can store arbitrary per-molecule data fields (e.g., experimental activities, calculated properties) alongside the structural data. [49]

The PDB (Protein Data Bank) format, with the .pdb extension, is the standard for archiving experimentally determined macromolecular structures in repositories such as the RCSB PDB. Each PDB file contains a series of "ATOM" and "HETATM" records that specify, for every atom in the structure, its element, residue name, chain identifier, residue sequence number, 3D coordinates, where available – formal charge or occupancy. Additional record types capture connectivity, secondary structure annotations, and metadata such as experiment type and resolution. [50]

The MAE (Maestro) format, with .mae extension, is a proprietary file format used by most of the tools from the Schrodinger suite. It is "the primary (and only lossless) Schrodinger output format", from which the other formats may be generated using tools provided by Schrodinger. [51]

For all non-proprietary molecule formats (SMARTS, SDF, PDB) mentioned and also many more[4] there is e.g. Open Babel tool, which is an open and

[4]For all molecule file formats handled by Open Babel see `https://openbabel.org/docs/FileFormats/Overview.html`, last accessed on 26th April 2025.

extendable software used for molecule files conversion and simple cheminformatic operations. [44] Also RDKit is a plausible choice for conversion of most of the file formats. [33]

1.5 Quantum Mechanics

Two major branches of quantum mechanical (QM) methods are those based on wavefunction, such as Hartree-Fock and those based on electron density, i.e., density functional theory (DFT). Both classes have methods which can treat hundreds of atoms with a reasonable accuracy. However, for description of protein binding sites with ligands, thousands of atoms need to be treated. This can be achieved by use of SQM methods which approximate Hartree-Fock or DFT calculations [52–54], which are less accurate but take minutes for even large systems of thousand atoms. [55] The simplification is based on substituting parameters (constants) for some integrals needed for the QM calculations. [53]

PM6-D3H4X/COSMO2 (parametric model 6 with corrections for dispersion, hydrogen and halogen bonding with conductor-like screening model 2), SQM method described in 1.5.1, slightly changed the narrative about lower accuracy SQM methods and shows accuracies comparable with the DFT methods, while maintaining the speed of computations. [22]

The energies computed using SQM2.20 have the following meanings. The electronic potential energy computed using the PM6 method is denoted as ΔE. The energy associated with the COSMO/COSMO2 solvation models represents a Gibbs free energy, ΔG. Their sum may be referred to as $\Delta G'$, but, for simplicity, ΔG is used throughout this project instead. Following the thermodynamic conventions, the lower the energy, the more probable is the given state of the system.

1.5.1 PM6

Parametric model 6 (PM6) is a popular [52] SQM method introduced by James Stewart in 2007. [56] It is a reparametrization of the original PM3 method and its modified neglect of diatomic overlap (MNDO) based approximations. [54]

The PM6 have known flaws especially for dispersion, hydrogen bonding and halogen bond. Those problems significantly restricted the applicability of the PM6 method and the first corrections did not allow the applicability of PM6 for geometry optimizations and molecular dynamics. This changed by development of PM6 with empirical X correction (PM6-X) by Řezáč and Hobza in 2011 which resulted in method with error of 10%. [57] In 2014, PM6 with D3H4 correction (PM6-D3H4) was developed by the same authors. The hydrogen bond correction was completely redesigned and the dispersion correction was based on previously developed DFT-D correction. [58] PM6-D3H4 has higher accuracy compared to other SQM methods and similar to DFT-based methods.[5]

New modifications for the PM-family of SQM methods are being developed, such as correction PM6-S for oxygen-sulfur interactions. [59] Recently published paper shows promising results using machine learning correction for PM6 in vacuum. [60] Also new parametrizations such as PM7 are being developed to describe some properties of the molecular interactions even better. However, the PM6-D3H4X still shows superior results. [61]

1.5.2 Solvents - COSMO/COSMO2

Solvent environment, usually water, can be added to the system in two ways. It can be either explicit, which means adding water molecules to the system or implicit model. Both approaches have its advantages and disadvantages and are able to produce reasonable results. The explicit solvent is computationally usually less efficient. To obtain reasonable results with the explicit solvent, several hundreds water molecules need to be added into a system and sampled regards their configuration space, which leads to highly increased computational cost. Implicit solvents approximate the solvent effect on the molecule behavior, such as hydrogen bonding or solvent polarization, in continuous manner. ([53], Chapter 16.3)

The conductor-like screening model (COSMO) is an implicit solvent method developed in 1993. [62] COSMO approximates solvent by an ideal conductor, which is reasonable for polar solvents like water. However, in case of less polar solvents (e.g. solvents with low dielectric constant) the performance of COSMO

[5] See Table 1 in [52].

decreases. The description of system given by the dielectric continuum models was considered as a fair advancement. This lead to development of new model - COSMO-RS (RS stands for real solvent) in 1995. [63] The second mentioned model is widely used in calculations until now. [64]

Meanwhile the need for accurate description of solvation systems, especially for systems of large P–L complexes in the context of CADD increased. COSMO solvent model was therefore recently reparametrized and modified by addition of nonpolar term. The new method, solvation model COSMO2, was developed in a way that it is available for PM6 and PM7 SQM methods. COSMO2 have shown improved accuracy on diverse datasets of small compounds, such as SAMPL1 or SAMPL4. It also performed well in large model systems of interaction of diverse ligands in the active site of carbonic anhydrase II. [65]

1.5.3 PM6-D3H4X/COSMO2

The use of PM6-D3H4 together with COSMO2 solvent model outperforms other scoring methods for different protein targets. It is also capable of high quality docking, virtual screening and other tasks of structure-based drug design. [55] The PM6-D3H4X/COSMO2 method was also successfully used in follow up research. [22, 66]

1.6 SQM2.20 Workflow

The accuracy and robustness of SFs is necessary for diverse tasks in CADD. Recently published article by the group of Associate Professor RNDr. Jan Řezáč, Ph. D., from the Institute of Organic Chemistry and Biochemistry in Prague presents a novel SQM based scoring function SQM2.20, which is robust, accurate and delivers results in time scale of minutes. This SF is based on state-of-the-art methods in SQM computation.[22]

Before the free energy computation, the P–L complex is treated as follows. The ligand is optimized using SQM (PM6-D3H4X/COSMO2 [57, 58, 65]). The protein residues in distance of 4 Å from ligand are optimized using molecular mechanics (MM), namely AMBER ff19SB [67] force field. Protein residues in

distance of 4 to 10 Å remain in their original positions without optimization. The rest of the protein is removed and the short peptide chains thus formed are capped if needed. The removal step allows the faster computations because of reduced number of atoms in a system. It was shown [55], that the computations on a smaller P–L model perfectly reproduce the results computed for non-reduced protein structure.

PM6-D3H4X/COSMO2 was selected as the main computational method to obtain the energies of the system, because it was recently shown to have accuracy at the level of the DFT computations while maintaining the speed of the SQM methods. It is also implemented in MOPAC together with MOZYME algorithm, which makes it faster compared to other SQM methods.

As we can see in Figure 1.3, the SQM2.20 workflow uses four types of energies to compute the final score. Each energy term has a strict physical meaning:

$$\mathrm{SQMScore} = \Delta \mathrm{E_{int}} + \Delta\Delta \mathrm{G_{solv}} + \Delta \mathrm{G_{conf}(L)} + \Delta \mathrm{G_{H+}} - \mathrm{T}\Delta \mathrm{S} \qquad (1.3)$$

The terms in equation 1.3 represent the gas phase interaction energy ΔE_{int}, the change of solvation free energy upon complex formation $\Delta\Delta G_{solv}$, the change of conformational free energy of the ligand in an aqueous environment $\Delta G_{conf}(L)$, the free energy of proton transfer between the ligand and the buffer ΔG_{H+} and the loss of ligand conformational entropy upon binding $T\Delta S$. The SQMScore is computed as a sum of terms at the PM6-D3H4X/COSMO2 level. The computations are performed using MOPAC2016 [68] with MOZYME algorithm [69] and the D3H4X correction is added using Cuby4 (see Chapter 3.1.3).

In the SQM2.20 paper, the ΔE_{int} term was computed also by other methods than PM6-D3H4X/COSMO2 for reference. The other methods included DFT-D, MM (using AMBER ff19SB force field [67, 70]) and other methods e.g. with machine-learning approach. The scores were computed on the PL-REX dataset (see Chapter 1.2.6). As you can see in Figure 1.6, the scores produced by different SFs vary a lot, however, the SQM2.20 outperforms all if them significantly in correlation of the scores with the experimental energies. The correlation was used instead of direct comparison of the scores, because not all SFs produce the score in energy units. The SQM2.20 score was computed using PM6-D3H4X/COSMO2.

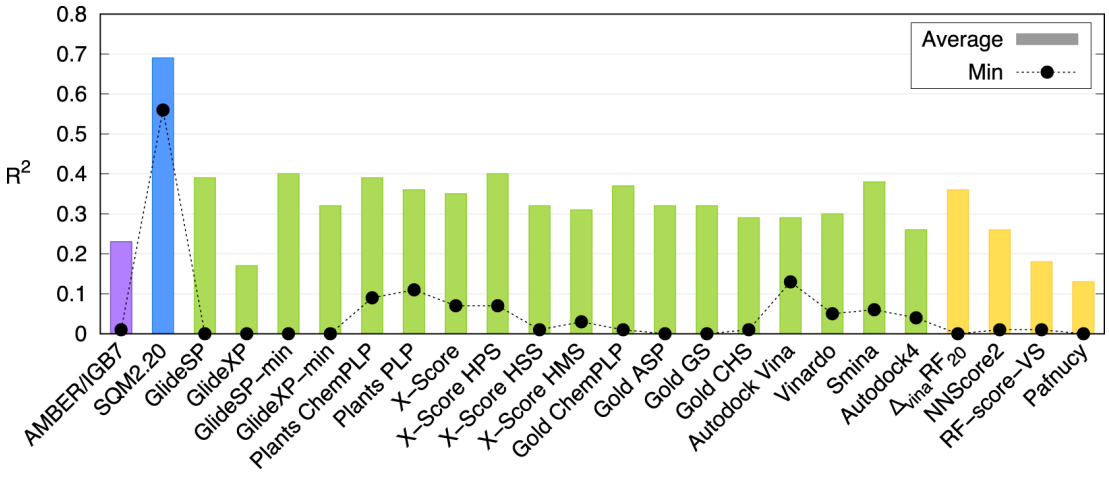


Figure 1.6 Correlation of scores produced by various scoring functions with the experimental energies on on the PL-REX dataset. The figure was obtained from [22].

In this project, only part of the SQM2.20 protocol shown in Figure 1.3 was used, namely the preparation phase augmented with free energy calculations. These entailed the sum of the electronic energy (E) of the P–L complexes plus its solvation free energy (ΔG_{solv}) to obtain the total stabilization energy. To be able to compare the stability of differently protonated P–L complexes, the free energies of the respective water species as defined in section Methods 3.5 were added.

1.7 Description of the Selected P–L Systems

The following four P–L structures are used in this project. The process of selecting these specific structures is described in 3.2. Here the structures and the binding sites of the ligands are described. The text focuses specifically on potential problems with the structures, especially in the P–L binding sites, which are crucial for this project.

1.7.1 PDB Structure 2INQ, DHFR–MT1 Complex

PDB entry 2INQ was produced using netron diffraction method without the use of X-ray diffraction. It has resolution of 2.2 Å and displays an *Escherichia coli* dihydrofolate reductase (DHFR) bound to methotrexate (PDB code MT1), which is an anti-cancer drug. There are two chains of the protein in the PDB structure, each with MT1 bound. [71] The first chain had a flaw of residuum ARG

71 with a broken connectivity to the rest of the peptide, therefore the second chain was taken for the further steps of the project. The second chain contained MT1 molecule named as MT1 1147. The structure did not contain any metal atoms. Please see the binding site of the P–L complex displayed in Figure 1.7. The ligand had two minor problems. It lacked one hydrogen on atom MT1 1147 C7, which is chemically incorrect - the C atom is part of an aromatic ring and should be bound by a single and double bond. It also had occupancy for MT1 1147 D1 of only 73%. That may imply that the D atom is present in only 73% and is missing in 27%. However, as discussed in results, the structure with this D atom missing is not probable and therefore it is considered as an error of the experimental structure. The corrected protonation state of MT1 from the neutron structure is displayed in Figure 1.8.

1.7.2 PDB Structure 2ZYE, PR–KNI Complex

The PDB entry 2ZYE presents a HIV-1 protease (PR) in complex with its inhibitor KNI-272 (which is listed as KNI in PDB). It was produced using neutron diffraction at a resolution of 1.9 Å. The structure consists of two peptide chains with the ligand KNI bound between them. [43] Therefore, both chains were taken into account for further steps of the project. The binding site of P–L complex is in Figure 1.9. The KNI ligand in 2D can be seen in Figure 1.10.

1.7.3 PDB Structure 4QXK, PKG I–PCG Complex

PDB entry 4QXK is a protein structure of cGMP dependent protein kinase (protein kinase G or PKG) I in complex with cGMP (PDB code PCG). The structure is solved by joint X-ray/neutron diffraction approach with final resolution of 2.2 Å. [72] The binding site of the P–L complex can be seen in Figure 1.11, the ligand in 2D can be seen in Figure 1.12

1.7.4 PDB Structure 6BQ8, PKG II–WNU Complex

PDB entry 6BQ8 is a protein structure of cGMP-dependent protein kinase II. The ligand with PDB code WNU is a modified cGMP (8-pCPT-cGMP). [73]

There is one glutamine residuum, GLN 336 in the binding site in a proximity to the ligand, which may be incorrectly solved in. It needs to be checked in the protein preparation step, so it is not changed to some other value, while some protein preparation pipelines may try to change its protonation. See the binding site in Figure 1.13.

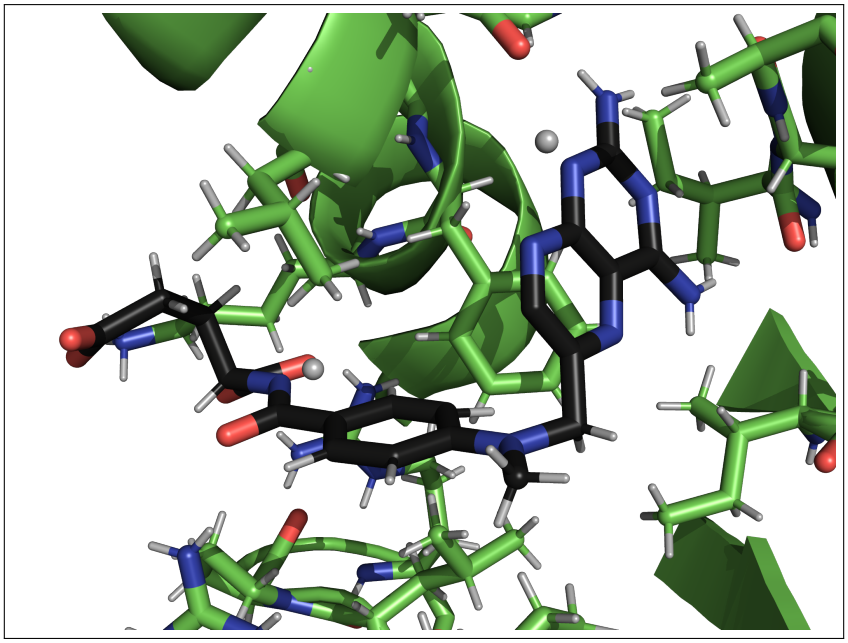


Figure 1.7 Ligand MT1 (1147) in the binding site of P–L complex with PDB code 2INQ. The deuterium atoms shown as balls did not have defined bonds. They, however are part of the ligand. The image was drawn using PyMOL, the ligand is displayed as black, protein parts are colored by green. The protein residues in a close proximity to the ligand (closer than 4 Å) are drawn as sticks, parts of protein in distance of 10 Å are drawn as a cartoon. The image settings apply also for other images of P–L binding site in this chapter.

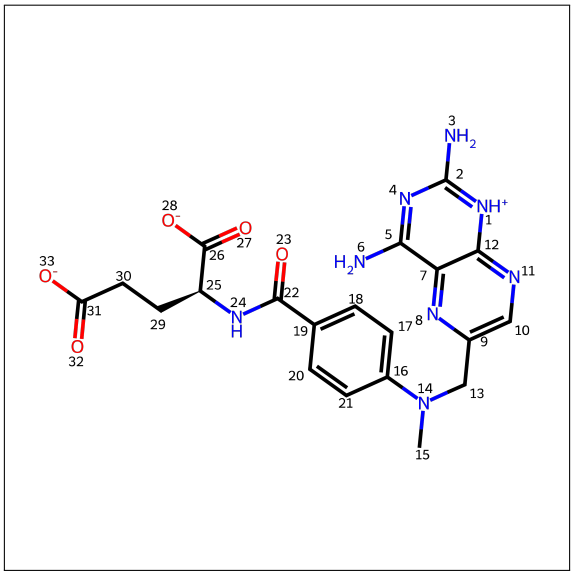


Figure 1.8 2D image of the MT1 ligand in a protonation state same as has MT1 1147 in 2INQ structure. The image was prepared in RDKit, the atom numbering comes from the algorithm for SMILES generation. The latter applies also for other 2D images in this chapter.

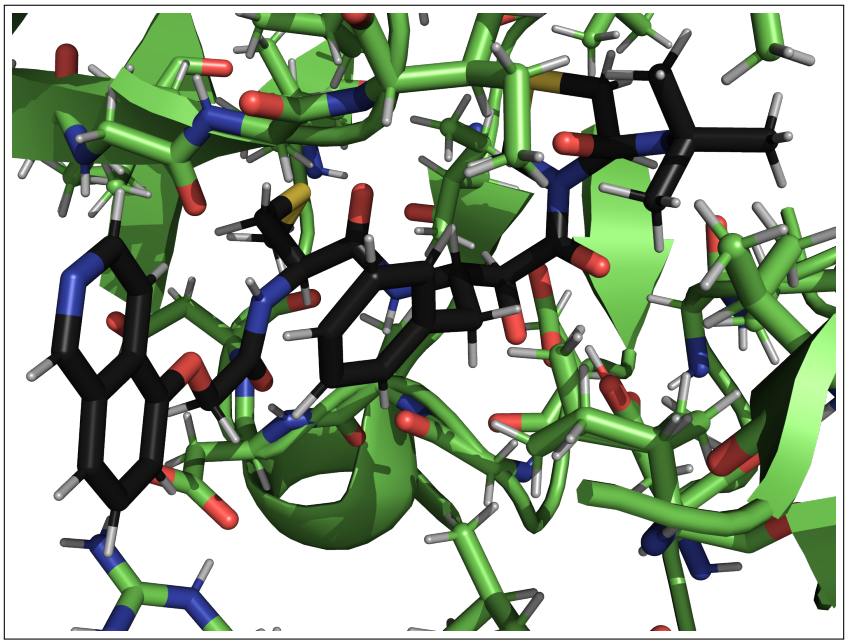


Figure 1.9 Ligand KNI in the binding site of P–L complex 2ZYE.

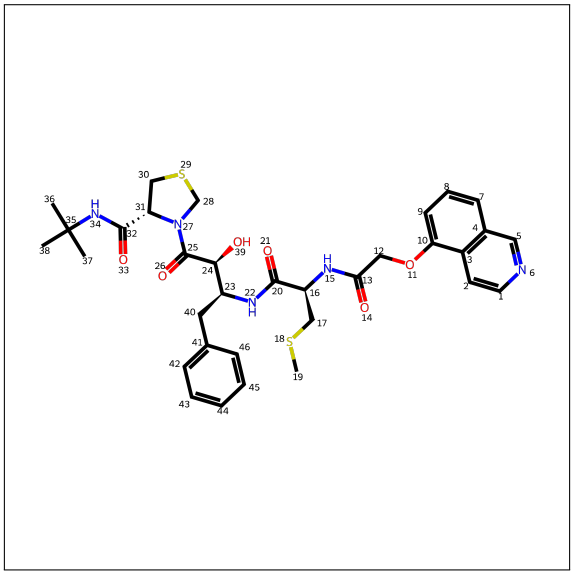


Figure 1.10 2D image of the KNI ligand in a protonation state from the 2ZYE structure.

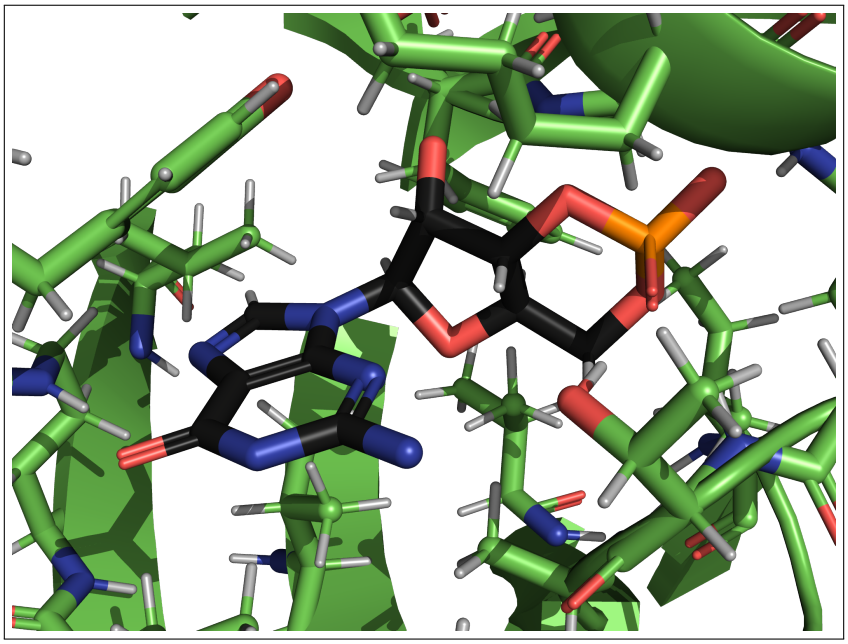


Figure 1.11 Ligand PCG in the binding site of P–L complex 4QXK.

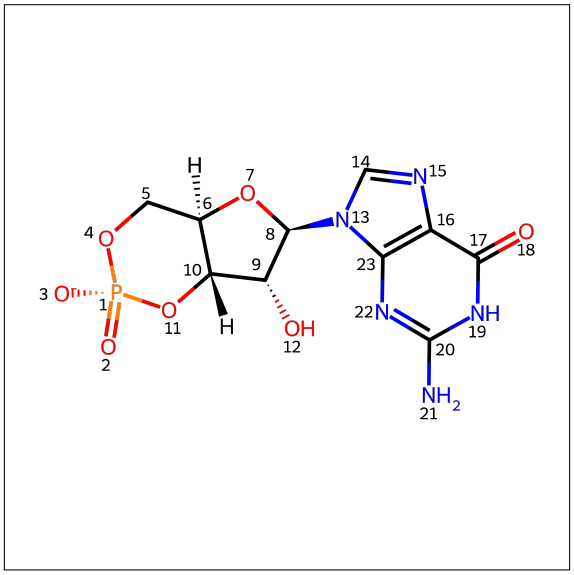


Figure 1.12 2D image of the PCG ligand in a protonation state from the 2ZYE structure.

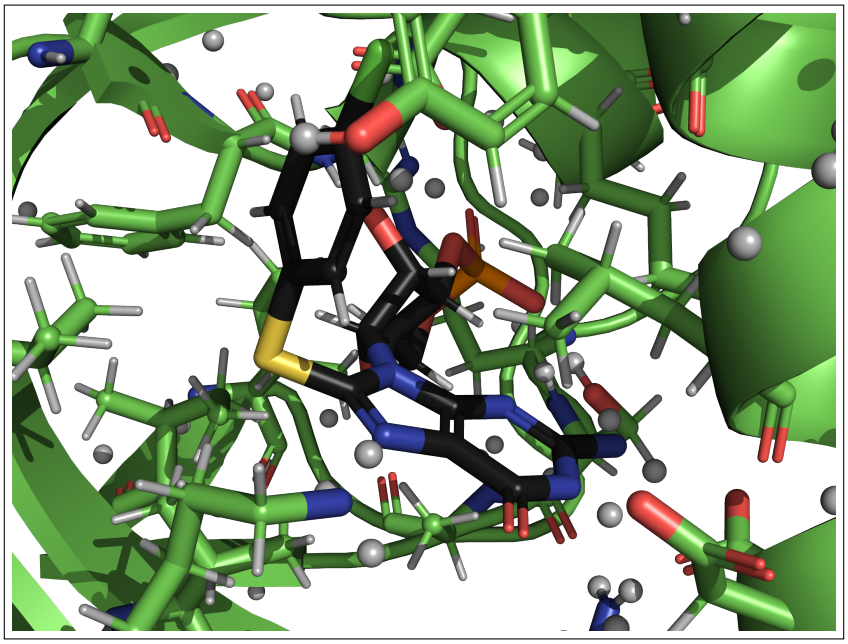


Figure 1.13 Ligand WNU in the binding site of P–L complex 6BQ8. The high amount of hydrogens/deuteria displayed as balls is in the figure due to the dual nature of the structure. The hydrogens solved by X-ray diffraction are shown with a correct connectivity, the deuteria solved by neutron diffraction are displayed as disconnected spheres.

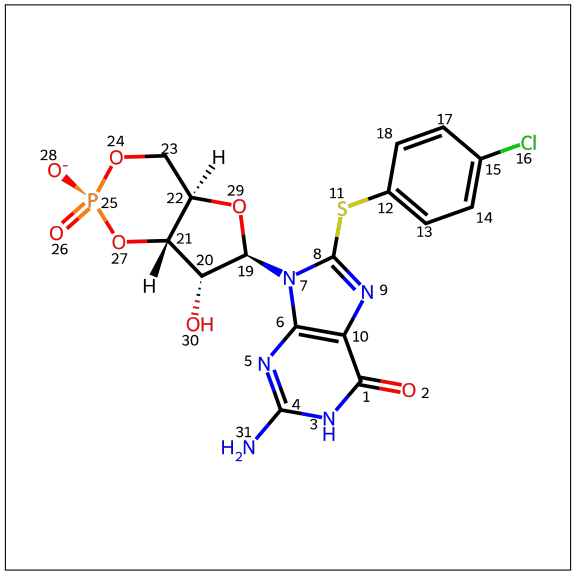


Figure 1.14 2D image of the WNU ligand in a protonation state from the 6BQ8 structure.

2 Aims

The primary objective of this thesis is to assess several available tools for defining protonation states of P–L complexes. Thus, we compare PROTOSS and Maestro of Schrodinger with the recently developed semiempirical quantum mechanical method PM6-D3H4X/COSMO2, as implemented in the SQM2.20 scoring function for their ability to recognize the experimental protonation state of P–L complexes. The individual steps to achieve this general goal consist of

- selection of neutron diffraction P–L structures with reliably solved hydrogen positions,
- preparation of alternative ligand protonation variants,
- preparation of models of the selected protein structures for computations, which includes addition of missing amino acid side chains and loops, removal of waters etc.,
- computation of free energies of prepared models of differently protonated P–L complexes and their comparison with respect to the experimental reference.

This work aims to contribute to the automation of the layered preparation pipeline – an essential goal in CADD for scalability and reproducibility. Specifically, the project aims to develop a workflow that automates the generation of multiple ligand protonation states and prepares P–L complexes suitable for SQM2.20-based scoring.

3 Methods

3.1 Software Tools Used in Project

3.1.1 Open Babel

Open Babel is an open-source chemical toolbox developed for converting between different molecular structure file formats. It can be used via both the command-line interface (CLI) and a graphical user interface (GUI). In this project, Open Babel version 3.1.0 was employed. [74]

3.1.2 RDKit

RDKit is an open-source cheminformatics toolkit developed primarily for applications involving small molecules. It is implemented in C++ and provides API for Python, Java, and C#. In this project, RDKit version 2024.9.5 was used. [33]

3.1.3 Cuby 4

Cuby 4 is an open-source cheminformatics framework written in the Ruby programming language. While it implements several tasks internally, its primary function is to orchestrate external tools and libraries and to integrate their functionalities. It provides access to software such as Amber, MOPAC, and ORCA, among others. Additionally, Cuby 4 offers utilities related to molecular geometry, such as the reconstruction of ligand geometries from SMILES strings and the conversion between different molecular file formats. [75, 76] In this project, Cuby 4 is employed as part of the SQM2.20 workflow, where it provides corrections for dispersion, hydrogen, and halogen bonds (D3H4X) for the PM6 method. [22]

3.1.4 Epikx

Epikx (also known as Epik 7) is a commercial tool developed by Schrodinger for generating protonation variants, tautomers, and for estimating pK_a values of molecules. [77] It can be operated through the Maestro GUI or as a standalone

CLI tool. Epikx represents a redesigned version of the earlier Epik software, now referred to as Epik Classic. [37] Unlike its predecessor, Epikx implements a machine learning approach, specifically employing convolutional neural networks for pK_a prediction, while micro-pK_a site recognition relies on SMARTS pattern-based rules. In this project, Epikx from Schrodinger suite release 2024-3 was utilized.

3.1.5 ProToss

ProToss is a fully automated tool for hydrogen prediction of P-L complexes, developed at the University of Hamburg. It is based on empirical scoring function, [7] and allows usage either from the web GUI[1] or using representational state transfer (REST) API[2]. ProToss was used to generate protonation variants of the ligands in a P-L complex. It is a simple and free-of-charge tool which can be incorporated into automated pipelines using the REST API. It, however, does not allow any user-specific configurations and can be used only for structures in PDB in its API version.

3.1.6 Protein Preparation Workflow

Protein Preparation Workflow (PPW) is a tool from Schrodinger, which serves as a customable pipeline to prepare protein structures for further steps of structure based CADD. It is available both from the Maestro GUI or from CLI. PPW integrates tools such as Prime, Epikx and other from the Schrodinger suite [78, 79].

3.1.7 Maestro

Maestro [80] is a commercial GUI tool developed by Schrodinger for the exploration and modeling of chemical structures across both life sciences and materials science domains. Maestro provides an entry point for accessing other Schrodinger tools such as:

- Epikx for protonation and tautomer variant generation and pK_a estimation,

[1]The tool is available at `https://proteins.plus/`.

[2]More information available at `https://proteins.plus/help/protoss_rest`.

- LigPrep for preparing ligands for diverse computational workflows,
- Prime for protein modeling, including modeling missing loops in crystal structures,
- PPW for preparation of protein model from the structure for further steps of CADD.

Actions performed within the Maestro interface are automatically accompanied by the generation of command-line equivalents and detailed log files. This feature enables task reproducibility and facilitates the integration of GUI-based operations into automated workflows. In this project, Maestro from Schrodinger suite release 2024-3 was utilized.

3.1.8 sdconvert

As mentioned, tools developed by Schrodinger use a proprietary MAE format as their default. Schrodinger suite provides tools to convert such files to their open analogs. One of them, sdconvert, serves to convert MAE files to SDF and back. In this project, sdconvert from Schrodinger suite release 2024-3 was utilized. [81]

3.1.9 PyMOL

PyMOL is a source-available molecular visualization system currently maintained, licenced and distributed by Schrodinger, Inc. It supports both a CLI and a GUI approach. PyMOL incorporates an integrated Python interpreter, which enables scripting and reproducibility of visualization workflows directly within the application. Its functionality can be further extended via plugins such as the Builder panel for molecular construction and editing or Wizard for various measurements within the displayed structure. In this project a user-created build (version 3.0) based on the open-source variant of the PyMOL software was used. [82]

3.1.10 Prime

Prime is a protein structure prediction tool developed by Schrodinger for modeling of protein loops. It is fully integrated into Maestro and its Protein Preparation Workflow. [83] The algorithm of Prime is from the ab initio class of protein structure prediction algorithms. It uses dihedral angle-based buildup procedure with side-chain optimization and energy minimization of selected loop structures. [84]

Prime was used in this project to correct the inputs structures from PDB with missing loops. The exact way how the loops were reconstructed was not of high importance - the loops are expected to be further than 10 Å from the binding site of the ligand. It was chosen because it is conveniently implemented as a part of the Protein Preparation Workflow. However, it would be enough to simply dismiss the missing part and cap the ends of the peptides to end up with a peptide structure with a correct bonding network.

In this project, Prime from Schrodinger suite release 2024-3 was utilized.

3.1.11 MOPAC and MOZYME

The Molecular Orbital PACkage (MOPAC) is a multi-platform software package that integrates various SQM methods. It was developed primarily by James Stewart, who is also recognized for the development of the PM3 and PM6 methods. MOPAC employs SQM approaches to obtain molecular orbitals and heats of formation. These results are subsequently used to calculate additional properties required for specific tasks, such as thermodynamic quantities and molecular force constants. The package also supports the configuration of solvation models, enabling the use of COSMO or COSMO2. [68]

MOPAC incorporates the MOZYME scaling algorithm, which enables efficient computation of large systems. MOZYME leverages localized molecular orbitals, resulting in calculation times that are ”almost proportional” to the system size—an important improvement, as traditional calculations typically scale with the cube of the system size. [69]

Both MOZYME and MOPAC are utilized in this project as part of the SQM2.20 workflow, where they perform SQM calculations.

3.1.12 Amber

Amber is an open-source toolkit for molecular dynamics simulations. It offers both a range of mechanical force fields, such as ff19SB, [67] and the implementations of algorithms necessary for molecular simulations. [70] Amber is used in this project as a part of the SQM2.20 workflow, providing both the ff19SB force field and the implementation required for optimizing the protein component of the P–L complex.

3.2 Selection of Experimental Structures

To obtain experimental structures as references for the protonation states of P-L complexes, neutron diffraction (ND) structures were retrieved from the Protein Data Bank (PDB).

At the time of the search, there were a total of 223 ND structures in the PDB. Successive rounds of filtering were performed in the following order

1. The structure has to have at least one ligand of molecular weight of 300 Da or more to remove small molecule fragments (see 1.3.2).

2. The resolution of the structure is at least 2.5 Å for the ND method.

    - The limit of 2.5 Å is necessary for reliable determination of hydrogen positions as described in Chapter 1.2.2.
    - The X-ray resolution in case of joint structures was high enough for all the entries (better than 2.2 Å in all the other cases), so there was no need to impose other criteria.

3. The protein is not an oxidoreductase, enzyme commision (EC) number 1.*. [85]

    - The criterion was established to exclude metalloproteins, as they may contain radicals and open-shell systems, which constitute a special case requiring a specific treatment.
    - An exception was made for proteins from EC 1.5.1.3 group (3 entries), because they were not metalloproteins.

4. At least one ligand is neither a cofactor, nor a peptide and does not contain metals.

5. The ligand is bound non-covalently.

6. The ligand has at least one protonatable group with pK_a in range $[4; 10]$ as determined by the Epik tool with a default setup. The chemical moieties outside this range will most likely not undergo protonation phenomena.

7. The ligand does not include the amine groups, which lowers the reliability of the protonation in ND structures, as described in Chapter 1.2.4. Also, for a better automatization, other ligands are not allowed in the binding site of the P–L complex.

3.3 System Setup

Prior to running the SQM2.20 workflow, the models of protonated P–L complexes were prepared.

3.3.1 Ligand Preparation

For each ligand, potential protonation variants were prepared using Epikx. The input was taken from RCSB PDB in SDF file format.[3] The the input files were first converted from SDF to MAE format using Schrodinger's sdconvert tool from a CLI.

After that, the epikx was used from a CLI, set as that all protonations of pH in range $[0, 14]$ were included. By default, the state population cutoff is set to 1%, so structures with theoretical occupation among the protonation variants under such cutoff were not included in the protonation variants. Also, by default the tool produces up to 16 structures. The output protonation variants are in MAE format, so they were converted to SDF format by sdconvert tool. The 3D conformation of the ligands had to be reconstructed based on the ligand conformer

[3]At the RCSB page, the file is denoted by label "Structure Data File (Ideal SDF)". For e.g. ligand MT1, the entry page is `https://www.rcsb.org/ligand/MT1` and the SDF file is directly reachable at `https://files.rcsb.org/ligands/download/MT1_ideal.sdf`.

in the P–L complex to fit the binding pose in a P–L complex. Because of this and also for a better visual control of the Epikx results, the SDF file was converted to SMILES format using Obabel tool.

The protonation variants in SMILES format were then used as an input for a custom made script, which employs RDKit's MCS module [86] to match the reconstructed molecule and the template structure obtained from P–L complex and to give the reconstructed molecule the template 3D coordinates. The hydrogen positions on the reconstructed molecule and the template molecule differ, because the reconstructed molecule was a protonation variant of the template. Therefore the hydrogens were treated separately and only the 3D positions of the matching hydrogens were reconstructed according to the template. The positions of the remaining hydrogens were computed automatically by RDKit. The specific positions of hydrogens, however, were changed afterwards, in the annealing step and the optimization step in the SQM2.20 protocol (see Chapter 3.4).

The resulting structures of ligand protonation variants were included in the steps of P–L preparation, so the preparation branched from one process for each initial PDB structure to up to 16 processes for one structure (there were up to 16 protonation variants of each ligand).

The ProToss tool was used to obtain alternative protonation of the ligands by a different tool than Epikx and to get another protonation reference in addition to the neutron structure. The tool was used with default settings using the REST API.

Similarly to ProToss, also Maestro was used to assign hydrogens to the P-L complex as an alternative. Its Protein Preparation Workflow tool was used similarly to as it is described in Chapter 3.3.2, however with hydrogen reassignment option checked in the configuration.

Both ProToss and Maestro prepared protonation variants included in the variants generated by Epikx, therefore the set of protonation variants generated by Epikx was not modified. ProToss and Maestro also prepared protonation of the protein binding site of the protein. This information, however, was not important.

3.3.2 Protein Preparation

The necessary steps to prepare the protein models included:

1. treatment of termini of the protein,
2. modeling of protein loops which are missing in the structure,
3. filling in missing side chains,
4. fixing side chains of asparagine, glutamine and histidine,
5. waters deletion.

These preparation steps were performed using Schrodinger's Protein Preparation Workflow. If missing the C-terminal oxygens were added. The missing loops were modeled using Schrodinger's Prime tool.

One glutamine residue was present in the binding site of the ligand WNU in structure 6BQ8. The structure was checked, however, its structure remained the same after the preparation process as it was in the neutron structure.

The Protein Preparation Workflow was used from the Maestro's GUI. Most of the configuration was left as default with following changed to fit the steps described in previous paragraph:

For binary options, the "☐" symbol is used for unmarked and "✓" is used for marked. For other configurations, the selected value is written after "..." string.

- **Global Settings:**
  - Simulation pH ... 7.4
  - **Small molecules ("hets") to process:**
    - ✓ Detected Ligands
    - ✓ Metals and Ions
    - ☐ Non-Water solvents
    - ☐ Others
- **Preprocess**
  - ☐ Cap termini

  - ✓ Fill in missing side chains
  - ✓ Fill in missing loops
  - **More Options:**
    - ✓ Add terminal oxygens to protein chains
    - ☐ Generate het states (with Epik)

- **Minimize and Delete Waters**
  - Minimize ... Hydrogens only
    - *Note: The minimization is done in the annealing step further on as well.*
  - Delete waters distant from ligands (”hets”) ... 0 Å
    - *Note: All waters should be deleted.*

Atoms of metals far from the active site (further than 10 Å) were excluded from the structure because they will not be considered in the later steps. Other metal atoms were included.

3.4 SQM2.20 Workflow

As described in Introduction 1.6, the sum of the electronic energies (E) and the solvation free energies (ΔG_{solv}) of the P–L complexes were obtained from the SQM2.20 workflow. Workflow SQM2.20 consists of both advanced model preparation and calculation steps as described in Figure 1.3. Some of the steps were performed here, namely

- annealing of hydrogens,
- preparation of the ligand binding site (removal of parts of the structure beyond 10 Å from the ligand),
- structure optimization,
- free energy calculation using PM6-D3H4X/COSMO2 method.

The prepared P–L complexes were treated with the tools included in the in-house SQM2.20 protocol in the same way as described in article [22]. The hydrogens were annealed and then the binding site of the ligand was defined as protein residues within the distance of 10 Å around the ligand. The other protein residues were removed. The remaining protein residues were capped by formyl at N-terminus, N-methyl amide at C-terminus and hydrogen on side chain. Next, the structure was optimized using a QM/MM approach - the ligand was optimized using the QM part, specifically by method PM6-D3H4X/COSMO [57, 58, 62], while the protein constituted the MM part - its 4 Å surroundings of the ligand was optimized using the Amber ff19SB [67] force field. The protein region between 4 - 10 Å was kept rigid. The COSMO approach was employed for the QM optimization of the ligand instead of COSMO2, the enhanced and more recent version of COSMO, as the latter currently lacks the necessary implementation of energy gradient. Finally, the free energies of the complexes were calculated using the PM6-D3H4X/COSMO2 [57, 58, 65] method which was employed using the MOPAC tool. [68]

3.5 Free Energy Comparisons

To compare the computed free energies of P–L complexes with different ligand protonation states, differently protonated water species need to be considered.

As a simple example, a ligand with two protonatable groups, -COOH and -NH_2, can show up to four different protonation variants, which can also have different charges:

1. -COO^-, -NH_2, total charge is -1, also denoted as L^-,
2. -COO^-, -NH_3^+, total charge is 0, also denoted as LH,
3. -COOH, -NH_2, total charge is 0, also denoted as LH, and
4. -COOH, -NH_3^+, total charge is 1, also denoted as LH_2^+.

Ligands with a different total charge have a different number of atoms. Specifically, they differ by one hydrogen cation, a proton, times the difference between their

charges. In the example, the molecule with -COO^-, -NH_2, with total charge of -1, has 2 protons less than the one with -COOH, -NH_3^+ and total charge of $+1$. That disallows to compare the free energies of the systems. To make different total charges comparable, a suitable correction must be added.

In this project, the following values of free energies as computed using PM6-D3H4X/COSMO2 method were used for water and products of its dissociation (RNDr. Jindřich Fanfrlík, Ph. D., personal communication):

- $\Delta G(H_2O) = -59.86$ kcal/mol
- $\Delta G(H_3O^+) = 38.82$ kcal/mol
- $\Delta G(OH^-) = -133.82$ kcal/mol

It is possible to add a molecule of H_3O^+ to a molecule with charge -1, a molecule of H_2O to a molecule with charge 0 and a molecule of OH^- to a system with charge 1 to make the number of atoms in a system equal. However, the energy of products of water dissociation consists of part from the reaction energy of the water dissociation. Therefore this energy needs to be subtracted from the correction. The water dissociation and its energy can be described as follows:

$$\Delta\Delta G_{\text{water dissociation}}: \quad 2H_2O \rightleftharpoons OH^- + H_3O^+$$
$$\Delta\Delta G_{\text{water dissociation}} = \Delta G(OH^-) + \Delta G(H_3O^+) - 2 \times \Delta G(H_2O)$$

One half of $\Delta\Delta G_{\text{water\_dissociation}}$ is subtracted, the final energy of the correction can be written as follows:

$$\Delta\Delta G_{\text{correction}} := \Delta G(H_3O^+) - \Delta G(H_2O) - \frac{1}{2}\Delta\Delta G_{\text{water\_dissociation}} = 86.32\,\text{kcal/mol}$$

Finally, to account for the absence of protons in the system, the correction energy is scaled by the difference between the system's total charge and the reference charge. In this project, the reference charge is set to zero. The example systems are corrected according to this approach as follows:

$$\Delta G_{\text{corr}}(L^-) = \Delta G(L^-) + 1 \times \Delta\Delta G_{\text{correction}}$$
$$\Delta G_{\text{corr}}(LH) = \Delta G(LH) + 0 \times \Delta\Delta G_{\text{correction}}$$
$$\Delta G_{\text{corr}}(LH_2^+) = \Delta G(LH_2^+) - 1 \times \Delta\Delta G_{\text{correction}}$$

This approach is similarly applicable for different charges of the system. The error of the correction is given by the error of the calculated energies of water species.

4 Results

For clarity, the naming convention for a given ligand protonation variant was defined as follows: `[ligand PDB code]_epikx_[variant number]`, where

- **ligand PDB code** refers to the identifier of the ligand used in the selected PDB structure (e.g., MT1, KNI),
- **epikx** indicates that the protonation variant was generated using the Epikx [28] tool from the Schrodinger suite,
- **variant number** denotes the sequential number assigned to the protonation variant as output by Epikx.

For simplicity, the same naming convention is used to refer to the corresponding P–L structure incorporating the specific protonation variant. Since each ligand (e.g., MT1) is associated with a single PDB protein structure (2INQ), the mapping between ligand code and protein structure is bijective, eliminating any ambiguity regarding the protein used in the model.

4.1 Selected Structures

The protein structures used in this theses were selected from the PDB database according to criteria described in Chapter 3.2.[1] The counts of found structures are listed in Table 4.1, together with those which passed the selection criteria and those which did not. The final selected structures and the associated ligands are

- 2INQ [71] and MT1,
- 2ZYE [43] and KNI,
- 4QXK [72] and PCG,
- 6BQ8 [73] and WNU.

More information about the selected structures and ligands is provided in Chapter 1.7.

[1] Data were retrieved from the RCSB [87] branch of PDB on 14th November 2024. Available at `www.rcsb.org`.

Filtering criteria	ND	Rejected
All in PDB	223	-
Ligand mass > 300 Da	70	153
Resolution < 2.5 Å	67	3
Not an oxidoreductase	52	15
Not exotic ligand or cofactor	38	14
No covalent bond	30	8
pK_a in range [4; 10]	16	14
No amine in ligand	5	11
No other ligand in binding site	4	1
Selected structures	4	219

Table 4.1 Number of structures from PDB which passed or were dismissed in the given steps of the filtering process. Each row corresponds to one criteria in Chapter 3.2, please see the Chapter 3.2 for a detailed description.

4.2 Ligand Protonation

Protonation variants of ligands MT1, KNI, PCG and WNU were prepared as described in Chapter 3.3.1. All the variants were generated using Epikx tool from Schrodinger suite. The variants of the ligands are described in detail in corresponding tables:

- MT1 – Table 4.2,
- KNI – Table 4.3,
- PCG – Table 4.4,
- WNU – Table 4.5.

The atom specific descriptions use atom numbering as included in images of the ligands in Chapter 1.7. Epikx offered also different tautomer variants for ligands KNI, PCG and WNU. They are described in following tables:

- KNI – Table 4.6,
- PCG – Table 4.7,
- WNU – Table 4.9.

In addition to the primary protocol based on Epikx, alternative protonation states were generated using two other approaches: the ProToss tool [7], and PPW (Protein Preparation Workflow) [78] with hydrogen reassignment enabled. For clarity and consistency in the following sections, protein structures resulting from the alternative PPW setup are referred to as the "Maestro" structures–named after the GUI software provided by Schrodinger from which this procedure is typically executed.

	acidic		*basic*				
	O28	**O33**	**N1**	**N4**	**N11**	**N14**	**Charge**
MT1_epikx_1	−	−	+	0	0	0	−1
MT1_epikx_2	−	0	0	0	0	0	−1
MT1_epikx_3	0	−	0	0	0	0	−1
MT1_epikx_4	−	0	+	0	0	0	0
MT1_epikx_5	−	−	0	+	0	0	−1
MT1_epikx_6	0	−	+	0	0	0	0
MT1_epikx_7	0	0	0	0	0	0	0
MT1_epikx_8	−	0	0	+	0	0	0
MT1_epikx_9	−	−	0	0	+	0	−1
MT1_epikx_10	0	−	0	+	0	0	0
MT1_epikx_11	0	0	+	0	0	0	+1
MT1_epikx_12	−	−	0	0	0	+	−1
MT1_epikx_13	−	0	0	0	+	0	−1

Table 4.2 Atomic charges for variants of molecule MT1 are listed here, using the same format and description as in Table 4.2. The charges are listed instead of the direct description of protonation on a given atom for simplicity. The number of hydrogens in given group can be deducted from the charge listed here, the connectivity of the atom in a molecule (please see 2D molecule images in Chapter 1.7) and possibly also tautomerization and bond reshuffling in a ring. No such phenomena were produced for the ligand MT1. For other ligands, tables with specific description is included.

	acidic				*basic*				
	O14	**O21**	**O33**	**O39**	**N6**	**N15**	**N22**	**N34**	**Charge**
KNI_epikx_1	0	0	0	0	0	0	0	0	0
KNI_epikx_2	0	0	0	0	+	0	0	0	+1
KNI_epikx_3	−	0	0	0	0	0	0	0	−1
KNI_epikx_4	0	−	0	0	0	0	0	0	−1
KNI_epikx_5	−	0	0	0	+	0	0	0	0
KNI_epikx_6	0	−	0	0	+	0	0	0	0
KNI_epikx_7	0	0	0	−	0	0	0	0	−1
KNI_epikx_8	0	0	0	−	0	0	0	0	−1
KNI_epikx_9	0	0	0	0	0	0	0	+	+1
KNI_epikx_10	0	0	0	0	0	0	+	0	+1
KNI_epikx_11	0	0	0	0	0	+	0	0	+1
KNI_epikx_12	0	0	0	−	+	0	0	0	0
KNI_epikx_13	0	0	0	0	+	0	0	+	+2
KNI_epikx_14	0	0	0	−	+	0	0	0	0
KNI_epikx_15	−	−	0	0	0	0	0	0	−2
KNI_epikx_16	−	0	−	0	0	0	0	0	−2

Table 4.3 Atomic charges for variants of molecule KNI are listed here, using the same format and description as in Table 4.2 Tautomer variants for molecule KNI are included in Table 4.6.

	acidic				*basic*		
	O3	**O12**	**O18**	**N19**	**N15**	**N22**	**Charge**
PCG_epikx_1	−	0	0	0	0	0	−1
PCG_epikx_2	−	0	0	0	0	0	−1
PCG_epikx_3	−	0	−	0	0	0	−2
PCG_epikx_4	−	0	0	0	0	0	−1
PCG_epikx_5	−	0	0	0	+	0	0
PCG_epikx_6	−	0	0	−	+	0	−1
PCG_epikx_7	−	0	0	−	+	+	0
PCG_epikx_8	0	0	0	0	0	0	0
PCG_epikx_9	−	0	0	0	0	+	0
PCG_epikx_10	−	−	0	0	0	0	−2
PCG_epikx_11	0	0	0	0	0	0	0
PCG_epikx_12	0	0	−	0	0	0	−1

Table 4.4 Atomic charges for variants of molecule PCG are listed here, using the same format and description as in Table 4.2. Tautomer states are listed in Table 4.7

	acidic				*basic*		
	O2	**O28**	**O30**	**N3**	**N5**	**N9**	**Charge**
WNU_epikx_1	0	−	0	0	0	0	−1
WNU_epikx_2	0	−	0	0	0	0	−1
WNU_epikx_3	−	−	0	0	0	0	−2
WNU_epikx_4	0	−	0	0	0	0	−1
WNU_epikx_5	0	−	0	0	0	+	0
WNU_epikx_6	0	−	0	−	0	+	−1
WNU_epikx_7	0	0	0	0	0	0	0
WNU_epikx_8	0	−	0	−	+	+	0
WNU_epikx_9	0	−	0	0	+	0	0
WNU_epikx_10	0	−	−	0	0	0	−2
WNU_epikx_11	0	0	0	0	0	0	0
WNU_epikx_12	−	0	0	0	0	0	−1

Table 4.5 Atomic charges for variants of molecule WNU are listed here, using the same format and description as in Table 4.2.

	C13, O14, N15	**C20, O21, N22**	**C32, O33, N34**
KNI_epikx_1	amide	amide	amide
KNI_epikx_2	amide	amide	amide
KNI_epikx_3	imide	amide	amide
KNI_epikx_4	amide	imide	amide
KNI_epikx_5	imide	amide	amide
KNI_epikx_6	amide	imide	amide
KNI_epikx_7	amide	amide	imide
KNI_epikx_8	amide	amide	amide
KNI_epikx_9	amide	amide	imide
KNI_epikx_10	amide	imide	amide
KNI_epikx_11	imide	amide	amide
KNI_epikx_12	amide	amide	imide
KNI_epikx_13	amide	amide	imide
KNI_epikx_14	amide	amide	amide
KNI_epikx_15	imide	imide	amide
KNI_epikx_16	imide	amide	imide

Table 4.6 Tautomer forms of three sites in different variants of the ligand KNI. The atom numbering matches the atom order as seen in Figure 1.10

	Guanine Variant
PCG_epikx_1	1
PCG_epikx_2	2
PCG_epikx_3	3
PCG_epikx_4	3
PCG_epikx_5	1
PCG_epikx_6	1
PCG_epikx_7	1
PCG_epikx_8	1
PCG_epikx_9	1
PCG_epikx_10	1
PCG_epikx_11	2
PCG_epikx_12	3

Table 4.7 Tautomers of generated variants of the ligand PCG. The numbers in "Guanine Variant" column correspond to the variant numbers shown in Table 4.8

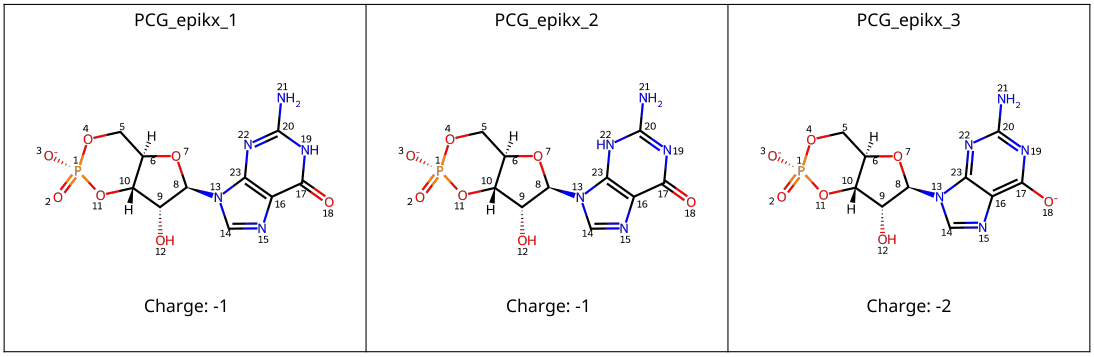


Table 4.8 There are three guanine tautomers included among the generated variants of the ligand PCG. The numbers of the structures are included in Table 4.7 to describe which tautomer is given variant.

	Guanine Variant
WNU_epikx_1	1
WNU_epikx_2	2
WNU_epikx_3	3
WNU_epikx_4	3
WNU_epikx_5	1
WNU_epikx_6	1
WNU_epikx_7	1
WNU_epikx_8	1
WNU_epikx_9	1
WNU_epikx_10	1
WNU_epikx_11	2
WNU_epikx_12	3

Table 4.9 Tautomers of generated variants of the ligand WNU. The numbers in "Guanine Variant" column correspond to the variant numbers shown in 4.10

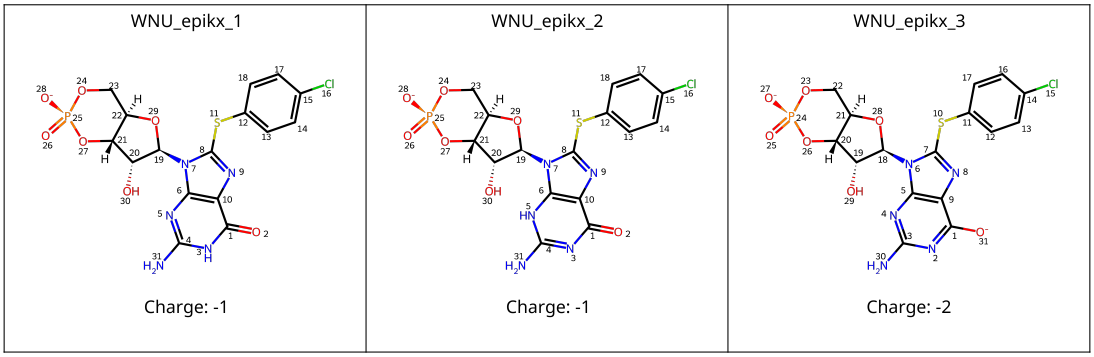


Table 4.10 There are three guanine tautomers included among the generated variants of the ligand WNU. The numbers of the structures are included in Table 4.9 to describe which tautomer is given variant.

4.3 P–L Complex Preparation

Protein models were prepared using the PPW [79], as described in Chapter 3.3.2. Missing side chains and loops were added, water molecules were removed and the ends of peptide chains were treated (oxygens added) so that they are not causing problems in further steps. The protonated variants of the ligands were inserted to the protein binding site in the exact conformation of the experimental structure of the P–L complex using self-written Python script which utilizes the RDKit library [33]. Finally, steps from the SQM2.20 workflow were applied (as described in more detail in Chapter 3.4), which consisted of, hydrogen annealing, removal of most of the protein (all atoms beyond 10 Å from the ligand), capping and structure optimization using QM/MM approach.

4.4 Free Energy Calculations

All protonation variants of ligands in complex with the protein were used for latter energy calculations using PM6-D3H4X/COSMO2 as described in Chapter 3.4. Two calculations failed at the MOPAC level due to incorrect charge of the input ligand protonation variant. Please see Table **??** in attachments for detailed results of all calculations. The comparison of computed energies for all four examined structures can be seen in the corresponding figures:

- 2INQ, MT1 - Figure 4.1,
- 2ZYE, KNI - Figure 4.2,
- 4QXK, PCG - Figure 4.3,
- 6BQ8, WNU - Figure 4.4.

As shown in Figure 4.1, we are able to recognize the correct protonation variant (ND structure) of 2INQ from the energies computed by a physics-based method. All other, non-experimental, protonation variants resulted with higher energy. ProToss tool predicted the protonation of MT1 identically to the experimental one, Maestro predicted a different protonation variant with both carboxyl groups protonated.

We can see similar pattern for the other structures. In all cases, the computed free energy of the experimental protonation is the lowest. For the P–L complex 2ZYE–KNI (Figure 4.2), both ProToss and Maestro also produced correct protonation variants. However, for 4QXK–PCG (Figure 4.3) and 6BQ8–WNU (Figure 4.4), neither ProToss nor Maestro generated correct protonation states.

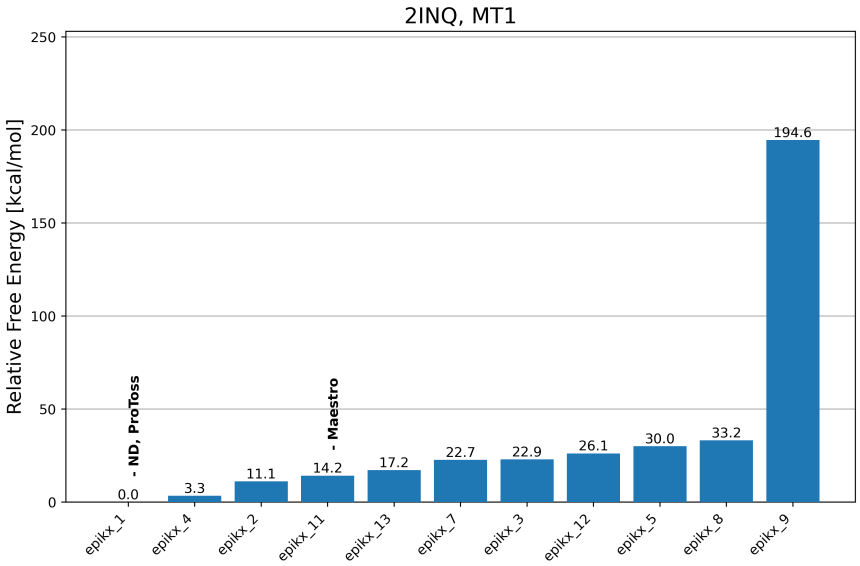


Figure 4.1 Free energies of the 2INQ P–L model with MT1 ligand protonation variants, computed using PM6-D3H4X/COSMO2 and referenced to the zero-charge state. Values are shown relative to the free energy of the protonation state determined by neutron diffraction (ND). Each bar represents one protonation variant of the P–L complex, generated with the Epikx tool. Variants from neutron diffraction (ND), Maestro, and ProToss are labeled above the corresponding bars.

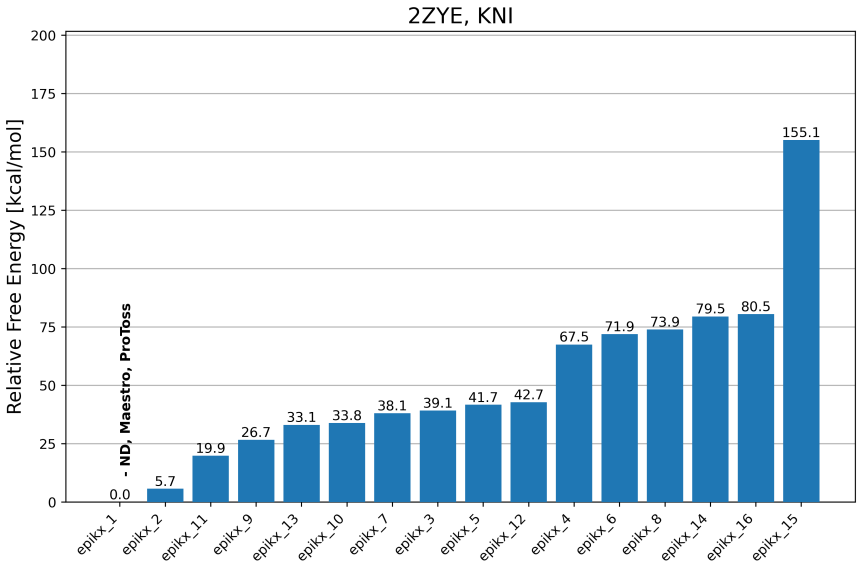


Figure 4.2 Free energies of the 2ZYE P–L model with KNI ligand protonation variants. Detailed description in Figure 4.1 applies here.

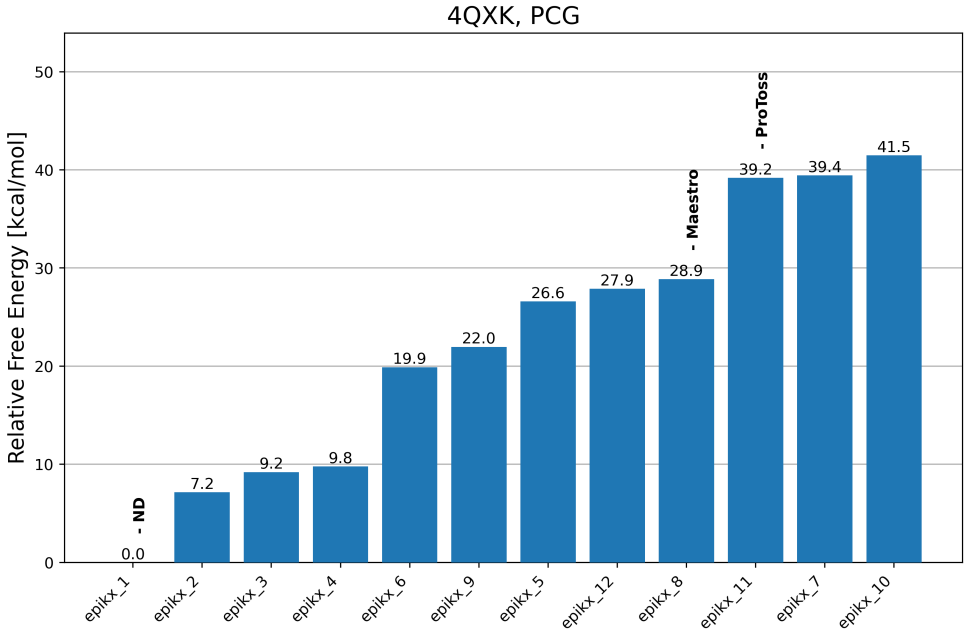


Figure 4.3 Free energies of the 4QXK P–L model with PCG ligand protonation variants. Detailed description in Figure 4.1 applies here.

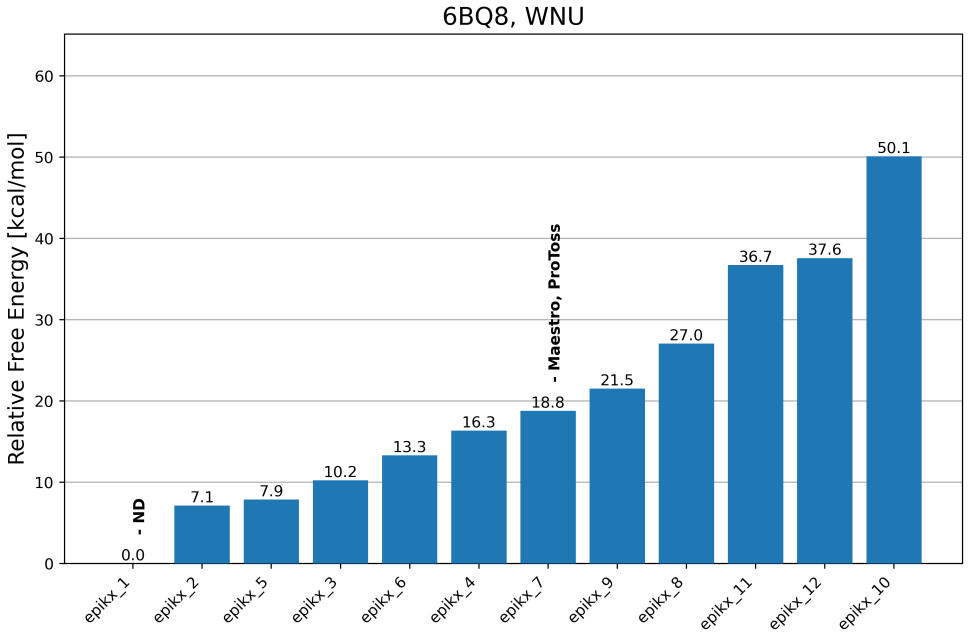


Figure 4.4 Free energies of the 6BQ8 P–L model with WNU ligand protonation variants. Detailed description in Figure 4.1 applies here.

5 Discussion

5.1 Protonation Variants Generation

There are multiple ways to determine protonatable sites. The approach used by Epikx used in this project is based on recognition of protonatable sites using SMARTS patterns. This is a state-of-the-art approach also used by tools for both protonation prediction and pK_a prediction such as MolGpKa [14] or Open Babel [74]. It is a reliable approach for protonations around the physiological pH, however, for extreme pH it may fail. There are not enough experimental data to recognize the pK_a micro-states which is a necessary precondition step in SMARTS site database construction. For purpose of CADD it is, however irrelevant, because such situation is not possible in physiological conditions.

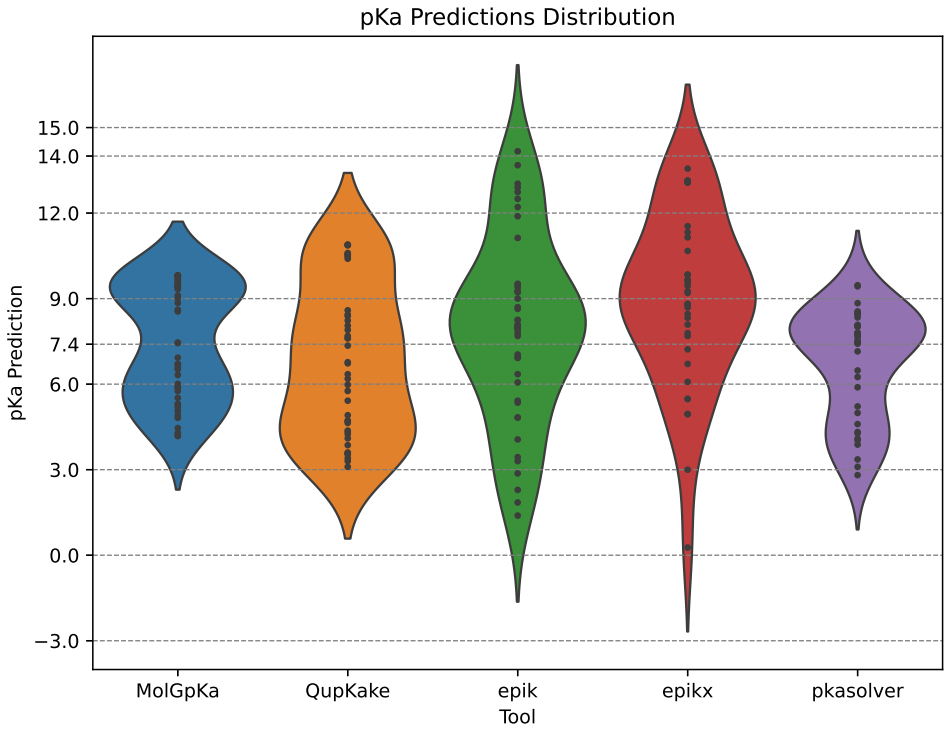


Figure 5.1 Violin plot showing the distributions of pK_a predictions for various pK_a prediction tools. The pK_a values are predicted for the 15 ligands that passed all filtration criteria except for the last two (reliable amines and no other ligand in a binding site). All tools were used with their default settings for pK_a prediction.

There are other protonation tools as well. One of them, Dimorphite, [46] uses an empirical rather than a machine-learning approach to protonation. It enumerates all possible protonation variants on a molecule in a given pH range. The tool uses an experimental database of pK_a values for such task. Alternatively, protonation can be prepared directly on the predicted pK_a by various pK_a tools. However, their results vary widely, as shown in an in-house project of the author of this thesis, the tools for pK_a prediction have quite a high variation regarding the pK_a prediction (see Figure 5.1). The tool Epikx was chosen because it allows the direct preparation of protonation variants of ligands together with the pK_a prediction.

5.2 P–L Complex Protonation

In the experiment setup, the focus was on the problem of ligand protonation within its complex with a protein, while the protonation state of the protein itself was kept fixed. The alternative tools for protonation of P–L complexes, ProToss and Maestro, were found to be unable to accurately prepare ligand protonation in all cases. This limitation may have occurred due to errors in the protein protonation process, which could have subsequently affected the ligand protonation.

5.3 Set of Tested Structures

The set of four selected proteins is rather narrow, hence the results of the SQM method presented in this theses, which was able to correctly predict all four protonation states of ligands in a P–L complex have limited demonstrability. The method needs further testing. The dataset, on which the method may be tested, needs to be of a high quality with a resolved protonation state. There are 12 neutron structures in PDB, which were dismissed from this project due to possibly unreliable protonation of amines in ligands or presence of a cofactor in a binding site of the protein. If the problematic sites of amines are reliably resolved, at least part of those structures may be used for further testing.

All structures included in this project together with structures which passed all

the filtration criteria from Chapter 3.2 besides the presence of unreliable amines and presence of different ligand in a binding site are listed in Attachments **??**.

Another option might be the PL-REX dataset [22] of 15 manually refined and reliable P–L structures with a resolved protonation or the PDBbind dataset [88].

It can be assumed, that the selected P–L structures used in this theses are either non-standard or hard regards their protonation state. If they were simple to solve, the structures would not be solved by neutron diffraction which is more difficult and expensive compared to broadly used X-ray diffraction as described in Chapter 1.2.2.

5.4 Applicability of the Presented Protocol

The approximate runtime of the SQM workflow is in the low tens of minutes for each protonation state on a single CPU. This runtime makes the method conveniently applicable for recognizing the correct protonation variant from several alternatives. For tens of ligands, each with around ten protonation variants, and one protein target, the total computation time can be expected to be under a week. However, since the computations can be run in parallel (with each variant on a separate CPU thread), the number of target ligands process is higher.

6 Conclusions

The general aim of this thesis was to explore available methods for defining protonation states in P–L complexes – a critical prerequisite for physics- and structure based CADD. To this end, we compare software PROTOSS and Maestro from Schrodinger with the newly developed SQM-based scoring function SQM2.20. The following steps were carried out:

- selection of P–L structures from the PDB solved by neutron diffraction which fulfill well-defined criteria of reliably protonated references,
- generation of multiple ligand protonation states using Epikx (Epik 7) and Maestro from the Schrodinger suite and ProToss,
- preparation of protein models using Schrodinger PPW and SQM2.20 protocols,
- free energy calculations of P–L complexes using SQM2.20, comparing them via the free energies of the water species.

P-L Complex	SQM2.20	ProToss	Maestro
2INQ, MT1	✓	✓	×
2ZYE, KNI	✓	✓	✓
4QXK, PCG	✓	×	×
6BQ8, WNU	✓	×	×

Table 6.1 The table shows summary results of this project. The SQM method was able to recognize the correct protonation in case of all four P–L complexes(PDB code, ligand code), the other tools had lower success rate.

In conclusion, the SQM2.20 workflow predicted the lowest (most favorable) energy of the P–L complex as the protonation variant identical to the experimental protonation solved by the neutron diffraction for all four studied structures. The tool ProToss was able to predict the correct protonation for structures 2INQ and 2ZYE and Maestro was only correct in case of 2ZYE. Please see the comprehensive table 6.1.

This work lays the foundation for broader validation of this SQM2.20-based prediction protocol for protonation of P–L complexes. The next steps beyond this thesis include adding ND structures which need manual curation. Further, the strategy will be employed to datasets with well defined structures (except hydrogens), such as PL-REX. [22] These advancements will be detailed in a forthcoming journal article.

This SQM2.20-based protonation preparation pipeline will be further integrated into academic and industrial CADD pipelines, supporting accurate and efficient molecular modeling in CADD applications.

Bibliography

(1) Young, D. C., *Computational Drug Design: A Guide for Computational and Medicinal Chemists*; Wiley-Interscience: Hoboken, NJ, 2009, p 344.

(2) *Drug Design and Discovery*, 4th ed.; Strømgaard, K., Krogsgaard-Larsen, P., Madsen, U., Eds.; CRC Press: Boca Raton, FL, 2009.

(3) Kolář, M.; Hobza, P. Molecular modelling in drug development, Ph.D. Thesis, Prague, Czech Republic: Univerzita Karlova. Přírodovědecká fakulta. Katedra fyzikální a makromolekulární chemie, 2013.

(4) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, P.; Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(5) Bütikofer, M.; Torres, F.; Kadavath, H.; Gämperli, N.; Abi Saad, M. J.; Zindel, D.; Coudevylle, N.; Riek, R.; Orts, J. NMR2-Based Drug Discovery Pipeline Presented on the Oncogenic Protein KRAS. *Journal of the American Chemical Society* **2025**, *147*, 13200–13209.

(6) Catapano, L.; Long, F.; Yamashita, K.; Nicholls, R. A.; Steiner, R. A.; Murshudov, G. N. Neutron crystallographic refinement with REFMAC5 from the CCP4 suite. *Biological Crystallography* **2023**, *79*, 1056–1070.

(7) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *Journal of cheminformatics* **2014**, *6*, 1–12.

(8) Borbulevych, O.; Martin, R.; Tickle, I.; Westerhoff, L. XModeScore: a novel method for accurate protonation/tautomer-state determination using quantum-mechanically driven macromolecular X-ray crystallographic refinement. *Acta Crystallographica Section D-Structural Biology* **2016**, *72*, 586–598.

(9) Afonine, P. V.; Mustyakimov, M.; Grosse-Kunstleve, R. W.; Moriarty, N. W.; Langan, P.; Adams, P. D. Joint X-ray and neutron refinement with phenix.refine. *Acta Crystallographica Section D-Structural Biology* **2010**, *66*, 1153–1163.

(10) Chen, J. C.-H.; Unkefer, C. J. Fifteen years of the Protein Crystallography Station: the coming of age of macromolecular neutron crystallography. *IUCrJ* **2017**, *4*, 72–86.

(11) Wlodawer, A.; Hendrickson, W. A. A procedure for joint refinement of macromolecular structures with X-ray and neutron diffraction data from single crystals. *Acta Crystallographica Section A* **1982**, *38*, 239–247.

(12) Shu, F.; Ramakrishnan, V.; Schoenborn, B. P. Enhanced visibility of hydrogen atoms by neutron crystallography on fully deuterated myoglobin. *Proceedings of the National Academy of Sciences* **2000**, *97*, 3872–3877.

(13) Chen, J. C.-H.; Hanson, B. L.; Fisher, S. Z.; Langan, P.; Kovalevsky, A. Y. Direct observation of hydrogen atom dynamics and interactions by ultrahigh resolution neutron protein crystallography. *Proceedings of the National Academy of Sciences* **2012**, *109*, 15301–15306.

(14) Pan, X.; Wang, H.; Li, C.; Zhang, J. Z.; Ji, C. MolGpka: A web server for small molecule p K a prediction using a graph-convolutional neural network. *Journal of Chemical Information and Modeling* **2021**, *61*, 3159–3165.

(15) Flachsenberg, F.; Ehrt, C.; Gutermuth, T.; Rarey, M. Redocking the PDB. *Journal of Chemical Information and Modeling* **2024**, *64*, 219–237.

(16) Siebenmorgen, T.; Zacharias, M. Computational prediction of protein–protein binding affinities. *WIREs Computational Molecular Science* **2020**, *10*, e1448.

(17) de Ruiter, A.; Oostenbrink, C. Advances in the calculation of binding free energies. *Current Opinion in Structural Biology* **2020**, *61*, Theory and Simulation – Macromolecular Assemblies, 207–212.

(18) Decherchi, S.; Cavalli, A. Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation. *Chemical Reviews* **2020**, *120*, 12788–12833.

(19) Ryde, U.; Söderhjelm, P. Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods. *Chemical Reviews* **2016**, *116*, 5520–5566.

(20) Berman, H. M.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nature Structural Biology* **2003**, *10*, 980.

(21) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative assessment of scoring functions: the CASF-2016 update. *Journal of chemical information and modeling* **2018**, *59*, 895–913.

(22) Pecina, A.; Fanfrlík, J.; Lepšík, M.; Řezáč, J. SQM2. 20: Semiempirical quantum-mechanical scoring function yields DFT-quality protein–ligand binding affinity predictions in minutes. *Nature Communications* **2024**, *15*, 1127.

(23) Jung, W.; Goo, S.; Hwang, T.; Lee, H.; Kim, Y.-K.; Chae, J.-w.; Yun, H.-y.; Jung, S. Absorption Distribution Metabolism Excretion and Toxicity Property Prediction Utilizing a Pre-Trained Natural Language Processing Model and Its Applications in Early-Stage Drug Development. *Pharmaceuticals* **2024**, *17*.

(24) Zhu, H.; Zhou, R.; Cao, D.; Tang, J.; Li, M. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nature Communications* **2023**, *14*, 6234.

(25) ChemAxon Marvin: Chemical Drawing and Visualization Software, `https://chemaxon.com/marvin`, Accessed on April 25, 2025.

(26) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *Journal of Chemical Information and Modeling* **2005**, *45*, 160–169.

(27) Schrödinger, LLC Schrödinger Software Suite, Release 2024-3, New York, NY.

(28) Schrödinger, LLC Schrödinger Release 2024-3: Epik 7 (Epikx), New York, NY, 2024.

(29) Schrödinger, LLC Schrödinger Release 2024-3: LigPrep, New York, NY, 2024.

(30) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **1997**, *46*, 3–25.

(31) Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004**, *1*, 337–341.

(32) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876–877.

(33) Landrum, G. et al. RDKit: Open-source cheminformatics, version Release: 2024.03.5 (Q1 2024), `https://www.rdkit.org`, 2024.

(34) Vacík, J., *Obecná chemie*, 2. vydání, aktualizované; Přírodovědecká fakulta Univerzity Karlovy: Praha, 2017, p 283.

(35) Abarbanel, O. D.; Hutchison, G. R. QupKake: Integrating Machine Learning and Quantum Chemistry for Micro-pKa Predictions. *Journal of Chemical Theory and Computation* **2024**, *20*, 6946–6956.

(36) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation* **2011**, *7*, 525–537.

(37) Schrödinger, I. Schrödinger solutions for small molecule protonation state enumeration and pKa prediction, Accessed: 4th April 2025, 2025.

(38) Todorov, N. P.; Monthoux, P. H.; Alberts, I. L. The Influence of Variations of Ligand Protonation and Tautomerism on Protein−Ligand Recognition and Binding Energy Landscape. *Journal of Chemical Information and Modeling* **2006**, *46*, 1134–1142.

(39) Hofer, F.; Kraml, J.; Kahler, U.; Kamenik, A. S.; Liedl, K. R. Catalytic Site pKa Values of Aspartic, Cysteine, and Serine Proteases: Constant pH MD Simulations. *Journal of Chemical Information and Modeling* **2020**, *60*, 3030–3042.

(40) The Open Babel Team OBPhModel Class Reference, `https://openbabel.org/api/3.0/classOpenBabel_1_1OBPhModel.shtml`, Accessed: 2025-04-14, 2019.

(41) Onufriev, A. V.; Alexov, E. Protonation and pK changes in protein-ligand binding. *Quarterly reviews of biophysics* **2013**, *46*, 181–209.

(42) Vatheuer, H.; Palomino-Hernández, O.; Müller, J.; Galonska, P.; Glinca, S.; Czodrowski, P. Protonation Effects in Protein-Ligand Complexes – A Case Study of Endothiapepsin and Pepstatin A with Computational and Experimental Methods. *ChemMedChem* **2025**, e202400953.

(43) Adachi, M. et al. Structure of HIV-1 protease in complex with potent inhibitor KNI-272 determined by high-resolution X-ray and neutron crystallography. *Proceedings of the National Academy of Sciences* **2009**, *106*, 4641–4646.

(44) The Open Babel Team Open Babel v3.1.0, `https://openbabel.org`, Version 3.1.0, 2020.

(45) Daylight - Chemical Information Systems, I. SMARTS - A Language for Describing Molecular Patterns, PO Box 7737, Laguna Niguel, CA 92607, 2022.

(46) Ropp, P.; Kaminsky, J.; Yablonski, S., et al. Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules. *Journal of Cheminformatics* **2019**, *11*, 14.

(47) Schrödinger, LLC Schrödinger Release 2024-3: Epik, New York, NY, 2024.

(48) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(49) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences* **1992**, *32*, 244–255.

(50) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.

(51) Schrödinger, I. maeparser: Maestro File Parser, `https://github.com/schrodinger/maeparser`, Accessed: 2025-04-22, 2025.

(52) Christensen, A. S.; Kubar, T.; Cui, Q.; Elstner, M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chemical Reviews* **2016**, *116*, 5301–5337.

(53) Jensen, F., *Introduction to Computational Chemistry*, 1st ed.; Wiley: 1998.

(54) Leach, A. R., *Molecular Modelling: Principles and Applications*; Prentice Hall: Harlow, England, 2001, pp 528–234.

(55) Pecina, A.; Eyrilmez, S. M.; Köprülüoğlu, C.; Miriyala, V. M.; Lepšík, M.; Fanfrlík, J.; Řezáč, J.; Hobza, P. SQM/COSMO Scoring Function: Reliable Quantum-Mechanical Tool for Sampling and Ranking in Structure-Based Drug Design. *ChemPlusChem* **2020**, *85*, 2362–2371.

(56) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling* **2007**, *13*, 1173–1213.

(57) Řezáč, J.; Hobza, P. A halogen-bonding correction for the semiempirical PM6 method. *Chemical Physics Letters* **2011**, *506*, 286–289.

(58) Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *Journal of Chemical Theory and Computation* **2012**, *8*, 141–151.

(59) Kříž, K.; Nováček, M.; Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets 3: Repulsive Contacts. *Journal of Chemical Theory and Computation* **2021**, *17*, Published online 2021/03/09, 1548–1561.

(60) Novacek, M.; Rezac, J. PM6-ML: The Synergy of Semiempirical Quantum Chemistry and Machine Learning Transformed into a Practical Computational Method. *Journal of Chemical Theory and Computation* **2025**, *21*, 678–690.

(61) Hostas, J.; Rezac, J.; Hobza, P. On the performance of the semiempirical quantum mechanical PM6 and PM7 methods for noncovalent interactions. *Chemical Physics Letters* **2013**, *568*, 161–166.

(62) Klamt, A.; Schuurmann, G. COSMO - A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradient. *Journal of the Chemical Society-Perkin Transactions 2* **1993**, 799–805.

(63) KLAMT, A. Conductor-Like Screening Model for Real Solvents - A New Approach to the Quantitative Calculation of Solvation Phenomena. *Journal of Physical Chemistry* **1995**, *99*, 2224–2235.

(64) Van Eygen, G.; Echezuria, C.; Buekenhoudt, A.; Coutinho, J. A.; Van der Bruggen, B.; Luis, P. COSMO-RS screening of organic mixtures for membrane extraction of aromatic amines: TOPO-based mixtures as promising solvents. *Green Chemical Engineering* **2025**, *6*, 263–274.

(65) Kriz, K.; Rezac, J. Reparametrization of the COSMO Solvent Model for Semiempirical Methods PM6 and PM7. *Journal of Chemical Information and Modeling* **2019**, *59*, 229–235.

(66) Yurenko, Y.; Muzdalo, A.; Černeková, M.; Pecina, A.; Řezáč, J.; Fanfrlík, J., et al. Multiscale Computational Protocols for Accurate Residue Interactions at Flexible Protein–Protein Interfaces. *ChemRxiv* **2024**, This content is a preprint and has not been peer-reviewed.

(67) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation* **2020**, *16*, 528–552.

(68) Stewart, J. J. P. MOPAC2016, `http://OpenMOPAC.net`, Colorado Springs, CO, USA, 2016.

(69) Stewart, J. J. P. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *International Journal of Quantum Chemistry* **1996**, *58*, 133–146.

(70) Case, D.; Belfon, K.; Ben-Shalom, I.; Brozell, S.; Cerutti, D.; Cheatham, T.; Cruzeiro, V.; Darden, T.; Duke, R.; Giambasu, G., et al. AMBER2020, university of California, San Fransisco. *J. Amer. Chem. Soc* **2020**, *142*, 3823–3835.

(71) Bennett, B.; Langan, P.; Coates, L.; Mustyakimov, M.; Schoenborn, B.; Howell, E. E.; Dealwis, C. Neutron diffraction studies of Escherichia coli dihydrofolate reductase complexed with methotrexate. *Proceedings of the National Academy of Sciences* **2006**, *103*, 18493–18498.

(72) Huang, G. Y.; Gerlits, O. O.; Blakeley, M. P.; Sankaran, B.; Kovalevsky, A. Y.; Kim, C. Neutron Diffraction Reveals Hydrogen Bonds Critical for cGMP-Selective Activation: Insights for cGMP-Dependent Protein Kinase Agonist Design. *Biochemistry* **2014**, *53*, PMID: 25271401, 6725–6727.

(73) Gerlits, O.; Campbell, J. C.; Blakeley, M. P.; Kim, C.; Kovalevsky, A. Neutron Crystallography Detects Differences in Protein Dynamics: Structure of the PKG II Cyclic Nucleotide Binding Domain in Complex with an Activator. *Biochemistry* **2018**, *57*, PMID: 29517905, 1833–1837.

(74) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 33.

(75) Řezáč, J. Cuby – ruby framework for computational chemistry, version 4, `http://cuby4.molecular.cz`, Accessed: 2025-04-26, 2016.

(76) Řezáč, J. Cuby – ruby framework for computational chemistry. *J. Comput. Chem.* **2016**, *37*, 1230–1237.

(77) Johnston, R. C.; Yao, K.; Kaplan, Z.; Chelliah, M.; Leswing, K.; Seekins, S.; Watts, S.; Calkins, D.; Chief Elk, J.; Jerome, S. V., et al. Epik: p K a and Protonation State Prediction through Machine Learning. *Journal of Chemical Theory and Computation* **2023**, *19*, 2380–2388.

(78) Schrödinger, LLC Schrödinger Release 2024-3: Protein Preparation Workflow, New York, NY, 2024.

(79) Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 221–234.

(80) Schrödinger, LLC Schrödinger Release 2024-3: Maestro, New York, NY, 2024.

(81) Schrödinger, LLC Schrödinger Release 2024-3: sdconvert, New York, NY, 2024.

(82) Schrödinger, LLC PyMOL Open-Source Repo. `https://github.com/schrodinger/pymol-open-source`, Accessed: 2025-04-26, 2025.

(83) Schrödinger, LLC Schrödinger Release 2024-3: Prime, New York, NY, 2024.

(84) Jacobson, M.; Pincus, D.; Rapp, C.; Day, T.; Honig, B.; Shaw, D.; Friesner, R. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function and Bioinformatics* **2004**, *55*, 351–367.

(85) Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Research* **2000**, *28*, 304–305.

(86) Landrum, G. et al. RDKit GitHub page - MCS.py docs, `https://github.com/rdkit/rdkit-orig/blob/master/rdkit/Chem/MCS.py`, Accessed on 14th April 2025, 2024.

(87) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.

(88) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. *Accounts of Chemical Research* **2017**, *50*, 302–309.

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
CADD	Computer-Aided Drug Design
CLI	Command-Line Interface
COSMO	COnductor-like Screening MOdel
COSMO-RS	COnductor-like Screening MOdel - Real Solvent
DFT-D	DFT with correction for Dispersion
DFT	Density Functional Theory
DHFR	Dihydrofolate Reductase
EC	Enzyme Commission
GUI	Graphical User Interface
LBDD	Ligand–Based Drug Design
MM	Molecular Mechanics
MNDO	Modified Neglect of Diatomic Overlap
MOPAC	Molecular Orbital PACkage
P–L	Protein–Ligand
PKG	Protein Kinase G
PL-REX	Protein–Ligand Refined EXperiment, a name of a dataset
PM6	Parametric Model 6
PM6-D3H4X	PM6 with corrections for Dispersion, Hydrogen and halogen (X) bonding
PDB	Protein Data Bank
PPW	Protein Preparation Workflow - a tool provided by Schrodinger
PR	HIV-1 protease
QM	Quantum Mechanics
RCSB	Research Collaboratory for Structural Bioinformatics
REST	REpresentational State Transfer
SBDD	Structure-Based Drug Design
SF	Scoring Function
SMILES	Simplified Molecular Input Line Entry System - a standardized string representation of small molecules
SMARTS	SMILES ARbitrary Target Specification
SQM methods	Semiempirical Quantum-Mechanical methods