MASARYKOVA UNIVERZITA
Přírodovědecká fakulta
Ústav biochemie

Diplomová práce

Brno 2021 Bc. Šimon Zouvala

MASARYKOVA
UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV BIOCHEMIE

Metody pro výpočet poloh atomů vodíků v proteinech

Diplomová práce

Bc. Šimon Zouvala

Vedoucí práce: doc. RNDr. Radka Svobodová, Ph.D., RNDr. Tomáš Raček

Brno 2021

Bibliografický záznam

Autor:	Bc. Šimon Zouvala Přírodovědecká fakulta, Masarykova univerzita Ústav biochemie
Název práce:	Metody pro výpočet poloh atomů vodíků v proteinech
Studijní program:	Biochemie
Studijní obor:	Chemoinformatika a bioinformatika
Vedoucí práce:	doc. RNDr. Radka Svobodová, Ph.D., RNDr. Tomáš Raček
Akademický rok:	2020/2021
Počet stran:	ix + 61
Klíčová slova:	proteiny, 3D struktura, souřadnice atomů vodíku, parciální atomové náboje

Bibliographic Entry

Author:	Bc. Šimon Zouvala Faculty of Science, Masaryk University Department of biochemistry
Title of Thesis:	Methods for calculation of hydrogen atom coordinates in proteins
Degree Programme:	Biochemistry
Field of Study:	Chemoinformatics and bioinformatics
Supervisor:	doc. RNDr. Radka Svobodová, Ph.D., RNDr. Tomáš Raček
Academic Year:	2020/2021
Number of Pages:	ix + 61
Keywords:	proteins, 3D structure, coordinates of hydrogen atoms, partial atomic charges

Abstrakt

V této diplomové práci se věnujeme metodám výpočtu souřadnic vodíků v proteinových strukturách. Atomy vodíků tvoří téměř polovinu všech atomů ve struktuře proteinu a hrají důležitou roli pro některé druhy aplikací, například výpočet parciálních atomových nábojů. V proteinových strukturách však často chybí vodíky, jelikož velká část struktur v Protein Data Bank je stanovena metodami rentgenové krystalografie nebo kryoelektronové mikroskopie, přičemž tyto metody polohy atomů vodíků neurčují.

V rámci své práce jsem provedl rešerši nástrojů pro výpočet souřadnic vodíků, uvedené nástroje jsem v práci popsal a provedl jejich porovnání. Kromě výčtu nástrojů na získání souřadnic vodíků jsou v této práci některé nástroje srovnány mezi sebou. Jedná se o softwarové nástroje PyMOL, HAAD, Reduce, Chimera, OpenBabel a GROMACS. Srovnání bylo provedeno na základě rychlosti a přesnosti výpočtu vodíků. Dále byl ještě analyzován vliv vypočítaných vodíků na parciální atomové náboje. Konkrétně šlo o porovnání hodnot nábojů na strukturách s dopočítanými vodíky a na strukturách s vodíky určenými experimentálně.

Ve srovnání si nejlépe vedly nástroje PyMOL a Reduce, které je možno integrovat do webového nástroje Atomic Charge Calculator II (ACC II), sloužícího k výpočtu nábojů.

Abstract

In this thesis, we deal with methods for calculating the coordinates of hydrogen atoms in protein structures. Hydrogen atoms make up almost half of all atoms in the protein structure and play an important role in some types of applications, such as the calculation of partial atomic charges. However, hydrogens are often absent in structures, as much of the structures found in the Protein Data Bank are determined by X-ray crystallography or cryoelectron microscopy methods, which do not determine the positions of hydrogen atoms.

As part of my work, I searched for tools for calculating the coordinates of hydrogen atoms, I described these tools in the work and compared them. In addition to the list of tools for obtaining hydrogen coordinates, some of them are compared in this work. These are the software tools PyMOL, HAAD, Reduce, Chimera, OpenBabel, and GROMACS. The comparison was made on the basis of the speed and accuracy of the hydrogen calculation. Furthermore, the effect of calculated hydrogens on partial atomic charges was analyzed. Specifically, it was a comparison of charge values on structures with calculated hydrogen and on structures with hydrogen determined experimentally.

The PyMOL and Reduce tools, which can be integrated into the Atomic Charge Calculator II (ACC II) web tool for charge calculation, performed best in comparison.

MUNI
SCI

MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
KOTLÁŘSKÁ 2, 611 37 BRNO
IČ: 00216224
DIČ: CZ00216224

ZADÁNÍ
DIPLOMOVÉ PRÁCE

Akademický rok: 2020/2021

Ústav:	Ústav biochemie
Student:	Bc. Šimon Zouvala
Program:	Biochemie
Obor:	Chemoinformatika a bioinformatika

Ředitel *ústavu* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje diplomovou práci s názvem:

Název práce:	Metody pro výpočet poloh atomů vodíků v proteinech
Název práce anglicky:	Methods for calculation of hydrogen atom coordinates in proteins
Jazyk závěrečné práce:	čeština

Oficiální zadání:
Díky vysoce výkonným metodikám pro stanovování proteinové struktury máme k dispozici velké množství dat o struktuře jednotlivých proteinů. Tato data jsou uložena v Protein Data Bank (PDB) a jsou volně přístupná. Proteinové struktury jsou důležitým vstupem pro další analýzy a simulace. Pro některé druhy aplikací (např. výpočet parciálních atomových nábojů) je nutné, aby proteinové struktury obsahovaly i souřadnice atomů vodíků. Ty však ve většině struktur z Protein Data Bank obsaženy nejsou, protože velká část těchto struktur byla stanovena metodami rentgenové krystalografie nebo kryoelektronové mikroskopie, které polohy atomů vodíků neurčují. Tato diplomová práce zaměřuje na metody pro výpočet poloh atomů vodíků v proteinech, u nichž informace o souřadnicích vodíků chybí. Cíle diplomové práce jsou:

- Rešerše metod a softwarových nástrojů pro výpočet poloh atomů vodíků v 3D strukturách proteinů
- Seznámení se s jednotlivými softwarovými nástroji a jejich funkcionalitou
- Vytvoření datové sady 3D struktur proteinů z PDB obsahujících informace o polohách atomů vodíků (půjde o proteiny získané pomocí nukleární magnetické rezonance)
- Analýza kvality a efektivity jednotlivých metod a softwarových nástrojů

Vedoucí práce:	doc. RNDr. Radka Svobodová, Ph.D. RNDr. Tomáš Raček
Datum zadání práce:	30. 4. 2021
V Brně dne:	31. 5. 2021

Zadání bylo schváleno prostřednictvím IS MU.

Bc. Šimon Zouvala, 18. 5. 2021
doc. RNDr. Radka Svobodová, Ph.D.

RNDr. Tomáš Raček, 20. 5. 2021
doc. Mgr. Jan Havliš, Dr., 21. 5. 2021

Poděkování

Na tomto místě bych chtěl poděkovat vedoucím mé diplomové práce doc. RNDr. Radce Svobodové, Ph.D. a RNDr. Tomášovi Račkovi za veškerou pomoc, trpělivost a cenné rady při tvorbě a sepisování mé diplomové práce. Dále bych chtěl poděkovat Mgr. Ondřeji Schindlerovi za pomoc při výpočtu parciálních atomových nábojů. Nakonec bych chtěl poděkovat své rodině za veškerou podporu a trpělivost při mém studiu.

Prohlášení

Prohlašuji, že jsem svoji diplomovou práci vypracoval samostatně pod vedením vedoucích práce s využitím informačních zdrojů, které jsou v práci citovány.

Brno 31. května 2021

. . . . . .

Bc. Šimon Zouvala

Obsah

Úvod **1**

1. Teorie **2**
1.1 Metody určování 3D struktury molekul – problém s polohou vodíků 2
1.1.1 Důležitost vodíků ve 3D struktuře 2
1.1.2 Důvody, proč nejsou polohy vodíku ve 3D struktuře uvedeny 3
1.1.3 Nukleární magnetická rezonance 3
1.1.4 Kryoelektronová mikroskopie 4
1.1.5 Rentgenová krystalografie 5
1.2 Metody dopočítávání poloh vodíků 5
1.3 Molekulová mechanika 6
1.4 Parciální atomové náboje 6

2. Metody **8**
2.1 Protein Data Bank 8
2.2 Statistické výpočty 9
2.2.1 Metriky porovnání 3D struktury 9
2.2.2 Metriky porovnání parciálních atomových nábojů 9
2.3 Silová pole 9
2.3.1 AMBER 10
2.3.2 CHARMM 10
2.3.3 GROMOS 10
2.4 MACH 11
2.5 Programy na dopočítání vodíků 11
2.5.1 Chimera 12
2.5.2 PyMOL 13
2.5.3 HAAD 14
2.5.4 Reduce 15
2.5.5 OpenBabel 15
2.5.6 GROMACS 17
2.5.7 CNS 19
2.5.8 Amber 20
2.5.9 CORINA 21
2.5.10 MCEE 21

2.5.11 WHAT IF . . . 22
2.5.12 YASARA . . . 22
2.5.13 CHARMM . . . 23
2.5.14 H++ . . . 24
2.5.15 Swiss-PdbViewer . . . 24

3. Implementace . . . 26
3.1 Požadavky na nástroje . . . 26
3.2 Automatizace . . . 27
3.3 Spouštění jednotlivých nástrojů před automatizací . . . 28
3.4 Výpočet parciálních nábojů . . . 30
3.5 Úprava výsledných souborů . . . 31
3.5.1 Automatické spárování vypočítaných a původních vodíků . . . 31
3.5.2 Převod do SDF . . . 31

4. Výsledky a diskuse . . . 33
4.1 Dataset . . . 33
4.1.1 Shrnutí informací o datové sadě . . . 34
4.2 Porovnání nástrojů . . . 34
4.2.1 Srovnání rychlosti . . . 34
4.2.2 Srovnání přesnosti . . . 35
4.2.3 Srovnání nábojů . . . 43
4.3 Diskuse . . . 44
4.3.1 Možnost využití v rámci ACC II . . . 47

Závěr . . . 48

Seznam použité literatury . . . 49

Příloha 57

A. Číselné parametry pro spuštění nástroje Gromacs . . . 57

B. Elektronické přílohy . . . 59

Úvod

Díky vysoce výkonným metodikám pro stanovování proteinové struktury máme k dispozici velké množství dat o struktuře jednotlivých proteinů. Tato data jsou uložena v Protein Data Bank (PDB) [1] a jsou volně přístupná. Proteinové struktury jsou důležitým vstupem pro různé chemické, biochemické, bioinformatické a biologické analýzy a další analýzy a simulace. Pro některé druhy aplikací (např. výpočet parciálních atomových nábojů [2, 3, 4]) je nutné, aby proteinové struktury obsahovaly i souřadnice atomů vodíků. Ty však ve většině struktur z Protein Data Bank obsaženy nejsou, protože velká část těchto struktur byla stanovena metodami rentgenové krystalografie nebo kryoelektronové mikroskopie, které polohy atomů vodíků neurčují.

Důležitost vodíků spočívá v tom, že tvoří téměř polovinu všech atomů v proteinové struktuře a rovněž hrají klíčovou roli při stabilizaci nativního stavu proteinu prostřednictvím hydrofobních interakcí a vodíkových vazeb [5].

Proto se tato diplomová práce se zaměřuje na metody pro výpočet poloh atomů vodíků v proteinech, u nichž informace o souřadnicích vodíků chybí.

Jednotlivé cíle této práce jsou:

- Rešerše metod a softwarových nástrojů pro výpočet poloh atomů vodíků v 3D strukturách proteinů
- Seznámení se s jednotlivými softwarovými nástroji a jejich funkcionalitou
- Vytvoření datové sady 3D struktur proteinů z PDB obsahujících informace o polohách atomů vodíků (půjde o proteiny získané pomocí nukleární magnetické rezonance)
- Analýza kvality a efektivity jednotlivých metod a softwarových nástrojů

Kapitola 1

Teorie

V této kapitole jsou popsány metody na určení 3D struktury molekul a problémy s určením polohy vodíků pomocí těchto metod. Dále se tato kapitola zabývá principy molekulové mechaniky a vysvětluje význam parciálních atomových nábojů.

1.1 Metody určování 3D struktury molekul – problém s polohou vodíků

1.1.1 Důležitost vodíků ve 3D struktuře

Vodíky tvoří téměř polovinu všech atomů v molekulách bílkovin a hrají důležitou roli v kinetice skládání (foldingu) proteinů a nativního stavu proteinu prostřednictvním hydrofobních interakcí a vodíkových vazeb [5]. Protože je vodíků v proteinech velké množství, je často užitečné je vyloučit v rámci některých časově náročných analýz a simulací. Nicméně jsou typy analýz, výpočtů a simulací, kde vodíky zanedbány být nemohou, protože by pak jejich výsledky byly nepoužitelné [6].

Důvody, proč jsou vodíky důležité, jsou následující:

Nepolární atomy vodíku v alkylových a aromatických skupinách přispívají k hydrofobním interakcím, zatímco polární atomy vodíku se přímo účastní vodíkových vazeb.

Vodíkové atomy tedy zprostředkovávají řadu důležitých interakcí. S ohledem na jejich energetický přínos, jsou důležité například ve studiích analýzy interakcí ligand-protein a protein-protein [5].

Přesné umístnění atomů vodíků hraje důležitou roli, jelikož vodíkové atomy obvykle jsou koncovými atomy v rámci struktury proteinu a jsou tedy hlavní složkou povrchu molekuly. Navíc koncentrace vodíkových iontů, neboli pH, je kontrolní proměnná pro chemické reakce, proto hrají atomy vodíku klíčovou roli v biologii a medicíně [7].

Vodíky jsou tedy nezbytné pro získání realistického pohledu na strukturu na úrovni atomů a analýzu interakcí vodíkových vazeb [5].

Tato vodíková vazba je způsobena elektrostatickou interakcí mezi částečně kladně nabitým atomem vodíku připojeným k vysoce elektronegativnímu atomu a dalším blízkým elektronegativním atomem [8]. Jinými slovy vodíkové vazby jsou typem interakce dipól-dipól vytvořeným mezi protonem skupiny X-H, kde X je elektronegativní atom, a dalšími elektronegativními atomy obsahujícími dvojici nevazebných elektronů [9]. Vodíkové vazby

jsou klasifikovány hlavně na základě síly interakce měřené velikostí interakčního potenciálu. Obvykle se rozlišují tři typy vodíkových vazeb a to slabé, střední a silné vazby [10].

1.1.2 Důvody, proč nejsou polohy vodíku ve 3D struktuře uvedeny

Hlavním důvodem absence vodíků v 3D struktuře je jejich způsob stanovení. Některé metodiky analýzy 3D struktury nedokáží kvůli malému rozlišení atomy vodíku detekovat. Mezi tři nejčastěji používané metodiky analýzy 3D struktury patří nukleární magnetická rezonance (NMR), která je blíže popsaná v části 1.1.3, kryoelektronová mikroskopie (cryo-EM) popsaná v části 1.1.4 a rentgenová krystalografie, jež mezi zmíněnými technikami patří k nejvíce používaným [11]. Věnuje se jí část 1.1.5.

1.1.3 Nukleární magnetická rezonance

Nukleární magnetická rezonance (NMR) [12] je nástrojem pro určení 3D struktury molekul v roztocích. Obecně platí, že NMR spektroskopie zkoumá lokální magnetické prostředí jádra s nenulovým magnetickým momentem ve smyslu jeho chemického posunu.

Nukleární magnetická rezonance vychází ze skutečnosti, že jádra atomů mají magnetické vlastnosti, které lze využít k získání chemických informací [13]. NMR je tedy technika spektroskopie, která je založena na absorpci elektromagnetického záření v radiofrekvenční oblasti 4 až 900 MHz jader atomů [14].

Jedná se o metodu, která dokáže stanovit struktury atomů v rámci biomakromolekul ve vodných roztocích za téměř fyziologických podmínek. NMR je technikou umožňující pro přímou detekci vodíkových atomů a vodíkových vazeb [13].

Princip NMR spočívá v tom, že většina jader má spin a všechna jádra jsou elektricky nabitá. Pokud je použito externí magnetické pole, je možný přesun energie mezi základní energií na vyšší energetickou hladinu. Přenos energie probíhá na vlnové délce, která odpovídá rádiovým frekvencím, a když se spin vrátí na svou základní úroveň, je energie emitována na stejné frekvenci. Signál, který odpovídá tomuto přenosu, se měří mnoha způsoby a zpracovává se, aby se získalo NMR spektrum pro dané jádro.

Fungování NMR je založeno na tom, že vzorek je umístěn v magnetickém poli. Poté je produkován NMR signál excitací vzorku jader rádiovými vlnami do nukleární magnetické rezonance, která je detekována citlivými rádiovými přijímači. Intramolekulární magnetické pole kolem atomu v molekule mění rezonanční frekvenci, čímž poskytuje přístup k podrobnostem o struktuře atomů v molekulách. Protože spektra jsou jedinečná nebo vysoce charakteristická pro jednotlivé sloučeniny, je NMR spektroskopie jednoznačnou metodou k identifikaci monomolekulárních organických sloučenin. Kromě identifikace poskytuje NMR spektroskopie podrobné informace o struktuře, dynamice, reakčním stavu a chemickém prostředí molekul [14].

Jinými slovy NMR umožňuje, že interakce jader a jeho okolí lze použít k získání informací o struktuře molekuly. Běžný přístup určování struktury pomocí NMR se dá klasifikovat na dva typy experimentů. Prvním typem lze přiřadit naměřené frekvence příslušným atomům a jeho výsledkem je tedy tabulka udávající rezonanční frekvence jednotlivých jader v molekule. Každé jádro v tabulce je tak označeno vlastní rezonanční frekvencí. Druhý

typ experimentu opět využívá přiřazování, ale tentokrát se již nevztahuje k topologii, ale k trojrozměrné struktuře. Výsledkem druhého typu experimentu je opět tabulka, která tentokrát shrnuje poznatky o 3D struktuře. Kombinace výsledků experimentů prvního a druhého typu nám však umožní stanovit, že takto blízko jsou ve struktuře například vodíky peptidových vazeb osmé a dvacáté aminokyseliny studovaného proteinu [15].

Nukleární magnetická rezonance je tedy technikou, která dokáže určit atomy vodíků při analýze. Nicméně tuto metodu lze použít jen pro stanovení struktury proteinů s nižším počtem reziduí. Pro rozsáhlejší proteiny tato metoda není aplikovatelná a proto je potřeba polohy vodíků rovněž dopočítávat.

1.1.4 Kryoelektronová mikroskopie

Jednočásticová Kryoelektronová mikroskopie(cryo-EM) [16] je metoda pro určování struktur biologických makromolekul [17]. Současná cryo-EM je díky technickým inovacím schopna řešit makromolekulární struktury při rozlišení, jež je blízké rozměrům atomu [16]. V roce 2017 byla dokonce udělena Nobelova cena za chemii Jacquesovi Dubochetovi, Joachimovi Frankovi a Richardu Hendersonovi za vývoj kryoelektronové mikroskopie pro stanovení struktury biomolekul v roztoku ve vysokém rozlišení [18].

V rámci cryo-EM se suspenze analyzovaných bílkovin zmrazí na nosných mřížkách transmisního elektronového mikroskopu (TEM), které jsou potaženy uhlíkovým filmem s pravidelnou perforační řadou. Mřížka se vzorky se poté okamžitě vysaje filtračním papírem s cílem vytvořit na mřížce film pufru nebo proteinu, který bude po zmrazení dostatečně tenký, aby takto vzniklým „ledem“ mohl elektronový paprsek proniknout. Optimalizace tloušťky ledu je zásadním krokem při přípravě vzorku, protože silnější vrstvy ledu zvyšují pravděpodobnost, že dopadající elektron projde nějakými případnými rozptyly, a tím sníží kvalitu obrazu. V případě extrémní tloušťky ledu elektronový paprsek vůbec neproniká.

Po vysátí se mřížka rychle ponoří do lázně kapalného etanu, což je velmi účinný kryogen, který dostatečně rychle zmrazí vodu, aby se zabránilo tvorbě ledových krystalů. Vytvoření skelné vrstvy ledu je základním krokem v cryo-EM a zachovává objekt zájmu v téměř nativním stavu. Výsledná vrstva skelného ledu se suspendovanými bílkovinnými molekulami pak musí během skladování a zobrazování v transmisním elektronovém mikroskopu zůstat blízko teploty kapalného dusíku.

Nevýhoda zmrazených hydratovaných přípravků je, že jsou extrémně citlivé na radiační poškození. Klíčovým aspektem cryo-EM je proto použití nízké elektronové dávky k minimalizaci ozáření vzorku před pořízením snímků. Předem kalibrované zobrazovací režimy mají tři nastavení. První nastavení spočívá ve vyhledání oblasti zájmu při malém zvětšení a malé elektronové dávce. Další možnost je zaostření při velkém zvětšení a velké elektronové dávce na sousední oblast. Posledním nastavením je zaznamenání obrazu při velkém zvětšení a při známé elektronové dávce. Jednou z několika inovací, které povýšily cryo-EM na metodu s vysokým rozlišením, je automatizace tohoto procesu [19].

Na rozdíl od rentgenové krystalografie, kde jsou atomy vodíků prakticky neviditelné, může cryo-EM identifikovat atomy vodíků při nízkém rozlišení, jelikož jádro atomu dobře odvádí elektrony. Bohužel, hlavním problémem cryo-EM je poměr signálu k šumu. Kvůli křehkosti proteinů musí být elektrony vyslané na ně pomalé a málo vzdálené od něho, aby se protein nerozpadl. Výsledné obrázky pak mohou být rozmazané [20].

1.1.5 Rentgenová krystalografie

Rentgenová krystalografie [21] je nejčastější technikou pro stanovení struktury proteinů a biologických makromolekul.

Cílem rentgenové krystalografie je získat z krystalu trojrozměrnou molekulární strukturu. K získání krystalu je nutné vyčistit vzorek zkoumané molekuly tak, aby při vysoké koncentraci krystalizoval. Výsledkem jsou difrakční difraktogramy, jejichž algoritmickým zpracováním lze určit souřadnice jednotlivých atomů [21].

Atomy vodíku však rozptylují rentgenové záření jen slabě. Z tohoto důvodu je obtížné přesně detekovat vodíky pomocí rentgenových paprsků a k jejich přesné lokalizaci se tato metoda nepoužívá [7].

Nevýhodou rentgenové krystalografie je rovněž to, že rozlišení je omezeno čistotou krystalu a vytváření těchto vzorků je časově velmi náročné a trvá až měsíce či dokonce roky. Mezi další nevýhodu lze zařadit i to, že některé proteiny se těžko krystalizují [22].

Jelikož v rentgenové krystalografii není možno odhalit atomy vodíků, chybí tyto atomy ve většině souborů v PDB [23].

1.2 Metody dopočítávání poloh vodíků

Jelikož nejčastější technikou analýzy je rentgenová krystalografie, která oproti NMR nedokáže určit atomy vodíků, je potřeba definovat polohy vodíků jiným způsobem. Existuje několik algoritmů sloužících k predikci pozic atomů vodíků. Obecně lze algoritmy pro dopočítávaní vodíků rozdělit do následujících kategorií:

- Geometrická metoda – umístění atomů vodíků pouze na základě geometrie.
- Vyhledávací metoda – systematické vyhledávání poloh vodíků pomocí silových polí molekulové mechaniky.
- Energetická metoda – minimalizace energie molekulové mechaniky.
- Heuristická metoda – optimalizace vodíkových vazeb a dalších vlastností pomocí heuristiky nebo empirických energetických funkcí. [6]

Jedná se však o volnou kategorizaci, jelikož některé nástroje kombinují více metod. Podrobnější popis vlastností některých nástrojů je uveden blíže v části 2.5. Nicméně obecně lze říci, že jsou atomy vodíku nejprve umístěny pomocí restrikce lokální geometrie a poté jsou jejich polohy optimalizovány konformačním vyhledáváním, které je prováděno energetickými funkcemi nebo heuristickými přístupy [5].

Vzhledem ke 3D souřadnicím těžkých atomů lze k určení pozic atomů vodíků použít prostorovou orientaci hybridního orbitalu. Poloha vodíku připojeného k těžkému atomu je určena ve vztahu k ostatním těžkým atomům připojeným ke stejnému centrálnímu těžkému atomu. V zásadě platí, že pokud těžký atom má sp^3 hybridní orbital, čtyři spojené atomy mají sklon tvořit čtyřstěn soustředěný se středem v tomto těžkém atomu. Pokud má sp^2 hybridní orbital, tři atomy s ním spojené mají sklon tvořit trojúhelník s těžkým atomem uprostřed. Při sp hybridním orbitalu má těžký atom a dva vázané atomy tendenci tvořit trojúhelník se třemi atomy na jeho vrcholech [24].

1.3 Molekulová mechanika

Molekulová mechanika (MM) [25] je empirická metoda pro výpočet vlastností molekul, jako je geometrie molekuly, potenciální energie, dipólový moment a vibrační frekvence. Různé MM softwarové nástroje používají různé sady parametrů a reprodukují tyto fyzikální vlastnosti s různým stupněm přesnosti.

Geometrické parametry používané v programech molekulové mechaniky jsou často odvozeny z krystalografických dat. Prediktivní hodnota metody je však omezena datovými sadami použitými k odvození parametrů. Přechodové stavy, neobvyklé stavy hybridizace a další elektronové efekty mohou vyžadovat zvláštní zacházení. Bez ohledu na tyto výhrady má MM značný úspěch [26].

Metody molekulové mechaniky jsou méně výpočetně náročné, rychlé a jsou schopny zvládnout velmi velké systémy včetně enzymů. Molekulová mechanika může poskytnout velmi přesné hodnoty energie, pokud jsou k dispozici správné parametry. Molekulová mechanika se rovněž využívá k optimalizaci geometrického uspořádání molekuly.

Molekulová mechanika využívá potenciálovou funkci, která popisuje potenciální energii molekuly na základě souřadnic jejích jednotlivých atomů. Grafem této rovnice je takzvaná hyperplocha potenciální energie. Potenciálová funkce spolu s jejími parametry se nazývá silové pole. Parametry silového pole mohou být optimalizovány tak, aby bylo dosaženo nejlepšího přizpůsobení vypočítaných a experimentálních vlastností molekul, jako je geometrie, konformační energie nebo jiné vlastnosti. Předpokládá se, že odpovídající parametry a silové konstanty jsou přenositelné z jedné molekuly do druhé.

Pro minimalizaci energie molekulové mechaniky je potřeba zahrnout postupné iterační výpočty, kde je počáteční konformace podrobena geometrické optimalizaci. Všechny souřadnice definující geometrii systému jsou upravovány malými přírůstky, dokud celková energie struktury nedosáhne lokálního minima. Lokální minimum však nemusí být globálním minimem. Vyhledávací metody se poté používají k hledání dalších konformací. Tyto metody mohou zahrnovat například systematické vyhledávání, které postupně mění všechny otočné vazby, aby se prozkoumal celý konformační prostor molekuly. Dále mohou metody obsahovat vzdálenostní geometrii nebo další přístupy náhodného vzorkování [25].

1.4 Parciální atomové náboje

Parciální atomové náboje [27] hrají důležitou roli ve výpočetní chemii, chemoinformatice a bioinformatice a nacházejí zde mnoho aplikací: Obvykle se uplatňují během molekulárních simulací [28], ve virtuálním screeningu, jako deskriptory molekulární struktury v QSAR (Quantitative Structure-Activity Relationship) [29] nebo QSPR (Quantitative Structure-Property Relationship) modelech [30].

Parciální atomové náboje charakterizují rozložení elektronové hustoty ve struktuře molekuly a jsou to reálná čísla. Jejich vznik je ovlivněn nerovnoměrným rozdělením elektronové hustoty v molekule, což je způsobeno tím, že elektrony jsou více přitahovány prvky s větší mírou elektronegativity. Existuje řada metod jak toto rozdělení prezentovat. Nejčastějším způsobem je komplex dílčích bodových nábojů nacházejících se v molekule. Jako parciální atomové náboje jsou pojmenovány ty náboje, které jsou lokalizovány ve středu atomů.

Parciální atomové náboje není možno experimentálně pozorovat, z tohoto důvodu je možné tyto náboje napočítat vícero metodami [31]. Nejznámější metodou pro výpočet nábojů je kvantová mechanika, která poskytuje kvalitní výsledky, ale je velmi časově náročná. Rychlejší alternativou jsou pak empirické metody pro výpočet nábojů. Příkladem těchto metod jsou: electronegativity equalization method (EEM) [32], Charge Equilibration (QEq) [33], Split-charge Equilibrium method (SQE) [34] apod.

Kapitola 2

Metody

V této kapitole je popsána databáze Protein Data Bank a způsob porovnávání 3D struktur. Dále jsou zde popsány některé programy na dopočítání vodíků a několik silových polí, které jsou využívány v těchto programech.

2.1 Protein Data Bank

Protein Data Bank (PDB, pdb.org) [1] byla založena v roce 1971 jako první otevřený zdroj 3D struktur makromolekul v biologii. PDB nadále slouží jako jednotné globální úložiště údajů o 3D struktuře na atomové úrovni a obsahuje experimentálně určené struktury proteinů, DNA, RNA a jejich komplexů s kovovými ionty, léky, a další malé molekuly volně dostupné bez omezení použití. Od roku 2003 je PDB spravována společně konsorciem Worldwide Protein Data Bank (wwPDB), jehož členy jsou Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB, rcsb.org [35]), Protein Data Bank in Europe (PDBe, pdbe.org [36]), Protein Data Bank Japan (PDBj, pdbj.org [37]) a Biological Magnetic Resonance Data Bank (BMRB, bmrb.io [38]).

Archiv PDB je úložiště atomových souřadnic a dalších informací popisujících proteiny a další důležité biologické makromolekuly. K určení umístění každého atomu v molekule se používají metody, jako je rentgenová krystalografie, nukleární magnetická spektroskopie a kryoelektronová mikroskopie. Získané informace o umístění se uloží a jsou následně anotovány a veřejně uvolněny do archivu wwPDB.

Primární informace uložené v archivu PDB se tvoří ze souřadnicových souborů pro biologické molekuly. Tyto soubory obsahují seznam atomů v každém proteinu a jejich 3D umístění v prostoru. Zmíněné soubory jsou k dispozici v několika formátech, příkladem může být formát PDB, mmCIF a XML. Typický soubor ve formátu PDB obsahuje velkou část textu jako „záhlaví“, které shrnuje protein, informace o citaci a podrobnosti o struktuře molekuly, následuje sekvence a seznam atomů včetně jejich souřadnic [39].

2.2 Statistické výpočty

Statistické výpočty pomáhají určit, zda jsou si dvě množiny výsledků podobné nebo se od sebe liší.

Níže jsou popsány statistické výpočty, neboli metriky, které se běžně používají k porovnání 3D struktur a parciálních atomových nábojů.

2.2.1 Metriky porovnání 3D struktury

Pro porovnání 3D struktur molekul se běžně používá Root mean square deviation (RMSD), jež se běžně uplatňuje jako míra odlišnosti mezi datovými sadami $X = (x_1, \ldots, x_n)$ a $Y = (y_1, \ldots, y_n)$. Rovnice 2.1 znázorňuje podobu RMSD.

$$RMSD(X,Y) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \tag{2.1}$$

Jedná se tedy o měřítko přesnosti, jelikož srovnává prognostické chyby odlišných modelů pro jednotlivá data. Větší chyby mají velký dopad, z tohoto důvodu RMSD reaguje na odlehlé hodnoty [40].

2.2.2 Metriky porovnání parciálních atomových nábojů

Metriky na porovnání parciálních atomových nábojů nejsou detailně popsány, jelikož byly použity stejné jako v bakalářské práci a jejich popis je možno dohledat tam [41].

Mezi tyto metriky patří Pearsonův korelační koeficient (PCC), jenž udává korelaci vztahu dvou náhodně spojitých proměnných, Mean absolute error (MAE), který určuje míru chybovosti mezi datovými sadami a Maximální absolutní odchylka (ABSMAX), jenž určuje maximální odchylku mezi datovými sadami v absolutní hodnotě. Jako poslední metrikou bylo použito i již zmíněné RMSD.

2.3 Silová pole

V rámci molekulové mechaniky (viz kapitola 1.3) jsou konkrétní potenciálové funkce spolu se svými parametry označovány jako takzvaná silová pole.

Silová pole je možno rozdělit podle způsobu, jakým popisují vodíky. První typ je označován jako „all–atom“. Jedná se o atomy vodíku, které jsou explicitně vyjádřeny. Občas je zbytečné počítat se všemi atomy a z toho důvodu existuje i druhý typ, tzv. sjednocené „united–atom“ silové pole. Tento druh silového pole má slabě interagující nepolární vodíkové atomy zahrnuty do struktury těžšího atomu uhlíku a tvoří se tzv. pseudoatomy CH, CH2, CH3 [31, 42].

Níže jsou blíže popsány tři typy silových polí, které se nějakým způsobem používají v některých nástrojích na určení poloh vodíků zmíněných v kapitole 2.5.

2.3.1 AMBER

Silové pole AMBER [43] je široce používané a má dostatečně definovanou vazebný úhel, úhel vazby, parciální náboj a van der Waalsovy parametry pro všechny běžné aminokyseliny a nukleotidy. Nicméně má nedostatečné parametry pro mnoho modifikací, jenž se nacházejí v nukleových kyselinách a proteinech.

Systematické studie prokázaly vynikající schopnost standardního silového pole AMBER reprodukovat na vysoké úrovni kvantově mechanická data pro vodíkové vazby a stohovací (stacking) interakce v plynné fázi. Celkově studie naznačují, že silové pole AMBER je fyzikálně významné a zachovává si správnou rovnováhu mezi intramolekulárními a intermolekulárními silami.

Verze silového pole AMBER parm94 a parm99 byly parametrizovány, když simulace byly v časovém měřítku 1 s a výpočty kvantové mechaniky byly omezeny na malé modelové systémy a na menší teoretické úrovně. Docela překvapivě si obě verze stále vedou dobře v simulacích v rozsahu 10 ns, což je v současné době běžné trvání simulace [44].

V závislosti na komponentech v systému je potřeba určit silové pole, které se má použít. K tomu existuje doporučený seznam silových polí. Nicméně pro další složky, které nejsou v seznamu zahrnuty, jako jsou modifikované aminokyseliny nebo nukleotidy a jiné ionty, se určuje silové pole podle potřeby [45].

2.3.2 CHARMM

Název CHARMM [46] se používá pro sadu silových polí. Zároveň je to název pro počítačový softwarový balíček pro simulaci a analýzu molekulární dynamiky, jenž je popsán v části 2.5.13.

Parametrizace empirického silové pole CHARMM je vytvářena společně se stejnojmenný nástrojem pro určení poloh vodíků, který byl původně vyvinut v laboratoři profesora Martina Karpluse z Katedry chemie na Harvardské univerzity [47].

Silové pole CHARMM má široké možnosti použití, jelikož jako aditivní „all–atom“ silové pole zahrnuje parametry pro proteiny, nukleové kyseliny, lipidy a sacharidy [46]. Vyjádření vazeb, úhlů, dihedrálních úhlů a nevazebnosti jsou podobné těm, které se nacházejí v jiných silových polích, jako je například AMBER. Silové pole CHARMM také zahrnuje výrazy pro nesprávné počítání ohybu mimo rovinu [48].

Doporučené silové pole CHARMM pro proteiny, které obsahuje sjednocené atomy („united-atom“), je CHARMM19 [49]. Pro „all-atom“ je vhodné využití verze CHARMM22 [48] a pro jeho dihedrální potenciál je možné použít korigovanou variantu CHARMM22/CMAP, stejně jako novější verze CHARMM27 a CHARMM36 a různé modifikace [50]. Samozřejmě pro další sloučeniny jako je DNA, RNA nebo lipidy existují různé varianty tohoto silového pole. Dokonce je možné některá silová pole kombinovat, například pro simulaci vazby protein-DNA [51].

2.3.3 GROMOS

GROMOS [52] je univerzální balíček sloužící k počítačovým simulacím jako je např. molekulární dynamika pro studium biomolekulárních systémů. Zahrnuje také vlastní silové pole

vhodné pro proteiny, nukleotidy, cukry a další. Zároveň může být aplikován na chemický i fyzikální systém.

GROMOS je silové pole sjednocených atomů („united-atom“), neboli bez explicitních alifatických (nepolárních) vodíků [53].

Základní filozofií silového pole GROMOS je snaha dosáhnout rovnováhy mezi přesným popisem potenciální energie jako funkce konformace na jedné straně a relativně jednoduchou funkční formou na straně druhé. Pro jednoduchou funkční formu je žádoucí minimalizovat počet požadovaných parametrů. Dále je vhodné omezit výpočetní náklady na vyhodnocení potenciální energie. Přesné parametry k popisu takzvaných vazebných interakcí, jako je funkce délek vazeb, vazebných a torzních úhlů, lze snadno získat z krystalografických a spektroskopických údajů pro malé molekuly. Nelze spojit interakce, které zahrnují elektrostatické a van der Waalsovy interakce, mezi (v zásadě) všemi páry atomů v systému. Je to z toho důvodu, že tyto interakce jsou velmi obtížné na parametrizování. První hodnoty nevazebných interakčních parametrů pro GROMOS byly získány z krystalografických dat a atomové polarizovatelnosti. Následně byly hodnoty upraveny tak, aby byly vytvořeny experimentální vzdálenosti a interakční energie jednotlivých párů funkčních atomových skupin pro minimální energetickou konfiguraci [52].

2.4 MACH

Softwarem MACH [3] se vypočítaly parciální atomové náboje. Jedná se o vysoce optimalizovaný nástroj pro parametrizaci empirických metod, který byl vyvinut v rámci diplomové práce magistra Ondřeje Schindlera. Tento software funguje ve 4 režimech. Ty umožňují parametrizaci, výpočet a porovnání parciálních atomových nábojů a poslední režim umožňuje získat informace o sadě molekul.

MACH má v sobě implementováno několik metod a je možné, že jejich počet v budoucnu poroste. V této práci se využila metoda EQeq. Jedná se o metodu pro odhad parciálních atomových nábojů, která využívá všechny naměřené ionizační energie pro každý atom v periodické tabulce. EQeq staví na základě Charge equilibration (QEq) [54], jež je často používaná v publikacích [4]. Rozdíl mezi nimi je ten, že QEq má původní parametry (zjednodušeně řečeno) odpovídající elektronegativitě atomu a tvrdosti. Nicméně existuje více různých implementací [55]. EQeq používá pouze ionizační energii a elektronovou afinitu, tzn. de facto tabulkové hodnoty. Nicméně implementace v nástroji MACH tyto tabulkové hodnoty nahrazuje parametrizovanými [3].

Metoda EQeq je zároveň součástí i webového nástroje Atomic Charge Calculator II (ACC II), který je možné používat na rychlý výpočet parciálních atomových nábojů [4].

2.5 Programy na dopočítání vodíků

Přidání atomů vodíků k modelům nebo experimentálním strukturám proteinů, které obsahují pouze nevodíkové atomy, je běžným krokem v upřesnění krystalografické struktury, v teoretických studiích proteinů a v predikci struktury proteinů. Přesná předpověď poloh vodíků je důležitá, protože tvoří přibližně polovinu atomů v proteinech a významně se tak podílejí na energetice dané molekuly [6].

Existuje mnoho výpočetních nástrojů pro určování pozic vodíků. Níže jsou blíže popsány některé z nich.

2.5.1 Chimera

UCSF Chimera [56] je program pro interaktivní vizualizaci a analýzu molekulárních struktur a souvisejících dat, včetně map hustoty, trajektorií a zarovnání (sequence alignments) sekvencí. Nástroj je k dispozici zdarma pro nekomerční použití.

Chimera používá pro přidání vodíků do struktury příkaz `AddH`. Tento příkaz nejenom že přidá atomy vodíků k molekulám, ale zároveň i doplní atomy OXT tam, kde chybí na peptidových C-koncích. K určení typu atomu a konektivity používá název atomu a residua. V případě, že se jedná o nestandardní residua, použije k jejich určení souřadnice atomů. Příkaz `AddH` pak používá typy atomů k určení počtu přidaných vodíků a jejich pozic. Pozice již existujících atomů se nezmění, ale všechny volné páry a atomy neidentifikovatelných prvků se odstraní.

Chimera umožňuje přidávat vodíky dvěma metodami. První z nich je sterické použití na základě typů atomů a vyhýbání se střetům. Druhá pomalejší metoda zvažuje vodíkové vazby. Tato metoda je tedy založena na základě typů atomů, vyhýbání se střetům a tvorbě vodíkových vazeb. Vzhledem k tomu, že braní v potaz vodíkové vazby prodlužuje dobu výpočtu, je tato metoda stále ve vývoji.

Cílem `AddH` je generovat protonační stavy přiměřené při fyziologickém pH. Například se k fosfodiesterovým částem DNA a RNA nepřidávají vodíky. Ve výchozím nastavení se předpokládá, že postranní řetězce kyseliny asparagové a kyseliny glutamové jsou negativně nabité, zatímco postranní řetězce s argininem a lysinem jsou považovány za pozitivně nabité (i když lze dosáhnout jiných stavů). Při biologickém pH jsou dvě chemické skupiny považovány za nejednoznačné. Tou první jsou imidazoly, jako jsou histidinové postranní řetězce. Stavy protonačního histidinu mohou být specifikovány uživatelem nebo odhadnuty metodou. Druhou skupinou jsou koncové fosfáty, kde je kyslík protonován, pokud je jedna vazba P–O alespoň o 0,05 Å delší než ostatní kolem stejného atomu fosforu.

Residua na koncích připojených peptidových řetězců se kontrolují, aby se určilo, zda jsou skutečně na konci. Kontrolují se na základě jakýchkoli informací SEQRES ve vstupním souboru PDB (nebo ekvivalentu mmCIF) a přítomnosti nebo nepřítomnosti dalších řetězců se stejnými ID. Skutečné N-konce se považují za kladně nabité ($^{+}H_3N$-) a skutečné C-konce se považují za záporně nabité (–CO_2–). Pokud na C-koncovém karboxylátu chybí kyslík (OXT), je atom kyslíku přidán. Koncová residua, která nejsou skutečně na konci, jsou ukončena jako ostatní residua uvnitř řetězce s N(H)- a –C(=O). Poloha „amidového“ vodíku na N-konci v takových případech není plně určena polohami existujících atomů. Vodík je tedy umístěn tak, aby vytvořil úhel φ rovný úhlu následného residua.

Délky vazeb pro X-H, kde X = C/N/O/S, jsou převzaty z parametrů Amber parm99 (viz. tabulka 2.1).

Délky vazeb k ostatním X jsou přibližné a jsou získány sečtením poloměrů kovalentních vazeb prvků X a H [57].

Chimera je napsaná primárně v Pythonu a kromě jádra je v něm implementováno vše. Jádro je vytvořeno pomocí jazyka C++. Díky tomuto rozdělení je možné, aby jádro zpracovávalo časově kritické operace [56]. Chimera je tedy konzolová aplikace a zároveň

umožňuje práci ve svém vlastním grafickém rozhraní. Pro práci pomocí terminálu je možné používat skripty napsané v programovacím jazyce Python.

X	Typ atomu	Délka vazby X-H [Å]
sp^3 uhlík	C3	1,0900
sp^2 uhlík	C2,Car	1,0800
sp uhlík	C1	1,0560
dusík	N3+, N3, Npl, Ng+	1,0100
sp^3 kyslík	O3 (kromě vody)	0,9600
sp^3 kyslík	O3 (voda)	0,9572
síra	S3	1.3360

Tabulka 2.1: Paremetry AMBER parm99 pro délky vazeb X-H, kde C3 = sp^3-hybridizovaný uhlík ,C2 = sp^2-hybridizovaný uhlík, Car = aromatický uhlík, C1 = sp-hybridizovaný uhlík, N3+ = sp^3-hybridizovaný dusík s formálním kladným nábojem, N3 = sp^3-hybridizovaný dusík, formálně neutrální, Npl = sp^2-hybridizovaný dusík navázaný na další tři atomy, formálně neutrální (amid, anilin), Ng+ = rezonanční ekvivalent dusíku sdílející formální pozitivní náboj (guanidinium, amidinium), O3 = sp^3-hybridizovaný kyslík a S3 = sp^3-hybridizovaná síra [57].

2.5.2 PyMOL

PyMOL [58] je multiplatformní molekulární grafický nástroj a je široce používán pro trojrozměrnou vizualizaci proteinů, nukleových kyselin, malých molekul, hustoty elektronů, povrchů a trajektorií. Je také schopen upravovat molekuly, realizovat ray tracing a vytvářet snímky nebo animace. V současnosti je PyMOL jedním z nejpoužívanějších makromolekulárních vizualizačních nástrojů. Vzhledem k tomu, že PyMOL je napsán v Pythonu, lze jej snadno rozšířit různými pluginy [59].

Jedná se o open-source nástroj [60] vytvořený Warrenem Lyfordem DeLanem a původně komercializovaný společností DeLano Scientific LLC. Od roku 2010 převzala PyMOL společnost Schrödinger [59]. Ta jej nabízí pro soukromé účely zdarma, komerční použití je možné po zakoupení jednoleté nebo tříleté licence [60].

Pro dopočítání vodíků do struktury se používá příkaz `h_add`, který používá primitivní algoritmus. PyMOL vyplňuje chybějící valence, ale neprovádí žádnou optimalizaci hydroxylových rotamerů. Mnoho krystalových struktur má špatné nebo náhodné orientace asparaginu, glutaminu nebo histidinu. Pro získání správných amidových rotamerů, imidazolových tautomerů a přiřazení protonačních stavů se používá netriviální výpočetní

chemický návrh zahrnující výpočty elektrostatického potenciálu a kombinatorické vyhledávání.

Obecně řečeno, PyMOL neobsahuje dostatek informací k pečlivému stanovení vodíkových vazeb, protože typické soubory PDB nejsou jednoznačné, pokud jde o vazby, valenční vazby a tautomery. V současné době jsou všechny tyto věci heuristicky odhadovány. Nástroj pro detekci vodíkových vazeb byl původně vyvinut za účelem napodobování algoritmu přiřazení sekundární struktury, který je založen na poměrně velkorysém pojetí vodíkových vazeb. Kritéria, která se používají při implementaci, mohou být založena na souřadnicích těžkých atomů.

Závěrem tedy je, že PyMOL pouze nabízí domnělé polární kontakty a ponechává na uživateli, aby určil, zda jsou přítomné interakce ve skutečnosti vodíkové vazby, solné můstky, polární interakce nebo pouze artefakty, jenž vznikají nesprávným přiřazením. Příkladem může být vazba dvou karbonylových vodíků, se kterými se zachází jako s hydroxyly.

PyMOL je možné používat v grafickém rozhraní nebo v terminálu jako konzolovou aplikaci. Pro práci v terminálu je možné využívat Python nebo PyMOL skripty se standardní příponou .pml. To jsou soubory prostého textu obsahující příkazy PyMOL, jeden na každý řádek, které se zobrazují tak, jak by byly zadány na příkazovém řádku PyMOL v grafickém rozhraní [61].

2.5.3 HAAD

HAAD (Hydrogen Atom ADdition) [62] je algoritmus, který predikuje polohy atomů vodíku na základě umístění těžkých atomů. Algoritmus je postaven na základních pravidlech hybridizace orbitalů s následnou optimalizací sterického odpuzování a elektrostatických interakcí.

Tento nástroj pro rychlou predikci atomů vodíku kombinuje omezení místní geometrie a hledání konformací. Účelem je snížit sterický odpor a zlepšit sítě vodíkových vazeb ve struktuře proteinu. Z toho důvodu HAAD atomy vodíků prvně umísťuje na základě lokální geometrie, která je určena hybridním orbitalem těžkého atomu, ke kterému je atom vodíku připojen. Obecně jsou pro fixaci prostorové polohy vodíku potřeba tři omezení. První dvě z nich jsou délka vazby a úhel vazby, které jsou konstantní a jsou převzaty ze silového pole CHARMM22. Třetí omezení je určeno na základě tříd daného atomu vodíku. Vodíky jsou klasifikovány na základě typu hybridního orbitalu a počtu atomů vodíků připojených k centrálnímu těžkému atomu.

Algoritmus byl jeho autory testován pomocí tří nezávislých datových sad. První sadu tvořily rentgenové struktury s vysokým rozlišením. Druhou vytvářely struktury určené neutronovou difrakcí a třetí tvoří vzdálenosti proton-proton nuklearního Overhauserova efektu (NOE).

Ve srovnání s široce používanými programy CHARMM a Reduce má HAAD výrazně vyšší přesnost, přičemž průměrná hodnota RMSD predikovaných atomů vodíku pro rentgenové a neutronové difrakční struktury poklesla o 26 % a 11 %. Atomy vodíku umístěné HAAD nástrojem mají navíc více shod s omezeními NOE a méně střetů s těžkými atomy. Průměrné vytížení procesoru u HAAD je 18krát a 8krát nižší než u CHARMM a REDUCE.

Významná výhoda HAAD v přesnosti i rychlosti přidávání vodíku činí tento nástroj vhodný pro podrobné studium struktury a funkce proteinu.

HAAD je samostatný program napsaný ve FORTRAN90 [5]. Kromě toho je možné tento nástroj použít jako online server. Online server, spustitelný soubor a dokonce i zdrojový kód programu jsou volně dostupné [62].

2.5.4 Reduce

Reduce [63] je program pro přidávání vodíků do souboru molekulární struktury formátu PDB. Vodíky se přidávají ve standardizované geometrii s optimalizací u orientací OH, SH, NH_3^+, methioninových methylů, asparaginových a glutaminových amidů postranních řetězců a histidinových prstenců. Tento nástroj umožňuje pracovat s proteiny i nukleovými kyselinami. Skupiny HET lze také zpracovávat, pokud je zajištěno propojení atomů, například přes CONECT záznamy.

Molekuly vody jsou zpracovány za předpokladu, že se mohou vždy orientovat tak, aby představovaly vše, co je pro každou interakci potřeba. Pokud jsou molekuly vody extrémně blízko, omezí se jejich skóre pro vodíkové vazby na rozumnou hodnotu [64].

Reduce umisťuje vodíky na základě očekávaných délek a úhlů v rámci geometrie atomu. Tam, kde je to vhodné, vypočítává kontaktní sítě atomů a pokouší se umístit vodíky tak, aby usnadnily vodíkové vazby a zabránily stérickým střetům [65].

Vodíky jsou tedy umístěny s ohledem na kovalentně vázané sousedy a jsou identifikovány jménem. Nestandardní názvy atomů jsou primární příčinou chybějících nebo nesprávně umístěných vodíků. Pokud se Reduce pokusí zpracovat soubor, který obsahuje vodíky s nestandardními názvy, nemusí být existující vodíky rozpoznány a mohou interferovat s generováním nových vodíků [45].

Používané jaderné vzdálenosti X-H jsou od mateřského atomu k jádru vodíku a vzdálenosti „e-cloudu“ X-H jsou od mateřského atomu ke středu elektronového mraku vodíku [66]. Jejich velikosti jsou znázorněny v tabulce 2.2.

Reduce rozpoznává disulfidy, kovalentní modifikace a spojení páteře ribosfosfátové nukleové kyseliny a přeskočí veškeré vodíky eliminované vazbou. Pokud aminokyselinový dusík hlavního řetězce není připojen k předchozímu residuu nebo jiné skupině, tak s ním zachází jako s N-koncem a vytvoří NH_3^+ pouze v případě, že počet residuí je menší nebo roven nastavitelnému limitu (1 jako výchozí). V opačném případě považuje residuum za sledovaný začátek skutečně připojeného fragmentu a neprotonizuje dusík. Reduce dále neprotonizuje karboxyláty (včetně C-konce), protože specificky nezohledňuje pH, namísto toho modeluje neutrální prostředí [45].

Reduce je samostatná komponenta nástroje MolProbity [67] a je využíván nástrojem Phenix [68] pro dopočítání vodíků. Samostatně je možné tento nástroj použít jako konzolovou aplikaci.

2.5.5 OpenBabel

OpenBabel [69] je sada chemických nástrojů navržená tak, aby pracovala s různorodými chemickými daty. OpenBabel je volně dostupný nástroj umožňující komukoli vyhledávat, převádět, analyzovat nebo ukládat data z molekulárního modelování, chemie, materiálů v pevné fázi, biochemie nebo souvisejících oblastí [70].

x-H atomy	e-cloud [Å]	jádro [Å]	příklad
tetra CH_1	0,97	1,09	Val Hb
tetra CH_2	0,97	1,09	Glu $Hb_{2,3}$
tetra CH_3	0,97	1,09	Ala $Hb_{1,2,3}$
planarní CH	0,93	1,08	Phe Hd_1
tetra NH_3	0,89	1,04	Lys $Hz_{1,2,3}$
planarní NH	0,86	1,02	peptid H
tetra OH	0,84	0,98	Ser Hg
tetra SH	1,20	1,30	Cys Hg

Tabulka 2.2: Vzdálenosti X-H používané v Reduce [66].

OpenBabel má dvě hlavní části. První z nich tvoří programy připravené k použití pro převádění, vyhledávání, úpravy a analýzu chemických souborů. Druhou část tvoří kompletní programátorská sada nástrojů, která umožňuje snadný vývoj softwaru pro chemické použití [69].

Open source verze Babel 2.3 dokáže převést více než 110 souborových formátů. Pro konvertování tak široké škály chemických a molekulárních dat je potřeba knihovnu, která implementuje velké množství algoritmů chemoinformatiky. Tyto algoritmy obsahují funkce od přiřazení parciálního náboje a detekce aromatičnosti až po zjištění pořadí vazeb a kanonizaci [71]. OpenBabel tedy nabízí metody pro stanovení konektivity vazeb, určení pořadí vazeb, stanovení aromatičnosti a typizace atomů [70].

V tomto nástroji je propojování vazeb určováno algoritmem pro rozpoznání atomů s malou tolerancí (0,45 Å), která umožňuje delší vazby než ty, které jsou typické. Atomy jsou tedy blíže než součet jejich kovalentních poloměrů. V rámci zpracování poruch v krystalografických datech (např. soubory PDB nebo CIF) nejsou atomy, jenž jsou blíže než 0,63 Å, vázány. K zajištění zachování standardní valenční vazby je proveden další filtrační průchod. Každý prvek má maximální počet vazeb. Pokud je maximum překročeno, pak jsou nejdelší vazby k atomu postupně odstraněny, dokud není splněno valenční pravidlo. Pořadí vazeb se určuje na základě úhlů a geometrií vazeb.

K určování typu atomů se používá tzv. lazy evaluation. Dále se využívá porovnávání atomů se vzory SMARTS za účelem stanovení hybridizace, implicitní valence a typů externích atomů. Typy atomu jsou důležité pro dopočítání vodíku, což vyžaduje určení implicitní a explicitní valence. Typy atomů jsou důležité i pro export do souborového formátu, který je vyžaduje, nebo jako požadavek uživatele [71].

OpenBabel tvoří knihovnu v C++ a lze ji snadno používat pomocí C++. Navíc k OpenBabelu lze přistupovat z Pythonu, Perlu, Ruby, C# a Javy. Pro Python existuje

modul nazvaný Pybel, který usnadňuje přístup k vazebným funkcím [70]. Kromě knihoven pro různé programovací jazky je možné OpenBabel spustit i jako konzolovou aplikaci.

2.5.6 GROMACS

GROMACS [53] je univerzální balíček pro provádění molekulární dynamiky [72, 73].

Je primárně určen pro biochemické molekuly, jako jsou proteiny, lipidy a nukleové kyseliny, které mají hodně komplikovaných vazebných interakcí. Na druhou stranu GROMACS je extrémně rychlý při výpočtu nevazebných interakcí, jež obvykle dominují simulacím, a tak je možné tento nástroj používat také pro výzkum biologických systémů.

GROMACS určuje souřadnice pomocí ztrátové komprese, což poskytuje velmi kompaktní způsob ukládání dat o trajektorii. Míra přesnosti se může zvolit podle potřeb uživatele.

Balíček umožňuje plně automatizovaně vytvářet topologie pro proteiny a dokonce i pro multimerní struktury. K dispozici jsou stavební bloky pro 20 standardních aminokyselinových residuí i pro některá modifikovaná, 4 nukleotidová a 4 deoxinukleotidová residua, několik cukrů a lipidů a ještě některé speciální skupiny jako hemy a několik malých molekul [53].

GROMACS podporuje tato silová pole: GROMOS-87(96), AMBER94(96/99/99SB/99SB-ILDN/03/GS) a CHARMM [74]. Podrobnější přehled silových polí, která jsou podporována, lze nalézt v tabulce A.2 v příloze.

GROMACS je metoda založená na geometrii. Příkaz `pdb2gmx`, který se používá pro přidání vodíků, používá různé vazebné délky a úhly podle zvoleného silového pole. Atomy vodíku vody se určují umístěním vazeb O-H střídavě podél tří souřadných os v obou směrech. Energetická funkce zahrnuje podmínky van der Waalsovy, Coulombické a vazebné energie. GROMACS optimalizuje své protonační stavy tím, že se snaží saturovat sousední vodíkové vazby [6].

GROMACS ukládá vodíkovou databázi v souborech `hdb`. Tento soubor obsahuje informace o použitém příkazu `pdb2gmx`, například jak připojit atomy vodíku ke stávajícím atomům. Pokud je ke stejnému atomu připojen více než jeden atom vodíku, bude na konec názvu atomu vodíku přidáno číslo. Například přidáním dvou atomů vodíku k ND2 v asparaginu, budou atomy vodíku pojmenovány HD21 a HD22.

Pro přidávání atomů vodíků se používají následující metody:

1. Jeden planární vodík, např. kruhy nebo peptidová vazba:

   Generuje se jeden atom vodíku (n), ležící v rovině atomů (i, j, k) v rovině půlící úhel (j-i-k) ve vzdálenosti 0,1 nm od atomu i, takže úhly (n-i-j) a (n-i-k) jsou >90°.

2. Jeden jediný vodík, např. hydroxyl:

   Jeden atom vodíku (n) je generován ve vzdálenosti 0,1 nm od atomu i, takže úhel (n-i-j) = 109,5° a dihedrální úhel (n-i-j-k) = trans.

3. Dva rovinné vodíky, např. ethylen -C=CH_2, nebo amid -C(=O)NH_2:

   Dva vodíky (n_1, n_2) jsou generovány ve vzdálenosti 0,1 nm od atomu i, takže úhel (n_1-i-j) = (n_2-i-j) = 120 ° a dihedrální úhel (n_1-i-j-k) = cis a (n_2-i-j-k) = trans, takže jména odpovídají standardům IUPAC.

4. Dva nebo tři čtyřboké vodíky, např. -CH_3:

   Tři (n_1, n_2, n_3) nebo dva (n_1, n_2) vodíky jsou generovány ve vzdálenosti 0,1 nm od atomu i, takže úhel (n_1-i-j) = (n_2-i-j) = (n_3-i-j) = 109,47°, dihedrální úhel (n_1-i-j-k) = trans, (n_2-i-j-k) = trans + 120° a (n_3-i-j-k) = trans + 240°.

5. Jeden čtyřboký vodík, např. $C_3^*CH^*$:

   Jeden atom vodíku (n) je generován ve vzdálenosti 0,1 nm od atomu i v čtyřboké konformaci tak, že úhel (n-i-j) = (n-i-k) = (n-i-l) = 109,47°.

6. Dva čtyřboké vodíky, např. C-CH_2^*-C^*:

   Dva atomy vodíku (n_1, n_2) jsou generovány ve vzdálenosti 0,1 nm od atomu i v čtyřboké konformaci na rovině rozdělující úhel j-i-k s úhlem (n_1-i-n_2) = (n_1-i-j) = (n_1-i-k) = 109,47°.

7. Dva vodíky u vody:

   Dva atomy vodíku jsou generovány kolem atomu i podle geometrie vody SPC 80. Osa symetrie se střídá mezi třemi souřadnými osami v obou směrech.

8. Tři vodíky u vody:

   Dva atomy vodíku jsou generovány kolem atomu i podle geometrie vody SPC 80. Osa symetrie se střídá mezi třemi souřadnými osami v obou směrech. Kromě toho se na pozici kyslíku generuje další částice, přičemž první písmeno názvu je nahrazeno písmenem „M“. Tato metoda se používá u čtyřatomových vodních modelů, jako je TIP4P 128.

9. Čtyři vodíky u vody:

   Stejné jako výše, až na to, že na pozici kyslíku se vytvářejí další dvě částice se jmény „LP1“ a „LP2.“ Tohle je využito u modelů s pěti atomy vody, jako je TIP5P 130.

Parametry pro konstrukci virtuální polohy vodíkových atomů jsou odvozeny z parametrů silového pole přímo argumentem grompp při zpracování topologie souboru. Konstrukce pro aromatická residua jsou založeny na délkách a úhlech vazeb pro geometrii. Nicméně tyto parametry jsou pevně zakódovány v příkazu `pdb2gmx` a to kvůli složitosti konstrukce potřebné pro celou aromatickou skupinu [74].

Program GROMACS ještě provádí analýzu vodíkových vazeb na základě geometrických kritérií a analýzy životnosti vodíkových vazeb [75].

Po dokončení práce program vypisuje různé citáty nebo fráze od nejrůznějších osobností, filmů seriálů a dalších. Příklad hlášek je zobrazen na obrázku 2.1.

GROMACS je volně přístupný software a na oficiálních stránkách je ještě specificky uvedeno, že slovo GROMACS neznamená žádnou zkratku, nýbrž je to vlastní jméno [53]. Tento nástroj je možné spouštět jako konzolovou aplikaci.

```
Writing coordinate file...
                --------- PLEASE NOTE ------------
You have successfully generated a topology from: 1kdf/1kdf_removed.pdb.
The Amber03 force field and the tip3p water model are used.
                --------- ETON ESAELP ------------

GROMACS reminds you: "All models are wrong, but some are useful." (George Box)
```

```
Writing coordinate file...
                --------- PLEASE NOTE ------------
You have successfully generated a topology from: 3gbl/3gbl_removed.pdb.
The Amber03 force field and the tip3p water model are used.
                --------- ETON ESAELP ------------

GROMACS reminds you: "O My God, They Killed Kenny !" (South Park)
```

Obrázek 2.1: Ukázka dvou hlášek, které se vypsaly po ukončení práce s programem Gromacs v terminálu. Horní obrázek ukazuje citát od George Boxe a spodní nejznámější hlášku ze seriálu South Park.

2.5.7 CNS

Crystallography & NMR System (CNS) [76] je výsledkem mezinárodního společného úsilí několika výzkumných skupin. Program byl navržen tak, aby poskytoval flexibilní víceúrovňový hierarchický přístup k nejčastěji používaným algoritmům při určování makromolekulární struktury. Mezi nejdůležitější přístupy patří vyhledávání těžkých atomů, experimentální fázování, modifikace hustoty, krystalografické zdokonalení s cílem maximální pravděpodobnosti a výpočet struktury NMR pomocí dat NOE, J-vazby, chemického posunu a dipolárních vazebných dat.

Generování všech poloh atomů vodíků lze v CNS pomocí souborů `generate_easy.inp` nebo `generate.inp`. Tyto soubory přestavují pozice vodíku, i když pozice vodíku byly získány z jiného programu. Činí se tak, protože stereochemické přiřazování názvu vodíku může být nekompatibilní s konvencí IUPAC používanou v CNS pro všechny vodíkové topologie a soubory parametrů [77].

V CNS jsou vodíky zpočátku umístěny náhodně a jsou optimalizovány pomocí čtyř stupňů simulace molekulární dynamiky (až 0,25 ps při 300 K) a následně střídány s pěti stupni minimalizace energie Powell s až 500 kroky na iteraci. Nástroj používá upravenou verzi silových polí PARALLHDG a PARHCSDX11. Toto silové pole zahrnuje parametry pro vazby a úhly Engh a Huber. Soubor parametrů silového pole používaný CNS zahrnuje van der Waalsovy a vazebné podmínky, ale nejsou zahrnuty žádné Coulombické interakce, které jsou konzistentní se standardními postupy zdokonalování struktury v rentgenové krystalografii. CNS je však schopen vypočítat Coulombické interakce, pokud je dodán vhodný soubor parametrů silového pole [76].

2.5.8 Amber

Amber [45] je sada biomolekulárních simulačních programů. Tento nástroj se začal tvořit na konci 70. let a nyní je stále vyvíjen aktivní vývojovou komunitou.

Název Amber označuje dvě věci. Nejprve se jedná o soubor molekulárně mechanických silových polí pro simulaci biomolekul. Tato silová pole jsou veřejně dostupná a používají se v různých simulačních programech. Blíže se jim věnuje sekce 2.3.1. Za druhé se jedná o balíček programů molekulární simulace, který zahrnuje zdrojový kód a ukázky jeho použití [78].

Amber je rozdělený na dvě hlavní sbírky balíčků, Amber a AmberTools. AmberTools se skládá z nezávislé sady softwarových nástrojů, které byly většinou součástí Amber. Obecně se nástroje používají k přípravě souborů, například souřadnic a topologie pomocí silových polí Amber. Také je možné tento nástroj využít k analýze výstupních dat a trajektorií. Balíček AmberTools také zahrnuje programy, které provádějí mnoho simulačních úkonů. Klíčovou vlastností AmberTools je, že se jedná o zcela otevřený zdroj poskytující přístup k mnoha klíčovým funkcím Amber bez nutnosti licence [79]. Druhý balíček, označený jako Amber, je soustředěn kolem simulačního programu pmemd, který je licencován přísněji [45].

AmberTools tedy obsahuje řadu přípravných nástrojů, které jsou navrženy pro volnou spolupráci. Jedním z nich je nástroj LEaP.

LEaP je zkratka vytvořená z názvů starších softwarových modulů Amber. Nahrazuje tedy moduly link, edit a parm. LEaP lze použít k přípravě vstupních dat pro programy molekulární mechaniky Amber [45]. Jinými slovy LEaP je modul ze sady programů Amber, který lze použít ke generování souborů silového pole kompatibilních s balíčky Amber Molekulová dynamika a Nucleic Acid Builder (NAB). Od verze Amber v12 jsou v současné době k dispozici dvě varianty tohoto nástroje. První je program s názvem tleap, jenž je možné používat skrz příkazový řádek, druhá verze nese název xleap a má vlastní editor s příkazovým řádkem [79].

Programy teLeap a xaLeap se obvykle spouští prostřednictvím skriptů tleap a xleap. Tyto dva programy sdílejí společný jazyk pro příkazy, ale program xleap byl vylepšen přidáním grafického uživatelského rozhraní X-windows.

Při použití grafického rozhraní pomocí příkazu xleap je možné přidat vodíky do zvolené molekuly za využití rozbalovací nabídky. V nabídce je příkaz „Add H & Build“, který přidá vodíky do atomů, které nemají plnou valenci. Zároveň se vytvoří souřadnice pro vodíky a další atomy označené jako „unbuilt“. Počet vodíků přidaných do každého atomu je určen hybridizací a typem prvku každého atomu.

Většina souborů formtáu PDB vyžaduje nějakou opravu před použitím v Amber. Té lze dosáhnout programem pdb4amber a za použití Reduce jako volitelného argumentu. Za pomoci této kombinace se převede původní soubor PDB na jeden, který bude pravděpodobně vhodnější pro vstup do programu LEaP.

Pdb4amber analyzuje soubory PDB a čistí je pro další použití, zejména pro programy LEaP Amber. Obecně pomáhá při přípravě souborů ve formátu PDB pocházejících z jiných míst (například rcsb.org), aby byly kompatibilní s LEaP. Argument Reduce dopočítá chybějící vodíky do PDB souboru a je použit stejný nástroj, který je blíže popsán v sekci 2.5.4 [45].

2.5.9 CORINA

CORINA Classic [80] je rychlý a výkonný generátor 3D struktury pro malé a střední velikosti molekul, typicky léky. Jeho robustnost, rychlost, výkon a dodatečné očišťující metody pro struktury a standardizační metody vytváří z CORINA Classic vhodnou aplikaci pro převod velkých chemických datových souborů a databází.

CORINA Classic je používána většinou farmaceutickými a chemickými společnostmi a akademickými výzkumnými laboratořemi k převodu jejich 2D struktur do 3D [80].

CORINA interně přidává chybějící nebo implicitně dané atomy vodíku před generováním 3D souřadnic, aby získala struktury s vyšší kvalitou a lépe řešenými blízkými kontakty. Ve výchozím nastavení jsou přidané atomy vodíku odstraněny z výstupního souboru. Toto nastavení umožňuje, aby CORINA zapsala přidané atomy vodíku [81].

Délky a úhly vazeb jsou nastaveny na standardní hodnoty převzaté z tabulky. Délky vazeb závisí na typech atomů, stavech atomové hybridizace a pořadí vazeb uvažovaného atomového páru. Pro typy vazeb, které nejsou uvedeny v tabulce, se počítají vhodné hodnoty z kovalentních poloměrů atomů a elektronegativit. Délky vazeb v konjugovaných systémech se uvolňují pomocí Huckelova MO schématu. Vazebný úhel závisí pouze na typu atomu a stavu hybridizace centrálního atomu. Atomy až se šesti sousedy lze zpracovat pomocí jednoho typu základní geometrie podle modelu VSEPR. Tento model obsahuje typy terminální, lineární, planární, čtyřboký, trigonální bipyramidový nebo oktaedrický. Tabulky pro délky a úhly vazeb atomů jsou parametrizovány pro celou periodickou tabulku [82].

Nástroj CORINA je možné spouštět jako konzolovou aplikaci.

2.5.10 MCEE

MCCE (Multi-Conformation Continuum Electrostatics) [83] je biofyzikální simulační program vyvinutý v Gunner Lab na City College v New Yorku.

MCCE kombinuje elektrostatiku kontinua a molekulární mechaniku. V tomto programu jsou pohyby postranního řetězce proteinu simulovány, zatímco dielektrický účinek rozpouštědla a velikost proteinového materiálu je modelován elektrostatikou kontinua. MCCE může vypočítat residua pKa, kofaktor Em a protein PI v systémech protein-rozpouštědlo. Dále zvládne napočítat reakce proteinových struktur na změny náboje, zbytkové ionizační změny v reakci na změny struktur proteinů a navíc ještě umístění a stechiometrii přenosů vodíků spojených s tokem elektronů [84].

Síť vodíkových vazeb, jež je získaná z MCCE, je tvořena vodíkovými vazbami existujícími v proteinu v rámci Boltzmannova rozložení. Síť pomáhá odhalit translokační dráhy protonů a vodní cesty zahrnující protein [83].

Síť vodíkových vazeb v MCCE je tedy získávána z mikrostávů v Boltzmannově rozložení ve vzorkování Monte Carlo na základě definované vzdálenosti a úhlu D-H...D, kde D je donorový atom s volným párem. Prvně se uloží mikrostav s Boltzmannovým rozložením ve vzorkování Monte Carlo. Poté se vypočítají všechny možné vodíkové vazby mezi komformery resuidí. Tímto jsou získány matice vodíkových vazeb. Posledním krokem se spojí mikrostavy a matice vodíkových vazeb a vypočítají se vodíkové vazby vybrané v mikrostavech ve vzorkování Monte Carlo [84].

2.5.11 WHAT IF

WHAT IF [85] je univerzální balíček pro modelování molekul, který se specializuje na práci s proteiny a molekulami v jejich prostředí, jako je voda, ligandy, nukleové kyseliny a další.

Nástroj WHAT IF byl vyvíjen dlouhou dobu. Začal v roce 1987 v Groningenu a pokračoval v letech 1989-2000 na EMBL v Heidelbergu, v letech 2000-2006 na CMBI Radboud University Nijmegen a v roce 2006 se s CMBI přestěhoval do Medical Center Radijd University Nijmege [86].

Souřadnice atomu vodíku se počítají pomocí jednoduché geometrické konstrukce, kde atomy vodíku jsou umístěny v ideálních trojúhelníkových nebo tetraedrických polohách. Ostatní atomy vodíku, jako je alkoholový OH, lysinový NH_3 a N-terminální NH_3, se mohou volně pohybovat po kruhu. Atomy vodíku spojené s molekulami vody tvoří úhel H-O-H přibližně 110°, ale orientace molekuly je volná. Ve všech případech jsou atomy vodíku umístěny ve vzdálenosti 1,0 Å od donorového atomu [85].

Obsáhlé možnosti nástroje WHAT IF jsou k dispozici jako webová služba. Dále je tento nástroj kompatibilní s několika dalšími softwarovými balíčky bioinformatiky například s nástrojem YASARA [86].

2.5.12 YASARA

YASARA [87] je program molekulární grafiky, modelování a simulace pro Windows, Linux, MacOS a Android vyvinutý v roce 1993.

YASARA využívá technologii PVL (Portable Vector Language), což je vývojový framework. PVL umožňuje vizualizovat i ty největší proteiny a umožňuje skutečné interaktivní simulace v reálném čase s vysoce přesnými silovými poli na standardních počítačích a to s využitím GPU, pokud jsou k dispozici. Místo statických obrázků je možné molekuly posouvat a táhnout a tím pracovat s dynamickými modely.

Základní nástroj YASARA View je možno získat zdarma a umožňuje programovat vlastní makra a Python pluginy. Na druhou stranu nástroje s vyšší funkcionalitou, jako jsou YASARA Model, YASARA Dynamics a YASARA Structure, vyžadují licenční poplatek, který umožňuje garantovat nový vývoj, aktualizace a podporu i v budoucnu.

Chemická vazba v YASARA má určité pořadí, například jednoduché, dvojité a další. Nicméně na rozdíl od jiných modelových programů jsou podporovány i řádové podíly vazeb někde mezi „jednoduchým“ a „dvojitým“. To pomáhá zachovat symetrii a poskytnout lepší obrázek o základní chemii ve struktuře.

Pořadí vazeb a chybějící atomy vodíku jsou přiřazeny automaticky pro jakoukoli organickou molekulu a to způsobem závislým na pH. V počátečním průchodu se provádí typizace, která se vyvíjela ve spolupráci s týmem OpenBabel. Následně YASARA používá knihovnu SMILES řetězců k přiřazení hodnot pKa funkčním skupinám, které pak pomáhají doladit pořadí vazeb a protonační vzorce podle zvoleného pH.

Volně dostupný YASARA View je omezen na pH 7. Na druhou stranu placený YASARA Model umožňuje zvolit jakékoli pH od 0 do 14 nebo pracovat ve vakuu. V YASARA Dynamics můžete také rozšířit chemické znalosti YASARA přidáním vlastních řetězců SMILES do knihovny [87].

2.5.13 CHARMM

CHARMM [88] (Chemistry at HARvard Macromolecular Mechanics) je program pro molekulovou simulaci s širokým uplatněním v systémech s mnoha částicemi s komplexní sadou energetických funkcí, řadou vylepšených metod odběru vzorků a podporou vícerozměrných technik a řady implicitních modelů rozpouštědel [89].

CHARMM se primárně zaměřuje na biologické systémy včetně peptidů, proteinů, protetických skupin, ligandů s malými molekulami, nukleových kyselin, lipidů a sacharidů, které se vyskytují v prostředí roztoků, krystalů a membrán. Program také nachází široké uplatnění pro anorganické materiály s aplikacemi v materiálovém designu [90].

Vodíky jsou klasifikovány podle jejich prostředí. Pro jejich definování se používají následující třídy:

a) Vodík vázaný na donor s alespoň dvěma předchůdci těžkých atomů. Například (C, CA) -N-H.

b) Vodík vázaný na donor s jedním předchůdcem těžkého atomu a žádným jiným vodíkem. Například -OH-HH tyrosinu.

c) Vodík vázaný na donor s jedním předchůdcem těžkého atomu a jedním dalším vodíkem. Například skupina -NH_2- (HH_{21}, HH_{22}) argininu.

d) Vodík vázaný na donor s jedním předchůdcem těžkého atomu a dvěma dalšími vodíky. Například -NZ- (HZ_1, HZ_2, HZ_3) skupina lysinu.

Nejprve jsou všechny vodíky třídy a) umístěny pomocí rovnovážných vazebných délek, úhlů a dihedrálních úhlů. Takto je to určeno, pokud existuje více než jeden předchůdce těžkého atomu. Následně se provede průměrování všemi možnými způsoby, jak umístit vodík.

V dalším kroku jsou konstruovány vodíky všech ostatních tříd. Všechny třídy mají společné to, že existuje určitý stupeň volnosti pro umístění vodíků, například rotace kolem vazby lysinu CE-NZ. Pro nalezení optimální polohy je dihedrální úhel s osou symetrie předchůdce atomu upraven v malých krocích. Tyto kroky jsou v určitém rozsahu, který je určen symetrií skupiny atomů. Pro každý úhel stoupání jsou vodíky atomů umístěny podle jejich rovnovážné geometrie a je vyhodnocena relativní energie odpovídající konfiguraci. Energie se určuje pomocí potenciálu vodíkové vazby, van der Waalsova členu, elektrostatického členu a dihedrálního členu odvozeného z úplného vyjádření energie CHARMM. Dihedrální úhel s nejnižší energií je odebrán a vodíky atomové skupiny jsou umístěny s optimálním dihedrálním úhlem. Tento postup se provádí v pořadí daném sekvencí residuí proteinu. Konstrukce vodíků nemá žádný vliv na stávající hodnocení energie.

Po konstrukci všech explicitních proteinových vodíků jsou konstruovány vodní vodíky. Nejprve se určí sekvence molekul vody nezávisle na jakékoli vstupní sekvenci. Molekuly vody jsou uspořádány s ohledem na minimální vzdálenost vodního kyslíku k jakémukoli atomu proteinu. Nejprve jsou vytvořeny vodní vodíky poblíž proteinu.

V současné době existují tři třídy molekul vody řešené v CHARMM:

a) voda schopná tvořit dvě různé vodíkové vazby na akceptorové atomy

b) voda schopná tvořit pouze jednu vodíkovou vazbu k akceptorovému atomu

c) voda netvoří vůbec žádné vodíkové vazby na akceptorových atomech.

V případě a) jsou vodíky umístěny provedením rotace molekuly vody v rovině definované dvěma nejlepšími vodíkovými vazbami a provedením konfigurace minimální energie.

V případě b) je jeden proton umístěn na (lineární) vodíkové vazbě a voda se otáčí kolem této osy a druhý vodík je umístěn pomocí rovnovážné geometrie. Znovu se vezme konfigurace minimální energie. Vyhodnocená relativní energie je součtem van der Waalsových, elektrostatických a vodíkových vazeb. Nakonec se vodíky v rámci vody z případu b) umístí standardním způsobem po umístění všech ostatních vodíků. Standardní způsob umístění je H_1 na osu x, H_2 v rovině x, y. Molekuly vody ST2 jsou považovány za normální vody pro konstrukci protonu. Poloha volných párů je odvozena od pozic protonů [88].

2.5.14 H++

H++ [91] je automatizovaný systém, který počítá hodnoty pK ionizovatelných skupin v makromolekulách a přidává chybějící atomy vodíku podle specifikovaného pH prostředí. Vzhledem k tomu, že na vstupu je struktura v souboru PDB, umožňuje H++ dostat na výstupu strukturu v několika běžných formátech, například PDB, PQR, AMBER .inpcrd/-prmtop, a poskytuje sadu nástrojů pro analýzu elektrostatických molekulárních vlastností [92].

Kromě toho jsou generovány odpovídající soubory souřadnic a topologie ve formátu, který je podporován balíkem molekulárního modelování AMBER. Tento balíček umožňuje provádět ještě několik dalších užitečných výpočtů. Systém je určen pro širokou komunitu biochemiků, molekulárních modelářů, strukturních biologů a návrhářů léků. Systém může být také použit jako vzdělávací nástroj v kurzech biochemie.

H++ přidá chybějící těžké atomy a vodíky, za předpokladu standardních protonačních stavů titrovatelných skupin. Dále přidá parciální náboje atomů a zároveň přiřadí poloměry pomocí modulů PROTONATE a LEaP Amber. Optimalizace pozic přidaných vodíků se provádí na základě molekulové dynamiky, která se skládá se ze tří po sobě jdoucích fází, během nichž se pohybují pouze vodíky. První fáze obsahuje 100 kroků, kde se minimalizuje přechod elektronů v konjugovaných systémech. Druhá fáze zahrnuje 500 kroků molekulové dynamiky při 300 K, přičemž všechny torzní potenciály, které zahrnují vodíky, jsou nastaveny na nulu. Třetí v sobě ukrývá 100 kroků minimalizace konjugovaných gradientů s torzním potenciálem k návratu na normální hodnoty. Pro všechny výpočty se používá silové pole AMBER parm99 [91].

H++ je možné využít pouze jako server. Lze tedy využívat pouze webové rozhraní [92] a nelze jej nainstalovat do vlastního zařízení.

2.5.15 Swiss-PdbViewer

Swiss-PdbViewer [93] nebo také DeepView je aplikace poskytující uživatelsky přívětivé rozhraní, které umožňuje analyzovat několik proteinů současně. Proteiny mohou být superponovány, aby bylo možné odvodit zarovnání struktur a porovnaly se jejich aktivní místa

nebo jiné relevantní části. Pomocí intuitivní grafiky a rozhraní nabídky lze snadno získat aminokyselinové mutace, vodíkové vazby, úhly a vzdálenosti mezi atomy [94].

Aplikace přidává chybějící polární atomy vodíku podle topologie GROMOS96 [93].

Swiss-PdbViewer vyvíjí od roku 1994 Nicolas Guex. Aplikace je úzce propojena se SWISS-MODEL, což je automatizovaný server pro modelování homologie vyvinutý ve Švýcarském institutu pro bioinformatiku ve Skupině strukturní bioinformatiky v Biozentru v Basileji [94].

Práce s touto aplikací je možná pouze skrz grafické rozhraní a neumožňuje práci pomocí příkazového řádku.

Kapitola 3

Implementace

V této kapitole je popsán postup k nalezení programu na určení poloh vodíků, který by byl součástí workflow na výpočet parciálních atomových nábojů. K nalezení vhodného programu bylo tedy zapotřebí provést výpočet a srovnání pro netriviální datovou sadu. V případě nutnosti i úpravy výstupních souborů z jednotlivých nástrojů. Nakonec bylo potřeba převést formát PDB do SDF formátu pro výpočet parciálních atomových nábojů.

3.1 Požadavky na nástroje

Prvním požadavkem je dostupnost daného nástroje zadarmo bez omezení. Dalším požadavkem bylo úspěšné nainstalování na zařízení Lenovo ThinkPad E550 se 64bitovým operačním systémem Ubuntu 20.04.2 LTS, procesorem Intel Core i5-5200U CPU @ 2.20GHz × 4 a grafickou kartou AMD Oland/Mesa Intel HD Graphics 5500 (BDW GT2). Třetím kritériem byla možnost spouštění pomocí příkazového řádku v terminálu. Poslední důležitou proměnou byl rychlý výpočet poloh atomů vodíků.

Právě z důvodu nedostatečné rychlosti byl vyřazen nástroj Corina (placený nástroj nainstalovaný v době testování na počítačích Masarykovy univerzity). Vzhledem k tomu, že při výpočtu jedné molekuly měl čas 7,5 minuty, což bylo více jak 2,5krát pomalejší než druhý nejpomalejší nástroj při testování, byl z testování vyřazen.

Na základě požadavků na nástroje bylo možné využít pro výpočet poloh vodíků následující programy: PyMOL (2.5.2), Chimera (2.5.1), OpenBabel (2.5.5), HAAD (2.5.3), Reduce (2.5.4) a GROMACS (2.5.6).

Ostatní nástroje, které nesplňovaly předem stanovené podmínky, slouží spíše pro přehled a kompletní popis problematiky zabývající se určování poloh vodíků. Výčet těchto nástrojů a odůvodnění vyřazení je znázorněn v tabulce 3.1. Samozřejmě výčet programů a algoritmů není úplný.

Nástroj	Důvod vyřazení z porovnávání
CNS	Nepodařilo se nainstalovat
AMBER	Obtížné vytvoření skriptu pro automatizaci
CORINA	Pomalé výpočty, placený
MCCE	Nepodařilo se nainstalovat
WHAT IF	Pouze webový nástroj, nemožnost spouštět přes terminál
YASARA	Placený
CHARMM	Nepodařilo se nainstalovat
H++	Pouze webový nástroj, nemožnost spouštět přes terminál
Swiss-Pdb Viewer	Nemožnost spouštět přes terminál

Tabulka 3.1: Přehled vyřazených nástrojů a důvod jejich vyřazení z porovnávání.

3.2 Automatizace

Kvůli možnosti porovnat nástroje bylo nutné je vyzkoušet na netriviální sadě struktur, což by manuálně zabralo mnoho času. Proto bylo vytvořeno automatické workflow, které pro danou vstupní sadu provede:

1. vytvoření kopie sady, kde budou vynechány původní vodíky,
2. přidání vodíků pomocí konkrétního nástroje,
3. výpočet RMSD mezi původními a nově přidanými vodíky.

Pro bod 1 ve workflow byl použit PyMOL pro odstranění vodíků a následného uložení do složky pak Python skript. Jednotlivé molekuly byly načteny do PyMOLu. Pro každou strukturu se odstranily vodíky a následně se uložila se do jiné složky.

U bodu 2 byl použit převážně Bash skript a u PyMOLu a Chimery Python skript. Tyto skripty načítají jednotlivé struktury bez vodíků. Následně určí polohy vodíků za pomoci daného nástroje a strukturu uloží do požadované složky ve formátu PDB. Výpočet vodíků u jednotlivých nástrojů je ve skriptu proveden podobným způsobem jako je znázorněno v části 3.3.

Bod 3 byl opět proveden za pomoci PyMOLu, jenž dokáže snadno vypočítat RMSD dvou struktur, a Python skriptu. Ve skriptu se využila funkce *align()* [95] z knihovny `pymol`. Tato funkce je totožná s příkazem `align` v PyMOLu. Funkce tedy provádí sekvenční zarovnání, poté vykoná strukturální superpozicování a nakonec provede cykly očištění, aby bylo možné vynechat odlehlé hodnoty ve struktuře nalezené během fitování. Tím se docílí lepšího zarovnání daných sekvencí.

Získané výsledky RMSD při použití knihovny `pymol` a v ní funkce aling() obsahují následující informace:

- RMSD po očištění – hodnota RMSD po vynechání odlehlých hodnot

- Počet seřazených atomů po očištění – počet atomů seřazených po vynechání odlehlých hodnot
- Počet cyklů očištění – počet cyklů, které se vykonaly. V každém cyklu jsou vynechány určité odlehlé hodnoty, které se poté nevyužívají ve výpočtu RMSD.
- RMSD před očištěním – hodnota RMSD bez vynechání odlehlých hodnot
- Počet seřazených atomů před očištěním – počet atomů bez vynechání odlehlých hodnot
- Raw score – součet skóre záměny a mezer získaných během zarovnání. Záměna je přítomnost neidentické aminokyseliny v dané poloze. Mezera je prostor vložený do zarovnání ke kompenzaci vložené nebo odstraněné části sekvence v jedné sekvenci vzhledem k druhé.
- Počet zarovnaných residuí – počet residuí, které se porovnávaly

Všechny výše zmíněné hodnoty RMSD se poté uloží pro jednotlivé molekuly do souboru s formátem CSV, jenž nese název daného nástroje.

3.3 Spouštění jednotlivých nástrojů před automatizací

Předtím než se vytvořily skript pro automatizaci spouštění, měly jednotlivé nástroje jiný způsob spouštění pro dopočítání vodíků.

PyMOL a Chimera mají dost podobné spouštění, kde se jedná o vytvořený program v Pythonu, kde ale každá metoda používá jinou knihovnu. PyMOL využívá knihovnu `pymol` a Chimera zase knihovnu `chimera`. Pomocí těchto knihoven lze používat nástroje stejně jako za pomoci příkazového řádku, který oba programy obsahují.

Pro názornou ukázku je níže použit vzorový soubor input.pdb pro jednu molekulovou strukturu, pro kterou budou dopočítány vodíky. Pokud nějaké nástroje potřebují uložit soubory, použije se vzorový soubor output.pdb. Jednotlivé nástroje se spouští těmito příkazy:

- PyMOL:

  ```
  pymol -cq pymol_script.py input.pdb,
  ```

  kde parametr `-cq` se používá, aby se nespustilo vestavěné PyMOL GUI a tím se urychlila doba výpočtu. Ukázka podobného kódu, který se využil je zobrazen na obrázku 3.1.

- Chimera:

  ```
  chimera --nogui --script chimera_script.py input.pdb,
  ```

  kde se parametr `--nogui` opět využije, aby se zabránilo spuštění vestavěného GUI v nástroji Chimera. Parametr `--script` udává, aby se k počítání použil připravený skript vytvořený v Pythonu. Ukázka kódu v chimera_script.py je znázorněn na obrázku 3.2.

```
1  import sys
2  import pymol
3
4  soubor = sys.argv$[3]$ #nacteni input.pdb
5
6  slozka_bez_vodiku = "cesta, kde se nachazi slozka se strukturami
   bez vodiku"
7
8  slozka_ulozit = "cesta, kde se ulozi struktura s vypocitanymi
  vodiky za vyuziti PyMOLu"
9
10 pymol.finish_launching() #spusti PyMOL
11 pymol.cmd.load(slozka_bez_vodiku + soubor) #nacte pozadovanou
  strukturu
12 pymol.cmd.h_add() #vypocita a prida vodiky do struktury
13 pymol.cmd.save(slozka_ulozit + soubor) #ulozi strukturu s vodiky
14 pymol.cmd.delete("all") #ukonci praci se vsemi otevrenymi
  strukturami
15 pymol.cmd.quit() #ukonci praci PyMOLU
```

Obrázek 3.1: Ukázka kódu pymol_script.py.

```
1  import sys
2  from chimera import runCommand
3
4  soubor = sys.argv$[3]$ #nacteni input.pdb
5
6  slozka_bez_vodiku = "cesta, kde se nachazi slozka se strukturami
   bez vodiku"
7
8  slozka_ulozit = "cesta, kde se ulozi struktura s vypocitanymi
  vodiky za vyuziti Chimery"
9
10 runCommand("open " + slozka_bez_vodiku + soubor) #nacte
  pozadovanou strukturu
11 runCommand("addh") #vypocita a prida vodiky do struktury
12 runCommand("write 0 " + slozka_ulozit + soubor) #ulozi strukturu
   s vodiky
13 runCommand("close all") #ukonci praci se vsemi otevrenymi
  strukturami
```

Obrázek 3.2: Ukázka kódu chimera_script.py.

- HAAD:

  `./HAAD input.pdb`,

  kde soubor s dopočítanými vodíky je poté uložen s příponou .pdb.h v adresáři, ve kterém se spustil. V rámci automatizace je soubor přejmenován tak, že se odstraní přípona .h, a je přesunut do požadované složky. Z tohoto důvodu nástroj HAAD, oproti ostatním nástrojům, nevytváří zcela validní PDB formát souboru.

- Reduce:

  `reduce input.pdb > output.pdb`.

  Reduce vypisuje výsledek na standardní výstup, takže je pak nutné jej uložit do požadovaného souboru.

- OpenBabel:

  `babel -ipdb input.pdb -opdb output.pdb -h`,

  kde -ipdb a -opdb značí vstupní, respektive výstupní soubor s příponou .pdb a parametr -h slouží k tomu, aby se určily polohy vodíků.

- GROMACS:

  `gmx pdb2gmx -f input.pdb -o output.pdb`,

  kde pdb2gmx je parametr pro dopočítání vodíků, -f značí vstupní soubor a -o výstupní soubor. Parametry pro silové pole a model vody musí buď uživatel zadat ručně nebo přidat do příkazu postupné vypisování požadovaných parametrů. Při ručním zadávání se v příkazovém řádku postupně možnosti voleb vypíší. Prvně se vybírá silové pole a jako druhé se volí model vody. V tabulkách A.1 a A.2 jsou znázorněny všechny možné volby.

3.4 Výpočet parciálních nábojů

Pro výpočet parciálních nábojů se využil nástroj MACH a metoda Extended charge equilibration (EQeq) v něm, viz část 2.4.

Z vypočítaných parciálních nábojů se provádělo srovnání mezi náboji z originálních struktur a struktur, jež se získaly jednotlivými nástroji na určení poloh vodíků. Pro toto porovnání byl použit nástroj vytvořený v rámci mé bakalářské práce [41], jenž je volně dostupný na adrese https://github.com/SimonZouvala/MGC-OGC. Tento nástroj v sobě obsahuje funkci pro statistické porovnání dvou sad nábojových hodnot. Nástroj taktéž vytváří graf za pomoci knihovny Matplotlib. Pro srovnání využívá statistických metod MAE, ABSMAX, RMSD a PCC. Pro jeho použití však bylo nutné provést drobné úpravy pro čtení souborového formátu CHG, ve kterém byly získané náboje za pomoci nástroje MACH.

3.5 Úprava výsledných souborů

Pro další práci se soubory bylo potřeba provést menší úpravy na výsledných souborech každého nástroje. Bylo nutné sjednotit názvy a pořadí atomů vodíků v získaných souborech, jelikož se tyto hodnoty lišily v jednotlivých nástrojích. Dále se převedly soubory z PDB formátu do jednoho souboru ve formátu SDF, aby bylo možné vypočítat parciální náboje za pomoci nástroje MACH.

3.5.1 Automatické spárování vypočítaných a původních vodíků

K co nejvíce přesným výsledkům, jak u RMSD tak i u výpočtu hodnot parciálních nábojů pomocí nástroje MACH, potřebujeme vědět, který vodík v nové struktuře odpovídá kterému v té původní. Nástroje ale mají jiný způsob pojmenování vodíku a tyto názvy se od původních lišily. Z tohoto důvodu bylo nejdříve nutné názvy sjednotit a seřadit. K tomuto účelu byl vytvořen skript.

Tento skript bere originální struktury a struktury, kde se vodíky predikovaly. Ukázka pseudokódu pro lepší pochopení je znázorněna ve schématu 3.3. Postupně se pro jednotlivá residua vezmou všechny atomy vodíků (2,3). Následně se pro jednotlivé originální vodíky v residuu napočítají vzdálenosti dle 3D souřadnic od všech vypočítaných vodíků (6). V takto vytvořené matici se spárují vodíky vždy podle nejmenší hodnoty mezi originálním a spočítaným (9). Pokud nastane situace, že pro dva vodíky se určí jako nejlepší jeden a ten samý v původní struktuře, tak se původní spáruje s tím dopočítaným, který má menší vzdálenost (13). Ten nespárovaný se poté spáruje tak, že se zjistí, které vodíky nejsou propojeny (15) a provede se to samé hledání nejmenších vzdálenosti (17). Tímto způsobem se naleznou atomy vodíků, které si odpovídají. Po nalezení dvojic vodíků se následně upraví číslo atomu a název vypočítaného vodíků podle původního atomu vodíku (18). Takto je docíleno co nejlepší srovnání u všech nástrojů.

3.5.2 Převod do SDF

Jelikož nástroj MACH pracuje se souborovým formátem SDF, bylo do něj potřeba převést všechny soubory PDB. Pro převod se použil nástroj OpenBabel, jež je pro tyto operace určen. Více o tomto nástroji je v části 2.5.5.

Jakmile byl získán jeden SDF soubor pro struktury, kde byly určeny polohy vodíků, a pro originální struktury, bylo možné vypočítat parciální atomové náboje.

```
1  for residua v struktuře do
2  |  vytvoří se matice M x N, kde M je počet vodíků v původní struktuře a N ve
   |   vypočítané struktuře for i=1 to M do
3  |  |  for j=1 to N do
4  |  |  |  vypočítá se matice o velikosti [i,j] kde budou vzdálenosti původních a
   |  |  |   vypočítaných vodíků
5  |  |  end
6  |  |  for j=1 to N do
7  |  |  |  pár = nejmenší vzdálenost v daném řádku i matice
8  |  |  end
9  |  |  pár se přidá do nezkontrolované páry
10 |  end
11 |  nezkontrolované páry se zkontrolují, aby byl použit původní a vypočítaný
   |   vodík pouze jednou
12 |  if chybí původní a vypočítaný vodík ve zkontrolované páry then
13 |  |  vypočítá vzdálenosti mezi všemi chybějícími původními a vypočítanými
   |  |   vodíky a přidají do zkontrolované páry
14 |  end
15 |  zkontrolované páry se opět projdou a vezmou se jen ty nejmenší vzdálenosti.
   |     // vypočítané vodíky již nemůžou mít stejný původní vodík
16 |  ∀ vypočítané vodíky ∈ vypočítaném residuu: změníme název a atomové číslo
   |   vodíku podle propojení z zkontrolované páry
17 |  ∀ atomy ∈ vypočítaném residuu: seřadíme dle atomového čísla.
18 end
```

Obrázek 3.3: Ukázka kódu pro automatické spárování vypočítaných a původních vodíků a následné přejmenování názvů vodíků a atomového čísla.

Kapitola 4

Výsledky a diskuse

V této kapitole je popsán použitý dataset molekulových struktur a uvedeno porovnání použitých nástrojů mezi sebou. Srovnání probíhá na základě rychlosti predikce poloh vodíků a jejich míry podobnosti s původní polohou. Dále je popsána míra ovlivnění parciálních atomových nábojů a nejčastější typ residua, u kterého nebyl predikován stejný počet vodíků. Kapitolu uzavírá diskuse, kde jsou shrnuty všechny získané výsledky a problémy jednotlivých nástrojů.

4.1 Dataset

Cílem vytvoření datasetu bylo porovnání jednotlivých nástrojů mezi sebou. Právě kvůli zajištění co největší možnosti porovnání jednotlivých nástrojů mezi sebou bylo zapotřebí mít v datasetu dostatečně velký počet struktur a bylo nutné, aby struktury byly různorodé. Za tímto účelem byla využita databáze RCSB PDB. Celkový počet struktur získaných z této databáze je 2199.

Pro porovnání dopočítaných poloh vodíků s polohami experimentálně naměřenými a zjištění jejich míry odchylky bylo zapotřebí, aby experimentálně naměřené struktury z RCSB PDB vodíky opravdu obsahovaly. Z tohoto důvodu byl použit filtr pro vyhledávání v databázi. Pomocí filtru byly získány struktury molekul, které byly stanoveny metodou NMR. Žádný další filtr nebyl použit, aby se docílilo různé variability struktur.

Dataset tedy obsahuje různorodé struktury molekul. Nicméně různorodá je i kvalita, jelikož některé struktury z 2199 molekul neobsahují dostatečné informace pro některé nástroje. Velikost počtu struktur, které se porovnávají za pomoci statistické metody RMSD, je tedy pro každý nástroj odlišná.

Obecně z datové sady byly vyřazeny struktury, které měly jeden z těchto dvou nedostatků: Prvním z nich je to, že i když byl použit filtr pro hledání struktur získaných pomocí NMR analýzy, tak ve struktuře vodíky chyběly, příkladem může být molekula 3PAT. Druhým nedostatkem je, že struktura obsahovala pouze nestandardní residua, jinými slovy v PDB souboru byly přítomny pouze atomy označené jako HETATM nikoli jako ATOM. Příkladem takové struktury je 2RMQ. Konkrétně šlo o 35 molekul, kde buď chyběly vodíky nebo měly jen jeden vodík v residuu, a nebo se nejednalo o standardní residua.

Dataset obsahuje i molekuly, které mají ve svých residuích neobvykle velmi malý počet vodíků, ukázkou takové struktury je molekula 2MRB. Objevily se i molekuly, jenž měly

místo třípísmenného názvu residua použito pouze jednopísmenné označení. Příkladem struktury obsahující pouze jednopísmenné názvy je 2HEM.

4.1.1 Shrnutí informací o datové sadě

Pro testování bylo použito 2199 struktur z PDB, které byly získány pomocí metody NMR. Jelikož nebylo stanoveno jiné kritérium, tak jejich výběr byl náhodný. V datasetu se nachází různorodé velikosti a typy struktur tak, aby se zjistili limity jednotlivých nástrojů při dopočítání vodíků.

Jelikož u struktur určených pomocí metody NMR se často objevuje více než jeden model [96], byl pro porovnání použit vždy ten první.

4.2 Porovnání nástrojů

Porovnávali jsme 6 softwarových nástrojů, konkrétně: PyMOL, HAAD, Reduce, OpenBabel, Chimera a GROMACS. Některé nástroje dokázaly vypočítat vodíky pro všechny struktury, jmenovitě PyMOL, Reduce, OpenBabel. Ostatní nástroje nezvládly zpracovat nebo dopočítat vodíky pro všechny struktury (viz tabulka 4.1. Přesněji nástroj HAAD zvládl napočítat 2007 struktur (konkretně zvládl vytvořit 2194 struktur, ale 187 struktur napočítal jen jako jeden glycin). Dále nástroj Chimera zvládl napočítat 292 struktur. Poslední nástroj Gromacs dopočítal vodíky pouze pro 64 struktur se silovým polem CHARMM, 75 struktur s AMBERG a 74 struktur s GROMOS.

Níže jsou srovnány rychlosti a přesnosti nástrojů vůči sobě.

4.2.1 Srovnání rychlosti

Měření, jak rychle jednotlivé nástroje zvládnou vypočítat vodíky pro 2199 struktur, se provádělo až poté, co bylo možné celý proces počítání automatizovat pomocí skriptů. Bližší popis automatizace pro vybrané nástroje je v části 3.2.

Jakmile byla automatizace dokončena pro všechny zvolené nástroje, bylo možné měřit, za jak dlouho daný nástroj dopočítá vodíky a uloží struktury s vodíky do předem určené složky.

Pro určení doby výpočtu se použil příkaz `time` [97] v příkazovém řádku, jenž se přidával před skript, který spouštěl daný nástroj.

Příkaz `time` vrací tři hodnoty. První určuje skutečně uběhnutý čas od spuštění do ukončení, který byl nazván jako reálný čas. Druhá hodnota značí čas uživatelské centrální procesové jednotky (CPU). Tato hodnota byla nazvána jako uživatelský čas. Poslední číslo vrací čas systémové centrální procesové jednotky (CPU). Poslední hodnota byla označena jako systémový čas.

Souhrn všech časů z jednotlivých nástrojů je znázorněn v tabulce 4.1.

Nejrychlejší nástroj co se týče systémového času je Chimera s 1,05 s, ale ta zvládla napočítat jenom 292 struktur. Právě to je nejspíše důvod nízké hodnoty systémového času. Na druhou stranu nástroj PyMOL s 1,13 s zvládl napočítat všechny struktury. Nástroj HAAD (7,46 s) následuje na dalším místě. Naopak nejpomalejší nástroj v systémovém času byl Reduce (1 min 23,15 s), ale i ten zvládl napočítat všechny struktury.

Nástroj	Počet napočítaných molekul	čas [s]			Poměr struktura/ sekunda
		uživatelský	systémový	reálný	
PyMOL	2199	48,47	1,13	53,16	0,0242
HAAD	2194	47,96	7,46	68,29	0,0311
Reduce	2199	441,84	83,15	540,07	0,2456
OpenBabel	2199	192,54	28,26	226,17	0,1029
Chimera	292	204,49	1,05	210,76	0,7218
GROMACS *CHARMM*	64	227,80	18,94	250,91	3,9205
GROMACS *GROMOS*	74	168,56	22,66	195,34	2,6397
GROMACS *AMBER*	75	242,71	20,47	267,81	3,5708

Tabulka 4.1: Rychlost výpočtů jednotlivých nástrojů.

Z hlediska uživatelského času je nejrychlejší HAAD s 47,96 s. Na druhém místě je PyMOL (48,47 s). Nejpomalejší byl opět Reduce (7 min 21,84 s).

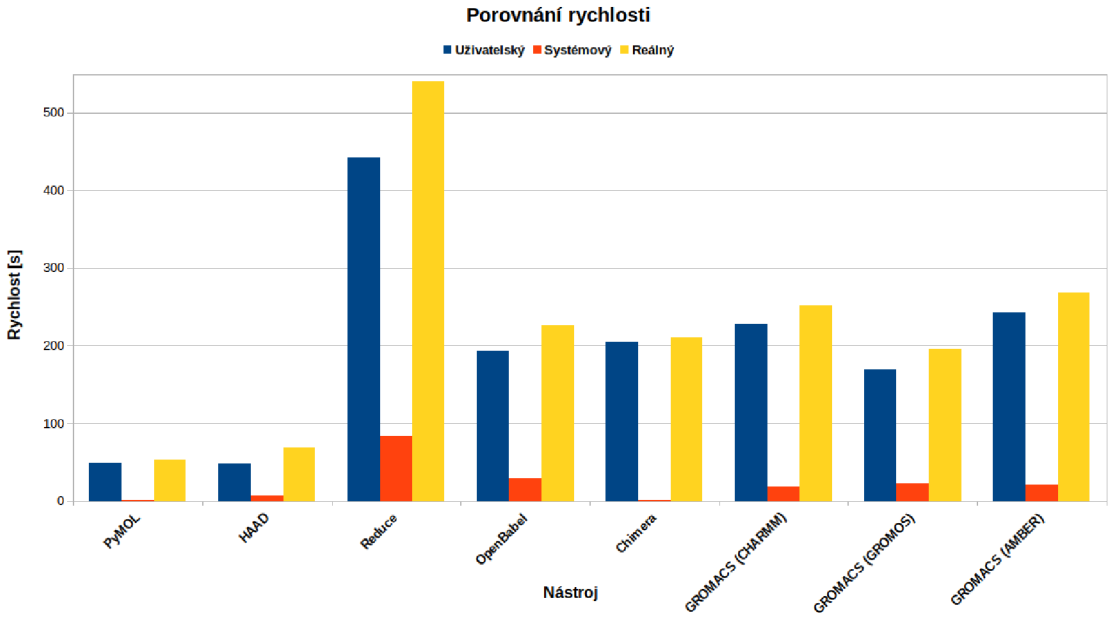
Celkově nejrychlejší nástroj, tedy v reálném čase, je PyMOL s 53,16 s. Druhý je HAAD (1 min 8,29 s). Nejpomalejší je již zmíněný nástroj Reduce (9 min 0,07 s).

Na obrázku 4.1 jsou graficky znázorněny rychlosti jednotlivých nástrojů. Na základě výsledků lze usuzovat, že PyMOL a HAAD vedou v rychlosti v poměru počtu struktur za sekundu (0,0242 a 0,0311 s na jednu strukturu). Nicméně ani Reduce s celkově nejdelším časem nemá tento poměr špatný (0,2456 s na jednu strukturu).

4.2.2 Srovnání přesnosti

Pro měření přesnosti daných nástrojů se použil program PyMOL, jenž dokáže snadno vypočítat RMSD atomů dvou struktur.

Jelikož v tak velkém množství struktur jako je 2199 molekul se nacházely struktury s různými hodnotami pH, nenastavovala se u žádného nástroje tato hodnota pro výpočet a ponechala se tedy defaultně nastavená. Toto nastavení u jednotlivých nástrojů způsobilo to, že se u některých molekul nevypočítaly atomy vodíků pro dané residuum ve stejném počtu jako v originální struktuře. Z tohoto důvodu je provedeno dvojí srovnání. V prvním se srovnávají molekuly, kde je počet vypočítaných vodíků shodný s originálem. Druhé srovnává i ty molekuly, kde počet vodíků neodpovídá.



Obrázek 4.1: Graf srovnání rychlostí pro jednotlivé nástroje. Modrá barva značí uživatelský čas, oranžová systémový čas a žlutá reálný čas. Počet napočítaných struktur jednotlivých nástrojů je blíže popsán v tabulce 4.1.

Srovnání struktur se shodným počtem vodíků

V rámci porovnání struktur, pro které testované nástroje napočítaly stejný počet vodíků jako byl v experimentálních strukturách, se nejlépe umístil Reduce. Tento nástroj vypočítal 354 struktur se stejným počtem atomů vodíků. Na druhém místě co se týče počtu struktur je HAAD s 326 molekulami. Jako třetí je PyMOL s 203. Chimera zvládla 78 a OpenBabel jen 3 struktury. Nástroj GROMACS ani v jednom silovém poli nezvládl napočítat stejný počet vodíků jako v původní struktuře.

Získané hodnoty RMSD bez očištění odlehlých hodnot ukazují, že nejmenší průměrnou a směrodatnou odchylku měl OpenBabel a Chimera. Nicméně oba tyto nástroje napočítaly oproti Reduce, PyMOLu a HAAD malý počet struktur. Proto dle počtu spočítaných molekul a přesnosti jsou na tom nejlépe Reduce a HAAD. Reduce napočítal nejvíce struktur a jeho průměrná hodnota RMSD je 0,277. Nicméně jeho směrodatná odchylka 0,907 je vyšší než je tomu u programu HAAD, který má průměr 0,293 a směrodatnou odchylku 0,235. U PyMOLu jsou hodnoty RMSD v průměru 0,357 s odchylkou 1,179.

V tabulce 4.2 jsou ukázány průměrné hodnoty RMSD a směrodatná odchylka těchto hodnot před očištěním. Tabulka 4.3 znázorňuje zase hodnoty po očištění.

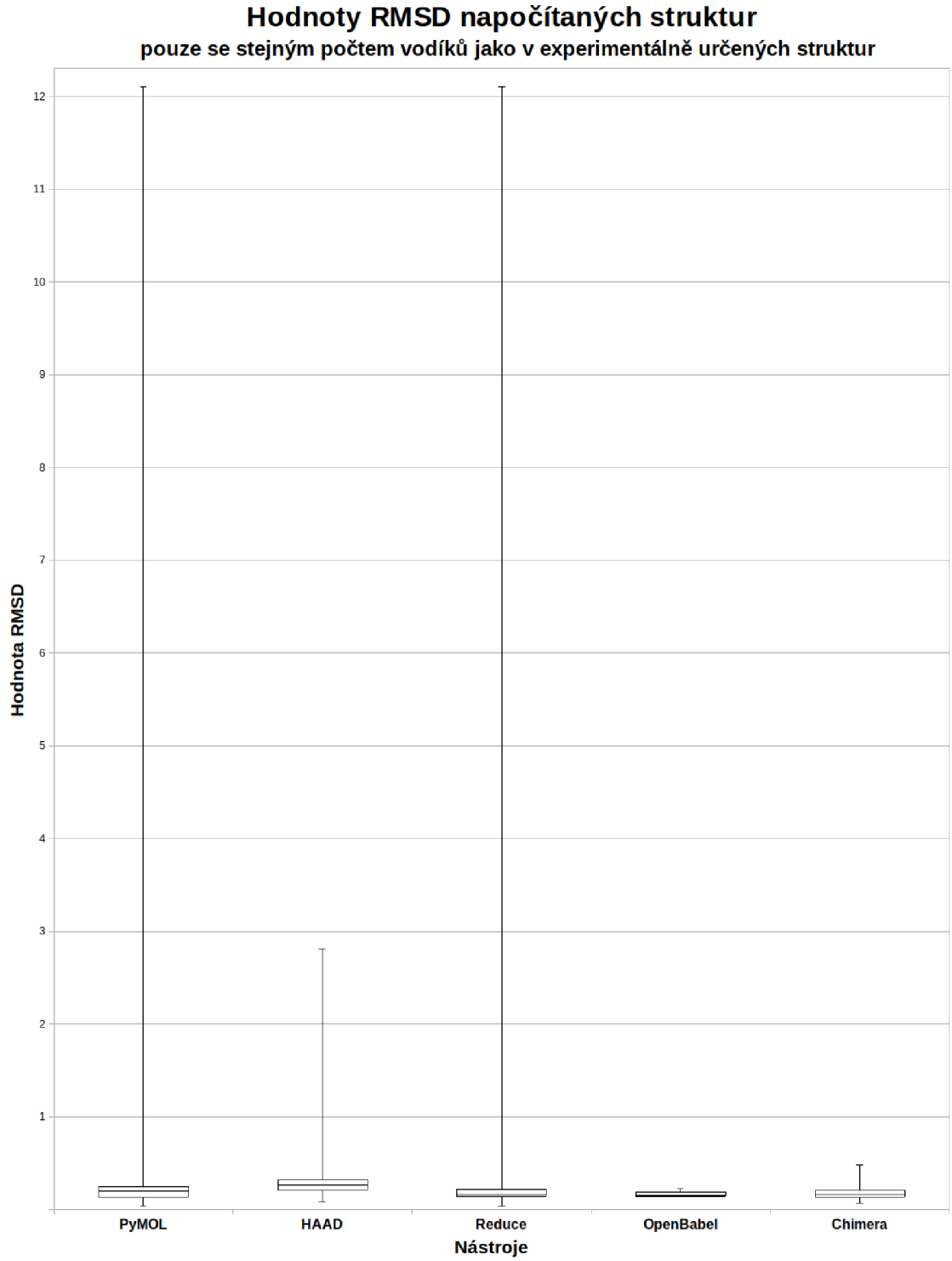
Krabicový graf znázorněný na obrázku 4.2 ukazuje, že PyMOL a Reduce mají větší průměrné hodnoty kvůli třem strukturám (2RQO, 2N9Q a 2KLW). U těchto struktur nastala chyba během zarovnání při výpočtu RMSD PyMOLem, jelikož byly odstraněny při párování vodíků residua, která se zařazují jako HETATM. Proto je hodnota RMSD těchto molekul tak vysoká. Ukázka špatně zarovnané struktury 2N9Q je na obrázku 4.3.

Nástroj	Počet napočítaných molekul	RMSD před očištěním	Seřazené atomy před očištěním	Raw score	Zarovnaná residua
PyMOL	203	0,357 ±1,179	716,966 ±513,460	224,192 ±158,932	25,946 ±17,799
HAAD	326	0,293 ±0,235	688,089 ±436,939	460,316 ±301,131	88,448 ±57,395
Reduce	354	0,277 ±0,907	908,918 ±1540,067	296,708 ±509,553	49,240 ±102,167
OpenBabel	3	0,172 ±0,047	460,667 ±413,384	169,000 ±117,154	32,000 ±27,713
Chimera	78	0,178 ±0,069	1410,205 ±1079,478	453,115 ±349,244	80,333 ±70,671

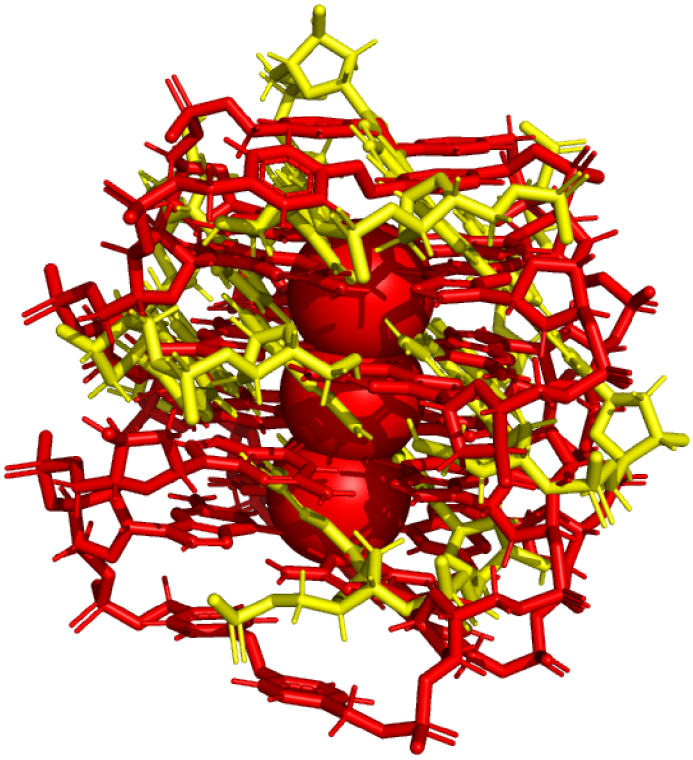
Tabulka 4.2: Průměrné hodnoty a směrodatné odchylky RMSD před očištěním odlehlých hodnot všech napočítaných struktur jednotlivých nástrojů.

Nástroj	Počet napočítaných molekul	RMSD po očištění	Seřazené atomy po očištění	Očišťovací cykly	Zarovnaná residua
PyMOL	203	0,081 ±0,818	498,138 ±366,015	4,956 ±0,363	25,946 ±17,799
HAAD	326	0,079 ±0,020	607,190 ±399,871	4,479 ±0,844	88,448 ±57,395
Reduce	354	0,112 ±0,616	796,466 ±1498,538	4,237 ±0,964	49,240 ±102,167
OpenBabel	3	0,006 ±0,004	314,000 ±295,437	5,000 ±0,000	32,000 ±27,713
Chimera	78	0,005 ±0,004	1007,064 ±775,447	5,000 ±0,000	80,333 ±70,671

Tabulka 4.3: Průměrné hodnoty a směrodatné odchylky RMSD po provedeném očištění odlehlých hodnot všech napočítaných struktur jednotlivých nástrojů.



Obrázek 4.2: Krabicový graf hodnot RMSD pro napočítané struktury se stejným počtem vodíků jako u experimentálně určených. Počty struktur pro jednotlivé nástroje jsou v tabulkách 4.2 a 4.3.



Obrázek 4.3: Struktura 2N9Q (přesněji molekula DNA), která se špatně zarovnala při měření RMSD PyMOLem. Červená barva značí experimentálně určenou strukturu a žlutá barva strukturu vypočítanou PyMOLem.

Srovnání včetně neshodného počtu vodíků

Při zahrnutí struktur, kde nebyl stejný počet spočítaných vodíků s experimentálně určeným počtem, se nebraly v potaz struktury, kde byl problém se spárováním těchto vodíků. Příkladem problému může být menší počet vodíků v původní struktuře. Z tohoto důvodu je počet struktur u každého nástroje nižší než jen při vynechání problémových struktur v datasetu.

I při tomto srovnání zvládl Reduce spočítat vodíky pro největší počet molekul a to pro 1572. Druhý nástrojem s největším počtem byl PyMOL se 1134 strukturami. Následoval HAAD se 611 a OpenBabel se 192 molekulami. Chimera napočítala vodíky pro 134 struktur. Gromacs dokázal spočítat nejméně a to 46 se silovým polem AMBERG, 38 s polem CHARMM a pouze 6 při využití GROMOSu.

Při porovnání průměrných hodnot RMSD, jejich směrodatných hodnot a při přihlédnutí počtu struktur, u kterých se zvládly napočítat vodíky, je možné říct, že Reduce je opět nejvíce přesný s průměrnou hodnotu 0,204 a směrodatnou odchylkou 0,449. Jako druhý nejpřesnější se dá označil PyMOL, i když jeho průměrná hodnota 0,289 a směrodatná odchylka 0,520 je vyšší než u programu HAAD s průměrem 0,269 a nejnižší odchylkou 0,192, zvládl napočítat téměř dvakrát více struktur než HAAD. V tabulce 4.4 jsou znázorněny i zbývající průměry

RMSD a jejich odchylky před očištěním pro ostatní nástroje. Tabulka 4.5 znázorňuje tyto hodnoty po očištění.

Nástroj	Počet napočítaných molekul	RMSD před očištěním	Seřazené atomy před očištěním	Raw score	Zarovnaná residua
PyMOL	1134	0,289 ±0,520	1516,042 ±1772,020	505,729 ±594,962	93,252 ±116,170
HAAD	600	0,269 ±0,192	766,248 ±440,014	511,099 ±298,978	98,072 ±57,110
Reduce	1572	0,204 ±0,449	1591,298 ±1675,575	531,450 ±562,961	98,760 ±109,247
OpenBabel	192	0,865 ±0,856	450,109 ±412,950	195,500 ±203,722	34,182 ±39,037
Chimera	134	0,184 ±0,138	1670,276 ±1414,712	545,022 ±454,621	100,888 ±91,945
GROMACS *CHARMM*	38	0,180 ±0,041	1528,658 ±1027,757	576,026 ±373,983	109,263 ±73,647
GROMACS *GROMOS*	6	0,193 ±0,039	576,000 ±495,825	298,667 ±263,874	58,333 ±51,008
GROMACS *AMBER*	46	0,180 ±0,046	1535,826 ±963,389	584,870 ±354,314	109,783 ±69,646

Tabulka 4.4: Průměrné hodnoty a směrodatné odchylky RMSD před očištěním odlehlých hodnot všech napočítaných struktur jednotlivých nástrojů.

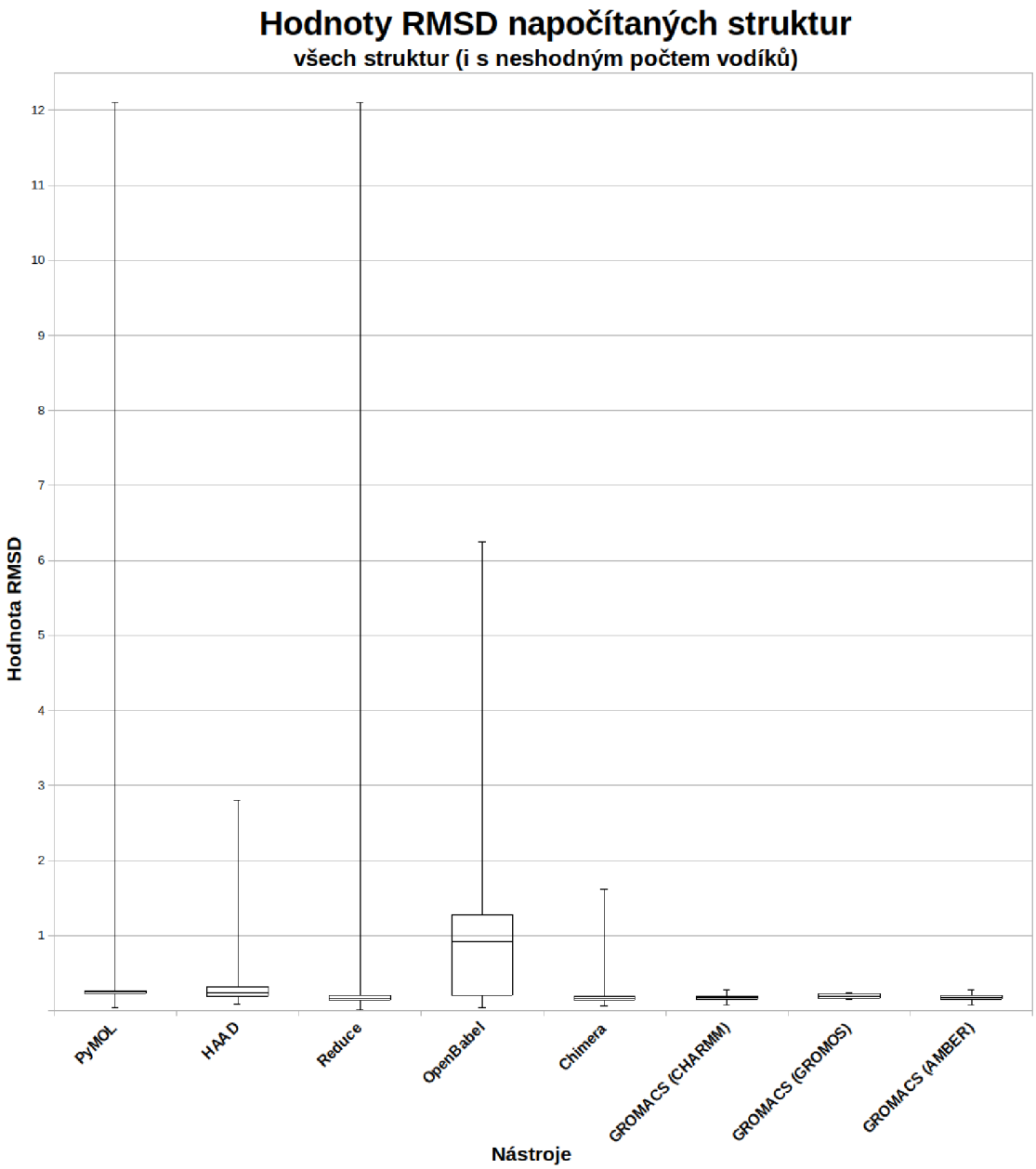
Krabicový graf na obrázku 4.4 opět znázorňuje ovlivnění PyMOLu a Reduce špatně zarovnanými strukturami, jak už bylo zmíněno v předešlém srovnání. Nicméně zde je ještě ukázáno, že u i OpenBablu se stala chyba se zarovnáním struktur. Přesněji se jedná o 2KP4, 2N1E, 2KP3 a 2N9Q.

Nejčastější residua s nestejným počtem vodíků

Jak již bylo zmíněno, nebyla nastavena přesná hodnota pH, se kterým se měly jednotlivé struktury počítat, a tak jednotlivé nástroje počítaly vždycky s defaultně nastavenou hodnotou pH. Důvodem bylo, že v datasetu se objevují struktury měřené při různém pH a hodnota pH u nich není uvedena. Na základě toho, bylo zjištěno, která residua se nejčastěji napočítala s nestejným počtem vodíků.

Počet struktur, které byly vzaty pro jednotlivé nástroje, je shodný s počtem v tabulkách 4.4 a 4.5.

Ve většině případů se jednalo o to, že původní struktury měly větší počet atomů vodíků než ty, kde se vodíky dopočítávaly. Kromě nástroje OpenBabel, kde 54,26 % residuí mělo



Obrázek 4.4: Krabicový graf hodnot RMSD pro všechny napočítané struktury jednotlivých nástrojů. Počty struktur pro jednotlivé nástroje jsou v tabulkách 4.4 a 4.5.

Nástroj	Počet napočítaných molekul	RMSD po očištění	Seřazené atomy po očištění	Očišťovací cykly	Zarovnaná residua
PyMOL	1134	0,021 ±0,347	1029,224 ±1200,205	4,992 ±0,154	93,252 ±116,170
HAAD	600	0,075 ±0,016	683,707 ±397,381	4,315 ±0,878	98,072 ±57,110
Reduce	1572	0,089 ±0,293	1467,485 ±1592,087	4,268 ±0,878	98,760 ±109,247
OpenBabel	192	0,060 ±0,453	328,526 ±309,129	4,714 ±0,676	34,182 ±39,037
Chimera	134	0,006 ±0,003	1208,239 ±1036,701	5,000 ±0,000	100,888 ±91,945
GROMACS *CHARMM*	38	0,047 ±0,007	1385,763 ±943,306	4,868 ±0,343	109,263 ±73,647
GROMACS *GROMOS*	6	0,000 ±0,000	460,167 ±401,703	3,667 ±0,516	58,333 ±51,008
GROMACS *AMBER*	46	0,045 ±0,012	1386,348 ±885,750	4,891 ±0,315	109,783 ±69,646

Tabulka 4.5: Průměrné hodnoty a směrodatné odchylky RMSD po provedeném očištění odlehlých hodnot všech napočítaných struktur jednotlivých nástrojů.

více atomů vodíku v originálních strukturách a 45,74 % napočítaných residuí mělo více vodíků. U tohoto nástroje byl celkově nejvíce neshodný počet atomů v residuu arginin v 11,23 % ze všech residuí. Residuem, u kterého OpenBabel určil menší počet vodíků, byl glycin v 14,41 % ze všech residuí, kde byl menší počet vodíků v originálu. Na druhou stranu se přepočítal u lysinu v 13,39 % ze všech residuí, kde byl větší počet vodíků než v původním residuu. Jelikož procenta nejvíce problémových residuí nejsou moc vysoká oproti ostatním, nelze úplně jednoznačně říct, které residuum dělalo nástroji OpenBabel největší problém. Je to způsobeno malým počtem vypočítaných struktur. Celkově se tedy u OpenBabel ve 192 struktur napočítalo 3249 residuí.

Podobný příklad, kde nelze určit nejproblémovější residuum, je u nástroje GROMACS. Při použití silového pole CHARMM byla v 19,70 % největší neshoda v počtu u residua leucin a zároveň je to se stejným procentem i residuum, kde tento nástroj napočítal méně vodíků než v aminokyselinách experimentálně naměřených struktur. Při použití CHARMM se tedy napočítalo 38 struktur z toho bylo 1726 residuí napočítáno neshodně v počtu atomů. V daném silovém poli a v silovém poli GROMOS tento nástroj nenapočítal více atomů vodíků v žádném residuu. Proto při použití pole GROMOS bylo napočítáno 6 struktur, z toho 349 residuí mělo problém s počtem atomů. Z toho nejčastěji v 10,32 % byl problém také s leucinem. Při použití silového pole AMBERG se napočítalo 46 struktur z toho 2121

residuí mělo nestejný počet atomů. Ve dvou případech se nástroj GROMACS přepočítal ve vodících a to jen u lysinu. Residuum s největším počtem neshodných atomů a zároveň s méně určeným počtem než u původních residuí, byl opět leucin.

Chimera sice napočítala pouze 134 struktur, z toho měla neshodu u 241 residuí. Poměrově bylo určeno v 86,72 % méně a 13,28 % více atomů vodíku. U residua histidinu měl tento nástroj 72,20 % v obou případech neshody a 82,30 % ze všech případů, kdy bylo napočítáno méně vodíků. V případě určení více atomů vodíků to bylo v 25 % případů u glycinu.

Histidin měl nejčastěji špatný počet atomů vodíku i u dalších nástrojů. Například u nástroje HAAD to byl právě histidin, který měl nestejný počet atomů a to v 95,07 % ze všech případů a 95,48 % ze všech případů, kdy bylo méně vodíků. V 5 případech z 10 to byl právě histidin, který měl více vodíků než v originále. HAAD napočítal 600 struktur a z toho 1095 residuí mělo nestejný počet atomů.

U PyMOLu šlo v 52,40 % ze všech případů a 54,26 % z případů, kdy bylo méně vodíků, také o histidin. Největší problém s větším počtem určených vodíků bylo u residua deoxythymidinu a to ve 14,91 % případů. PyMOL napočítal 1134 struktur, z toho měl nestejný počet atomů u 3109 residuí. Poměrově to bylo 96,33 % méně a 3,67 % více určených vodíků.

Nakonec Reduce, jenž napočítal 1572 struktur, z toho 5229 residuí mělo neshodu v atomech. Poměrově to bylo 97,25 % méně a jen 2,75 % více určených atomů vodíku. Reduce dopadl stejně jako PyMOL, nicméně s více jasnými hodnotami pro histidin a to 74,47 % ze všech případů a 76,48 % z případů, kdy bylo určeno méně vodíků. V 11,81 % případů více určených atomů vodíků to bylo opět u nukleové kyseliny deoxythymidinu.

Důvodem nestejného počtu vodíků u histidiny, který se objevoval jako nejčastější residuum, je pravděpodobně to, že jeho náboj je ve fyziologických podmínkách velmi citlivý na změny pH. Při nižším je histidin protonovaný a nese kladný náboj, naopak při vyšším pH není nabitý [98].

Shrnutí

Z výše popsaného srovnání je možné usoudit, že nástroj Reduce dosahuje nejmenších hodnot RMSD na počet spočítaných struktur. PyMOL a HAAD mají tyto hodnoty jen drobně vyšší něž Reduce. Co se týče Chimery, OpenBabelu a Gromoss, tak důvody jejich malého počtu spočítaných struktur jsou blíže určeny v části 4.3.

Na druhou stranu residuem, kde se nejčastěji nesešel stejný počet vodíků jako v původní struktuře, je amynokyselina histidin, u níž se objevoval problém u třech nástrojů s největším počtem napočítaných struktur. PyMOL a Reduce napočítaly nejčastěji více atomů vodíků u nukleové kyseliny deoxythymidinu.

4.2.3 Srovnání nábojů

Pro srovnání parciálních atomových nábojů se použily stejné molekuly jako v části 4.2.2 Srovnání se shodným počtem vodíků. Protože byl jako srovnávací metrika použit Pearsonův korelační koeficient, tak musel být stejný počet atomů jak u původní struktury, tak i u predikované.

Náboje se počítaly pomocí nábojové metody EQeq, která je součástí jak MACHu, tak ACC II. softwaru, viz část 3.4.

Byly tedy porovnány struktury získané z nástrojů PyMOL, HAAD, Reduce, OpenBabel a Chimera. Z toho měl OpenBabel nejlepší korelační koeficient, ale bylo možné porovnat pouze 3 struktury získané z tohoto nástroje, protože ostatní napočítané struktury neměly stejný počet struktur jako v experimentálně určené struktuře. Druhým nástrojem, kde již bylo porovnáno 326 molekul, je HAAD s R^2 = 0,9983. Třetí nejvyšší hodnotu, R^2 = 0,9919, dosáhla Chimera, ale zde se porovnávalo jen 78 struktur. Reduce dosáhl hodnoty R^2 = 0,9726 pro 354 molekul. PyMOL měl nejnižší hodnotu a to R^2 = 0,9686.

V tabulce 4.6 jsou všechny hodnoty získané v rámci porovnání nábojů.

Z tabulky vyplývá, že přidáním vodíků za pomoci nástrojů, byly nejvíce ovlivněny parciální náboje u atomu kyslíku. Hodnoty PCC pro kyslík jsou u PyMOLu a HAAD -0,0810 a 0,0435. U Reduce a Chimery je sice o něco větší hodnota (0,1340 a 0,2827), nicméně i zde je patrné, že atom kyslíku byl nejvíce ovlivněn. U nástroje OpenBabel nelze určit, který atom byl nejvíce ovlivněn.

Dále ještě byly často ovlivněny náboje atomů vodíků u nástrojů PyMOL a Reduce. Posledním atomem, jenž byl, i když méně, ovlivněn predikcí vodíků, je dusík a to při použití softwarů PyMOL, Reduce a Chimera.

Celkově můžeme konstatovat, že všechny použité nástroje doplňují struktury tak, že výsledné hodnoty nábojů jsou podobné s těmi napočítanými na původních strukturách.

4.3 Diskuse

Z výše popsaných 6 nástrojů, které byly v rámci této práce testovány, nejlépe dopadla trojice PyMOL, Reduce a HAAD.

Přestože Reduce v rychlosti skončil na posledním místě, řádově se jedná o minutové intervaly a dokázal napočítat všechny druhy struktur bez problému. Ve srovnání přesnosti určení vodíků vůči původním drží Reduce největší počty jak ve srovnání se stejným počtem vodíků, tak i v nestejném počtu vodíků. Průměrné hodnoty RMSD měl u obou srovnání zároveň nejnižší a směrodatné odchylky měl jako druhé nejnižší ze zmíněné trojice. Výhoda Reduce je i ta, že atomy vodíků pojmenovává dle konvencí IUPAC nomenklatury pro standardní aminokyseliny a nukleotidy [99]. Jedinou nevýhodou je, že všechny přidané vodíky mají atomové číslo nula.

PyMOL je nejrychlejším nástrojem, jelikož zvládl napočítat všechny struktury za necelou minutu. Nicméně v testu přesnosti určení vodíků se stejným počtem atomů vodíků skončil vždy jako třetí ze zmíněné trojice. Ovšem hodnoty RMSD jsou blízké hodnotám napočítaným pro Reduce. Na druhou stranu PyMOL měl jako druhý nejvíce struktur na porovnání s nestejným počtem vodíků a se stejným počtem těchto atomů byl za nástrojem HAAD. I přestože PyMOL používá asi nejjednodušší algoritmus ze všech, je tento nástroj velmi účinný. Nicméně i když přidává k přidaným vodíkům jejich atomové číslo, tak názvy doplňuje podle svého způsobu. Tento způsob zahrnuje znak vodíku (H) a číslo jako pořadí daného vodíku vždy s dvoučíselnou reprezentací. Příkladem je tedy H01, H02 a tak podobně.

Třetí ze zmíněné trojice je HAAD, jenž žádným způsobem nenapočítal 5 struktur, ale v rychlosti se umístil na druhém místě s časem něco málo přes minutu. HAAD si vedl

Nástroj	Počet napočítaných molekul	R^2	Atom	MAE	ABSMAX	RMSD	PCC
PyMOL	203	0,9686	H	0,0904	0,5139	0,1433	0,1924
			O	0,1094	0,4894	0,1760	-0,0810
			N	0,1139	0,4555	0,1584	0,2387
			S	0,0519	0,2082	0,0270	0,9288
			C	0,2612	0,6925	0,1602	0,8779
			P	0,0299	0,1166	0,0125	0,9833
HAAD	326	0,9983	H	0,0690	0,5143	0,0174	0,9869
			O	0,0317	0,3028	0,0544	0,0435
			N	0,0713	0,3016	0,0213	0,9858
			S	0,0768	0,2216	0,0353	0,9453
			C	0,3346	0,8517	0,0215	0,9992
Reduce	354	0,9726	H	0,0825	0,6138	0,1265	0,3210
			O	0,0895	0,4897	0,1375	0,1340
			N	0,1097	0,5544	0,1380	0,6082
			S	0,0618	0,1818	0,0215	0,9714
			C	0,3327	0,7433	0,1099	0,9658
			P	0,0267	0,1014	0,0096	0,9902
OpenBabel	3	0,9996	H	0,0652	0,1316	0,0128	0,9903
			O	0,0310	0,0920	0,0130	0,9379
			N	0,0399	0,0295	0,0061	0,9975
			S	0,0229	0,0022	0,0013	0,9998
			C	0,3152	0,0571	0,0095	0,9998
Chimera	78	0,9919	H	0,0735	0,2644	0,0129	0,9909
			O	0,0674	0,4561	0,1065	0,2827
			N	0,0833	0,4117	0,0843	0,7723
			S	0,0647	0,1438	0,0257	0,9571
			C	0,3276	0,6000	0,0701	0,9843
			P	0,0281	0,0779	0,0087	0,9918

Tabulka 4.6: Srovnání původních nábojů s náboji, které byly získány po určení poloh vodíků.

dobře i v testu přesnosti, kdy v počtu struktur se stejným počtem atomů jako v původní struktuře byl druhý a měl největší zastoupení. U obou způsobů testování v přesnosti měl jako druhý nejmenší průměrnou hodnotu RMSD a zároveň hodnotu směrodatné odchylky měl nejmenší ze zmíněných tří nástrojů. Velká nevýhoda nástroje HAAD je ale ta, že výsledek své operace ukládá jako soubor s příponou .h, jenž je velmi podobný souborovému formátu PDB. Tento soubor však není zcela validní, jelikož zde chybí symbol prvku a místo něho je zde pouze číslo 1 u přidaných vodíků (u původních atomů je na tomto místě prázdno). Dále u přidaných vodíků chybí atomové číslo, respektive je nastavena nula, a znak identifikace řetězce je prázdný. Pro označení přidaných atomů vodíků se používají předefinované názvy. Podle residua a umístění vodíku v residuu je stanoven název. Ten je podobný IUPAC nomenklatuře, jen s tím rozdílem, že je poslední číslo přesunuto na začátek. Dalším nedostatkem je, že HAAD může přidat stejné souřadnice dvěma a více predikovaným vodíkům, ukončuje své výpočty po skončení prvního řetězce (nedokáže napočítat všechny řetězce ve struktuře) a neumožňuje počítat nukleotidy, respektive všechny nukleotidy reprezentuje jako jeden glycin.

Co se týče nábojů, tak nebyl odhalen žádný nástroj, který by se nějak radikálně odlišoval. Jediný větší rozdíl v nábojích způsobuje HAAD, kdy nejvíce ovlivňuje jen atomy kyslíku. Ostatní nástroje ovlivňují podobným způsobem náboje atomů vodíku, kyslíku a dusíku. Nicméně celkové hodnoty korelace s původními náboji jsou u všech nástrojů dost podobné.

U všech tří nejlepších nástrojů je patrné, že největší problém způsobovala aminokyselina histidin, pro niž se ve většině případů napočítalo méně vodíků, než bylo v původní struktuře. Důvodem je, že histidin je citlivý na změny pH.

Poslední nástroj, který nějakým způsobem predikoval vodíky pro všechny struktury, je OpenBabel. Nicméně kvalitativně jich napočítal pouze 192 a z toho jen 3 se stejným počtem vodíků. Mnoho molekul bylo u tohoto nástroje vynecháno, protože OpenBabel přidává atomy vodíků vždycky jako vodíky patřící ke standardním aminokyselinám nebo nukleotidům. Místo toho, aby je správně označil jako patřící k nepolymerním nebo jiným „nestandardním“ chemickým látkám. Jinými slovy, atomy vodíků byly označeny jako ATOM, místo aby byly jako HETATM. Další problém u OpenBabelu je, že predikované vodíky označuje jednoznačně pouze jednopísmenným označením H a přidává je hromadně na konec. Nicméně zachovává propojení mezi vodíky a ostatními atomy. Co se týče rychlosti umístil se jako třetí ze čtyř nástrojů, které napočítaly téměř všechny struktury. Ohledně přesnosti je možné brát v potaz pouze hodnocení v rámci nestejného počtu vodíků, kde měl největší průměrnou hodnotu RMSD a zároveň i směrodatnou odchylku této hodnoty. Nejvíce zajímavé na tomto nástroji je to, že poměr mezi strukturami, kde bylo predikováno méně vodíků, a strukturami, jež měly predikováno více atomů vodíků, je téměř 50 na 50. Zároveň se nedá určit jen jeden typ residua, který by dělal největší problém. U všech ostatních nástrojů byl většinou problém, že určily méně vodíků než bylo v původní struktuře a bylo patrné, že hlavně jedno residuum dělá problém. OpenBabel je velmi dobrý nástroj na převod z jednoho souborového typu do jiného. Nicméně na určení vodíků je vhodnější použít jiný nástroj.

Ostatní nástroje není možné hodnotit stejným způsobem, jelikož napočítaných struktur bylo velmi málo. Z tohoto důvodu není možné porovnat rychlost s ostatními nástroji.

Chimera z 2199 struktur predikovala 292. Důvodem je, že Chimera je mnohem citlivější na chybu, nejednoznačnost ve struktuře molekuly. Příkladem můžou být jednopísmenné

označení residuí, záporné číslování atomového čísla a struktury obsahující pouze HETATM. V porovnání přesnosti si tento nástroj vedl lépe než OpenBabel a to jak při stejném počtu vodíků tak i v odlišném počtu. Co se týče residua s největším počtem chybně predikovaných vodíků, jde u Chimery nejčastěji o histidin, stejně jako u trojice nejlepších. Tento nástroj dodržuje IUPAC nomenklaturu u jmen vodíků. Narozdíl od Reduce, ale správně označuje atomové číslo jednotlivým atomům vodíků podle pořadí ve struktuře. Chimeru tedy nelze jednoznačně označit za vhodný nebo nevhodný nástroj na určení poloh vodíků. Nicméně pro jeho porovnání s dalšími nástroji je potřeba vytvořit mnohem specifičtější dataset, který nebude mít tak velkou různorodost struktur.

Posledním nástrojem, který byl zahrnut do porovnání, je GROMACS. U tohoto nástroje se porovnávaly predikce poloh vodíků za využití tří silových polí. Jednalo se vždy o nejnovější verze těchto polí, jmenovitě GROMOS jako GROMOS96 (přesněji jeho verze 54a7), CHARMM s verzí CHARMM27 a AMBER jako AMBER99SB-ILDN. U všech polí se poté využil model vody TIP3P TIP3, jenž je nástrojem doporučován. Jelikož GROMACS predikoval pokaždé nejmenší počet struktur, tak nebylo shledáno nutným, aby se testovaly ostatní verze těchto silových polí. Důvodem byl i to, že pro lepší porovnání tohoto nástroje by bylo nutné vytvořit podobný dataset jako v případě Chimery. GROMACS totiž, kromě stejných případů jako Chimera, nepredikuje polohy vodíků ve strukturách, kde se nalézá HETATM ve struktuře molekuly. Kvůli tomu má GROMACS pro porovnání pouze malý počet struktur a nemohlo se provést porovnání hodnot parciálních atomových nábojů. I když se však nedá GROMACS porovnat s ostatními 5 nástroji, je možné tento nástroj porovnat v rámci použitých silových polí. Z testovaných silových polí si nejlépe vedlo pole AMBER, jež mělo největší počet struktur. Na druhou stranu pole CHARMM mělo jen o pár struktur méně a výsledky v porovnání přesnosti a rychlosti mělo dost shodné s polem AMBER. Jako méně vhodné pole pro použitý dataset se ukázal GROMOS, jenž sice byl nejrychlejší z použitých silových polí, ale pro porovnání přesnosti se ukázalo použitelných jen 6 struktur. Co se týče nejčastějšího residua, kde bylo predikováno méně poloh vodíků, ukázalo se, že u nástroje GROMACS všechna pole měla nejvíce problém s aminokyselinou leucin. CHARMM a GROMOS u žádného residua neurčily více poloh vodíků než v původní struktuře. U AMBER se takto stalo pouze u všech případů lysinu. Nicméně jednalo se jen o 2 z 2121 případů nesprávně určeného počtu vodíků v residuu. Závěrem tedy i u GROMACS nelze jednoznačně určit zda je vhodný na predikci polohy atomu vodíků nebo není. Pro dané zjištění je zapotřebí zvolit dataset s většími specifikacemi ve výběru.

4.3.1 Možnost využití v rámci ACC II

Pro nasazení do nástroje ACC II nelze doporučit Chimeru, GROMACS a OpenBabel. Jelikož při použitém datasetu si tyto nástroje nevedly dobře. Dále nelze doporučit HAAD, jenž má větší množství nedostatků oproti PyMOLu a Reduce. Výběr mezi PyMOLem a Reduce však není jednoznačný. PyMOL sice není tak přesný jako Reduce, ale zato je rychlejší na výpočet velkého množství molekul. Navíc umožňuje vytváření dalších pluginů, jež mohou zlepšit přesnost predikce. Nebo ho lze použit přímo v Python programu. Proto bude záležet na požadavcích implementace nástroje na predikci vodíků v softwaru ACC II.

Závěr

V této diplomové práci je popsán princip fungování 15 nástrojů na určení souřadnic vodíků v proteinových strukturách. Z celkového počtu nástrojů bylo možné na základě definovaných kritérií porovnat mezi sebou 6 z nich, jmenovitě PyMOL, HAAD, Reduce, OpenBabel, Chimera a GROMACS. V rámci práce byla porovnávána rychlost výpočtu a dále přesnost vypočítaných souřadnic atomů vodíků vzhledem k souřadnicím získaným experimentálně pomocí metody NMR. Jako třetí aspekt bylo porovnáno ovlivnění parciálních atomových nábojů u struktury, kde byly vodíky určeny softwarovými nástroji vůči nábojům, kde byly vodíky určeny experimentálně.

Pro porovnání byla použita vytvořená datová sada, jež v sobě zahrnovala náhodně vybrané struktury molekul. Tato sada obsahovala 2199 struktur určených pomocí NMR. Všechny struktury byly získány z RCSD PDB.

Z měření, jak rychle nástroje dokáží vypočítat polohy vodíků, nejlépe vyšel PyMOL. Jako druhý nejrychlejší nástroj se umístil HAAD. Oproti tomu, nejpomalejším nástrojem byl Reduce.

Nicméně Reduce se ukázal jako nejlepší nástroj během porovnání přesnosti, kde dosahuje nejmenších hodnot RMSD vzhledem k počtu úspěšně zpracovaných struktur. PyMOL a HAAD dosahovaly jen mírně vyšších hodnot RMSD.

Při porovnávání nábojů pomocí Pearsonova korelačního koeficientu se ukázalo, že struktury s vodíky dopočítanými pomocí nástrojů PyMOL, HAAD, Reduce, OpenBabel a Chimera dosahují podobných hodnot parciálních atomových nábojů jako struktury, získané experimentálně. Z tohoto porovnání byl však vynechán GROMACS, jelikož nebyly získány žádné struktury vhodné pro výpočet parciálních nábojů.

V poslední fázi práce pak byly analyzovány nedostatky všech použitých nástrojů. Největší nedostatek měl HAAD, jelikož výsledek své operace ukládá jako soubor s příponou .h. Tento soubor však není validní formou souborového formátu PDB.

Z výsledků všech uvedených analýz vyplývá, že nejvhodnějšími nástroji jsou PyMOL a Reduce. Proto jsou právě tyto dva nástroje vhodné pro integraci do softwarového nástroje ACC II.

Seznam použité literatury

[1] Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic acids research*. 2019, **47**(D1), D520–D528.

[2] GEIDL, Stanislav; BOUCHAL, Tomáš; RAČEK, Tomáš; VAŘEKOVÁ, Radka Svobodová; HEJRET, Václav; KŘENEK, Aleš; ABAGYAN, Ruben; KOČA, Jaroslav. High-quality and universal empirical atomic charges for chemoinformatics applications. *Journal of cheminformatics*. 2015, **7**(1), 1–10.

[3] SCHINDLER, Ondřej. *Empirické metody pro výpočet parciálních atomových nábojů [online]*. 2019 [cit. 2021-05-13]. Dostupné také z: https://is.muni.cz/th/ts2kg/. Diplomová práce. Masarykova univerzita, Přírodovědecká fakulta, Brno. Vedoucí práce Tomáš RAČEK; Radka SVOBODOVÁ.

[4] RAČEK, Tomáš; SCHINDLER, Ondřej; TOUŠEK, Dominik; HORSKÝ, Vladimír; BERKA, Karel; KOČA, Jaroslav; SVOBODOVÁ, Radka. Atomic Charge Calculator II: web-based tool for the calculation of partial atomic charges. *Nucleic acids research*. 2020, **48**(W1), W591–W596. Dostupné z DOI: 10.1093/nar/gkaa367.

[5] LI, Yunqi; ROY, Ambrish; ZHANG, Yang. HAAD: a quick algorithm for accurate prediction of hydrogen atoms in protein structures. *PloS one*. 2009, **4**(8), e6701.

[6] FORREST, Lucy R; HONIG, Barry. An assessment of the accuracy of methods for predicting hydrogen positions in protein structures. *Proteins: Structure, Function, and Bioinformatics*. 2005, **61**(2), 296–309.

[7] WOIŃSKA, Magdalena; GRABOWSKY, Simon; DOMINIAK, Paulina M; WOŹNIAK, Krzysztof; JAYATILAKA, Dylan. Hydrogen atoms can be located accurately and precisely by x-ray crystallography. *Science advances*. 2016, **2**(5), e1600192.

[8] *Hydrogen Bonding — Introduction to Chemistry*. [Online] [cit. 2021-03-27]. Dostupné z: https://courses.lumenlearning.com/introchem/chapter/hydrogen-bonding/.

[9] SILVERMAN, Richard B. Chapter 3 - Receptors. In: SILVERMAN, Richard B. (ed.). *The Organic Chemistry of Drug Design and Drug Action (Second Edition)*. Second Edition. San Diego: Academic Press, 2004, s. 121–172l. ISBN 978-0-12-643732-4. Dostupné z DOI: https://doi.org/10.1016/B978-0-08-051337-9.50008-0.

[10] SZALEWICZ, Krzysztof. Hydrogen Bond. In: MEYERS, Robert A. (ed.). *Encyclopedia of Physical Science and Technology (Third Edition)*. Third Edition. New York: Academic Press, 2003, s. 505–538. ISBN 978-0-12-227410-7. Dostupné z DOI: https://doi.org/10.1016/B0-12-227410-5/00322-7.

[11] JR, Mark A. Herzik. Cryo-electron microscopy reaches atomic resolution. *Nature* [online]. 2020, **587**, 39–40 [cit. 2021-05-06]. ISSN 1476-4687. Dostupné z DOI: https://doi.org/10.1038/d41586-020-02924-y.

[12] FERRANDO, Riccardo. Synthesis and experimental characterization of nanoalloy structures. In: *Frontiers of Nanoscience*. Elsevier, 2016, sv. 10, s. 47–74.

[13] CHIRAYIL, Cintil Jose; ABRAHAM, Jiji; MISHRA, Raghvendra Kumar; GEORGE, Soney C; THOMAS, Sabu. Instrumental techniques for the characterization of nanoparticles. In: *Thermal and Rheological Measurement Techniques for Nanomaterials Characterization*. Elsevier, 2017, s. 1–36.

[14] ARYAL, Sagar. *Nuclear Magnetic Resonance (NMR) Spectroscopy — Instrumentation — Microbe Notes* [online]. 2021 [cit. 2021-05-05]. Dostupné z: https://microbenotes.com/nuclear-magnetic-resonance-nmr-spectroscopy/.

[15] ŽÍDEK, L et al. *Strukturní biochemie-Skripta k přednášce C9530 Přírodovědecké fakulty Masarykovy univerzity*. Brno, 2014.

[16] DOERR, Allison. Cryo-electron tomography. *Nature Methods*. 2017, **14**(1), 34–34.

[17] YIP, Ka Man; FISCHER, Niels; PAKNIA, Elham; CHARI, Ashwin; STARK, Holger. Atomic-resolution protein structure determination by cryo-EM. *Nature*. 2020, **587**(7832), 157–161.

[18] CRESSEY, Daniel; CALLAWAY, Ewen. Cryo-electron microscopy wins chemistry Nobel. *Nature* [online]. 2017, **550**, 167 [cit. 2021-05-06]. ISSN 0028-0836. Dostupné z DOI: doi:10.1038/nature.2017.22738.

[19] BHELLA, David. Cryo-electron microscopy: an introduction to the technique, and considerations when working to establish a national facility. *Biophysical reviews*. 2019, **11**(4), 515–519.

[20] *Better than Crystal Clear? Cryo-EM Achieves Atomic Resolution* [online]. 2020 [cit. 2021-05-06]. Dostupné z: https://www.alzforum.org/news/research-news/better-crystal-clear-cryo-em-achieves-atomic-resolution.

[21] SMYTH, MS; MARTIN, JHJ. x Ray crystallography. *Molecular Pathology*. 2000, **53**(1), 8.

[22] CALLAWAY, Ewen. Revolutionary cryo-EM is taking over structural biology. *Nature*. 2020, **578**(7794), 201.

[23] *Hydrogens in PDB Files*. [Online]. Amherst: University of Massachusetts Amherst, ©2021 [cit. 2021-03-26]. Dostupné z: https://www.umass.edu/microbio/chime/pe_beta/pe/protexpl/help_hyd.htm.

[24] PAULING, Linus. The nature of the chemical bond. Application of results obtained from the quantum mechanics and from a theory of paramagnetic susceptibility to the structure of molecules. *Journal of the American Chemical Society*. 1931, **53**(4), 1367–1400.

[25] GUND, TAMARA. 3 - Molecular Modeling of Small Molecules. In: COHEN, N. CLAUDE (ed.). *Guidebook on Molecular Modeling in Drug Design*. San Diego: Academic Press, 1996, s. 55–92. ISBN 978-0-12-178245-0. Dostupné z DOI: https://doi.org/10.1016/B978-012178245-0/50004-4.

[26] DESCHAMPS, Jeffrey R.; FLIPPEN-ANDERSON, Judith L. Crystallography. In: MEYERS, Robert A. (ed.). *Encyclopedia of Physical Science and Technology (Third Edition)*. Third Edition. New York: Academic Press, 2002, s. 121–153. ISBN 978-0-12-227410-7. Dostupné z DOI: https://doi.org/10.1016/B0-12-227410-5/00160-5.

[27] WANG, Bo; TRUHLAR, Donald G. Partial atomic charges and screened charge models of the electrostatic potential. *Journal of chemical theory and computation*. 2012, **8**(6), 1989–1998.

[28] MCCAMMON, JA; STRAATSMA, TP. Computational alchemy. *Annu. Rev. Phys. Chem*. 1992, **43**, 407.

[29] LEACH, Andrew R; GILLET, Valerie J. *An introduction to chemoinformatics*. Springer, 2007.

[30] JURS, Peter C. Quantitative structure-property relationships. *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes*. 2003, 1314–1335.

[31] LEACH, Andrew R; LEACH, Andrew R. *Molecular modelling: principles and applications*. Pearson education, 2001.

[32] CHAVES, J; BARROSO, Joan M; BULTINCK, Patrick; CARBÓ-DORCA, Ramon. Toward an alternative hardness kernel matrix structure in the Electronegativity Equalization Method (EEM). *Journal of chemical information and modeling*. 2006, **46**(4), 1657–1665.

[33] RAPPE, Anthony K; GODDARD III, William A. Charge equilibration for molecular dynamics simulations. *The Journal of Physical Chemistry*. 1991, **95**(8), 3358–3363.

[34] NISTOR, Razvan A; POLIHRONOV, Jeliazko G; MÜSER, Martin H; MOSEY, Nicholas J. A generalization of the charge equilibration method for nonmetallic materials. *The Journal of chemical physics*. 2006, **125**(9), 094108.

[35] BERMAN, Helen M; WESTBROOK, John; FENG, Zukang; GILLILAND, Gary; BHAT, Talapady N; WEISSIG, Helge; SHINDYALOV, Ilya N; BOURNE, Philip E. The protein data bank. *Nucleic acids research*. 2000, **28**(1), 235–242.

[36] MIR, Saqib; ALHROUB, Younes; ANYANGO, Stephen; ARMSTRONG, David R; BERRISFORD, John M; CLARK, Alice R; CONROY, Matthew J; DANA, Jose M; DESHPANDE, Mandar; GUPTA, Deepti et al. PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic acids research*. 2018, **46**(D1), D486–D492.

[37] KINJO, Akira R; BEKKER, Gert-Jan; SUZUKI, Hirofumi; TSUCHIYA, Yuko; KAWABATA, Takeshi; IKEGAWA, Yasuyo; NAKAMURA, Haruki. Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic acids research*. 2016, gkw962.

[38] ULRICH, Eldon L; AKUTSU, Hideo; DORELEIJERS, Jurgen F; HARANO, Yoko; IOANNIDIS, Yannis E; LIN, Jundong; LIVNY, Miron; MADING, Steve; MAZIUK, Dimitri; MILLER, Zachary et al. BioMagResBank. *Nucleic acids research*. 2007, **36**(suppl_1), D402–D408.

[39] *PDB-101: Learn: Guide to Understanding PDB Data: Introduction* [online]. RCSB PDB [cit. 2021-03-26]. Dostupné z: http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction.

[40] WILLMOTT, Cort J; MATSUURA, Kenji. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*. 2005, **30**(1), 79–82.

[41] ZOUVALA, Šimon. *Implementace metod MGC a OGC pro výpočet parciálních atomových nábojů [online]*. 2018 [cit. 2021-05-14]. Dostupné také z: https://is.muni.cz/th/jb4h2/.

[42] ŠREJBER, Martin. *Srovnání silových polí pro studium biologických membrán [online]*. 2014 [cit. 2021-03-21]. Dostupné také z: https://theses.cz/id/ptghxd/.

[43] ADURI, Raviprasad; PSCIUK, Brian T; SARO, Pirro; TANIGA, Hariprakash; SCHLEGEL, H Bernhard; SANTALUCIA, John. AMBER force field parameters for the naturally occurring modified nucleosides in RNA. *Journal of Chemical Theory and Computation*. 2007, **3**(4), 1464–1475.

[44] PÉREZ, Alberto; MARCHÁN, Iván; SVOZIL, Daniel; SPONER, Jiri; CHEATHAM III, Thomas E; LAUGHTON, Charles A; OROZCO, Modesto. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical journal*. 2007, **92**(11), 3817–3829.

[45] CASE, David A; BELFON, Kellon; BEN-SHALOM, Ido; BROZELL, Scott R; CERUTTI, David; CHEATHAM, Thomas; CRUZEIRO, Vinícius Wilian D; DARDEN, Tom; DUKE, Robert E; GIAMBASU, George et al. Amber 2020. 2020.

[46] VANOMMESLAEGHE, Kenno; HATCHER, Elizabeth; ACHARYA, Chayan; KUNDU, Sibsankar; ZHONG, Shijun; SHIM, Jihyun; DARIAN, Eva; GUVENCH, Olgun; LOPES, P; VOROBYOV, Igor et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry*. 2010, **31**(4), 671–690.

[47] *MacKerell Lab* [online]. Univesity of Meryland [cit. 2021-03-25]. Dostupné z: http://mackerell.umaryland.edu/charmm_ff.shtml.

[48] MACKERELL JR, Alex D; BASHFORD, Donald; BELLOTT, MLDR; DUNBRACK JR, Roland Leslie; EVANSECK, Jeffrey D; FIELD, Martin J; FISCHER, Stefan; GAO, Jiali; GUO, H; HA, Sookhee et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*. 1998, **102**(18), 3586–3616.

[49] REIHER III, WH. Theoretical studies of hydrogen bonding. *Ph. D. Thesis at Harvard University*. 1985.

[50] CHEN, Jianhan; IM, Wonpil; BROOKS, Charles L. Balancing solvation and intramolecular interactions: toward a consistent generalized Born force field. *Journal of the American Chemical Society*. 2006, **128**(11), 3728–3736.

[51] MACKERELL JR, Alexander D; BANAVALI, Nilesh; FOLOPPE, Nicolas. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers: Original Research on Biomolecules*. 2000, **56**(4), 257–265.

[52] OOSTENBRINK, Chris; VILLA, Alessandra; MARK, Alan E; VAN GUNSTEREN, Wilfred F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry*. 2004, **25**(13), 1656–1676.

[53] *GROMACS - Gromacs*. 2020. Dostupné také z: http://www.gromacs.org/.

[54] WILMER, Christopher E; KIM, Ki Chul; SNURR, Randall Q. An extended charge equilibration method. *The journal of physical chemistry letters*. 2012, **3**(17), 2506–2511.

[55] ODA, Akifumi; HIRONO, Shuichi. Geometry-dependent atomic charge calculations using charge equilibration method with empirical two-center Coulombic terms. *Journal of Molecular Structure: THEOCHEM*. 2003, **634**(1-3), 159–170.

[56] PETTERSEN, Eric F; GODDARD, Thomas D; HUANG, Conrad C; COUCH, Gregory S; GREENBLATT, Daniel M; MENG, Elaine C; FERRIN, Thomas E. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*. 2004, **25**(13), 1605–1612.

[57] *UCSF Chimera Home Page*. [Online]. ©2018 [cit. 2021-02-28]. Dostupné z: https://www.cgl.ucsf.edu/chimera/.

[58] SCHRÖDINGER, LLC. *The PyMOL Molecular Graphics System, Version 1.8*. 2015.

[59] YUAN, Shuguang; CHAN, H.; HU, Zhenquan. Using PyMOL as a platform for computational drug design. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 2017, **7**, e1298. Dostupné z DOI: 10.1002/wcms.1298.

[60] *PyMOL — pymol.org*. [Online]. Schrodinger., © 2021 [cit. 2021-03-01]. Dostupné z: https://pymol.org/2/.

[61] *PyMOLWiki* [online] [cit. 2021-03-06]. Dostupné z: https://pymolwiki.org/index.php/Main_Page.

[62] *HAAD: predict hydrogen atoms in protein structures*. [Online] [cit. 2021-02-25]. Dostupné z: https://zhanglab.ccmb.med.umich.edu/HAAD/.

[63] *Reduce Software : Kinemage Website*. © 2000. Dostupné také z: http://kinemage.biochem.duke.edu/software/reduce.php.

[64] WORD, J Michael; LOVELL, Simon C; RICHARDSON, Jane S; RICHARDSON, David C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology*. 1999, **285**(4), 1735–1747.

[65] *Reduce Guide - MolProbity Help.* [Online] [cit. 2021-03-02]. Dostupné z: http://molprobity.biochem.duke.edu/help/reduce_guide/reduce_guide.html.

[66] WILLIAMS, Christopher J; HEADD, Jeffrey J; MORIARTY, Nigel W; PRISANT, Michael G; VIDEAU, Lizbeth L; DEIS, Lindsay N; VERMA, Vishal; KEEDY, Daniel A; HINTZE, Bradley J; CHEN, Vincent B et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science*. 2018, **27**(1), 293–315.

[67] *Main page - MolProbity* [online]. Department of Biochemistry, Duke University [cit. 2021-03-02]. Dostupné z: http://molprobity.biochem.duke.edu/.

[68] *Phenix.reduce: tool for adding hydrogens to a PDB model* [online] [cit. 2021-03-02]. Dostupné z: https://www.phenix-online.org/documentation/reference/hydrogens.html.

[69] *Open Babel:About - Open Babel* [online]. 2009 [cit. 2021-03-04]. Dostupné z: http://openbabel.org/wiki/Open_Babel:About.

[70] HUTCHISON, Geoffrey R; MORLEY, Chris; JAMES, Craig; SWAIN, Chris; DE WINTER, Hans; VANDERMEERSCH, Tim; O'BOYLE, Noel M. Open Babel Documentation. 2011.

[71] O'BOYLE, Noel; BANCK, Michael; JAMES, Craig; MORLEY, Chris; VANDERMEERSCH, Tim; HUTCHISON, Geoffrey. Open Babel: An Open Chemical Toolbox. *Journal of cheminformatics*. 2011, **3**, 33. Dostupné z DOI: 10.1186/1758-2946-3-33.

[72] ABRAHAM, Mark James; MURTOLA, Teemu; SCHULZ, Roland; PÁLL, Szilárd; SMITH, Jeremy C.; HESS, Berk; LINDAHL, Erik. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 2015, **1-2**, 19–25. ISSN 2352-7110. Dostupné z DOI: https://doi.org/10.1016/j.softx.2015.06.001.

[73] LINDAHL, Erik; HESS, Berk; VAN DER SPOEL, David. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual*. 2001, **7**(8), 306–317.

[74] LINDAHL; ABRAHAM; HESS; SPOEL, van der. *GROMACS 2019.3 Manual*. Zenodo, 2019. Ver. 2019.3. Dostupné z DOI: 10.5281/zenodo.3243834.

[75] VAN DER SPOEL, David; LINDAHL, Erik; HESS, Berk; GROENHOF, Gerrit; MARK, Alan E; BERENDSEN, Herman JC. GROMACS: fast, flexible, and free. *Journal of computational chemistry*. 2005, **26**(16), 1701–1718.

[76] BRÜNGER, Axel T; ADAMS, Paul D; CLORE, G Marius; DELANO, Warren L; GROS, Piet; GROSSE-KUNSTLEVE, Ralf W; JIANG, Jian-Sheng; KUSZEWSKI, John; NILGES, Michael; PANNU, Navraj S et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D*. 1998, **54**(Pt 5), 905–921.

[77] *Crystallography & NMR System* [online]. Yale University, 1997-2010 [cit. 2021-03-19]. Dostupné z: http://cns-online.org/v1.3/.

[78] *The Amber Molecular Dynamics Package.* [Online]. 2021 [cit. 2021-03-21]. Dostupné z: http://ambermd.org/.

[79] SALOMON-FERRER, Romelia; CASE, David A; WALKER, Ross C. An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 2013, **3**(2), 198–210.

[80] *CORINA Classic - High-Quality 3D Molecular Models — MN-AM.* [Online]. Molecular Networks GmbH, Germany a Altamira, LLC, USA., © 2021 [cit. 2021-03-18]. Dostupné z: https://www.mn-am.com/products/corina.

[81] CORINA Classic Generation of Three-Dimensional Molecular Models: Version 4.3.0 Program Description. In: [online]. Molecular Networks GmbH, 2019, s. 162 [cit. 2021-03-18]. Dostupné z: https://www.mn-am.com/files/docs/corina/corina_classic_manual.pdf.

[82] SADOWSKI, Jens; GASTEIGER, Johann. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chemical Reviews*. 1993, **93**(7), 2567–2581.

[83] CAI, Xiuhong; HAIDER, Kamran; LU, Jianxun; RADIC, Slaven; SON, Chang Yun; CUI, Qiang; GUNNER, MR. Network analysis of a proposed exit pathway for protons to the P-side of cytochrome c oxidase. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*. 2018, **1859**(10), 997–1005.

[84] *MCCE Documentation.* [Online]. Gunner Lab, ©2020 [cit. 2021-03-14]. Dostupné z: https://gunnerlab.github.io/Stable-MCCE/.

[85] HOOFT, Rob WW; SANDER, Chris; VRIEND, Gerrit. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins: Structure, Function, and Bioinformatics*. 1996, **26**(4), 363–376.

[86] *WHAT IF homepage* [online] [cit. 2021-03-13]. Dostupné z: https://swift.cmbi.umcn.nl/whatif/.

[87] *YASARA - Yet Another Scientific Artificial Reality Application.* [Online]. ©1993-2018 [cit. 2021-03-14]. Dostupné z: http://www.yasara.org/index.html.

[88] *Home - CHARMM* [online]. ©2020 [cit. 2021-03-14]. Dostupné z: https://www.charmm.org/charmm/.

[89] BROOKS, Bernard R; BRUCCOLERI, Robert E; OLAFSON, Barry D; STATES, David J; SWAMINATHAN, S a; KARPLUS, Martin. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*. 1983, **4**(2), 187–217.

[90] BROOKS, Bernard R; BROOKS III, Charles L; MACKERELL JR, Alexander D; NILSSON, Lennart; PETRELLA, Robert J; ROUX, Benoît; WON, Youngdo; ARCHONTIS, Georgios; BARTELS, Christian; BORESCH, Stefan et al. CHARMM: the biomolecular simulation program. *Journal of computational chemistry*. 2009, **30**(10), 1545–1614.

[91] ANANDAKRISHNAN, Ramu; AGUILAR, Boris; ONUFRIEV, Alexey V. H++ 3.0: automating p K prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research*. 2012, **40**(W1), W537–W541.

[92] *H++ (web-based computational prediction of protonation states and pK of ionizable groups in macromolecules)* [online]. Virginia Tech [cit. 2021-03-11]. Dostupné z: http://biophysics.cs.vt.edu/index.php.

[93] *DeepView–The Swiss-PdbViewer User Guide*. Geneva, 13.09.2001. Dostupné také z: https://spdbv.vital-it.ch/Swiss-PdbViewerManualv3.7.pdf.

[94] *Swiss PDB Viewer - Home* [online]. Swiss Institute of Bioinformatics [cit. 2021-03-13]. Dostupné z: https://spdbv.vital-it.ch/.

[95] *Align - PyMOLWiki.* [Online]. 2009 [cit. 2021-03-01]. Dostupné z: https://pymolwiki.org/index.php/Align.

[96] BERMAN, Helen; HENRICK, Kim; NAKAMURA, Haruki. Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*. 2003, **10**(12), 980–980.

[97] *Ubuntu Manpage: time - změří čas vykonání příkazu nebo vypíše využití zdrojů* [online]. Canonical Ltd., ©2019 [cit. 2021-03-12]. Dostupné z: http://manpages.ubuntu.com/manpages/precise/cs/man1/time.1.html.

[98] WHITFORD, David. *Proteins: Structure and Function.* 1. vydání. United States: John Wiley & Sons Inc, 2005. ISBN 978-0471498940.

[99] MARKLEY, John L; BAX, Ad; ARATA, Yoji; HILBERS, CW; KAPTEIN, Robert; SYKES, Brian D; WRIGHT, Peter E; WÜTHRICH, Kurt. Recommendations for the presentation of NMR structures of proteins and nucleic acids–IUPAC-IUBMB-IUPAB Inter-Union Task Group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. *Journal of biomolecular NMR*. 1998, **12**(1), 1–23.

Příloha A

Číselné parametry pro spuštění nástroje Gromacs

Číselná hodnota	Model vody
1	TIP3P TIP 3-point, recommended
2	TIP4P TIP 4-point
3	TIP4P-Ew TIP 4-point optimized with Ewald
4	TIP5P TIP 5-point
5	SPC simple point charge
6	SPC/E extended simple point charge
7	None

Tabulka A.1: Číselný parametr pro jednotlivé modely vody.

Číselná hodnota	**Silové pole**
1	AMBER03 protein, nucleic AMBER94 (Duan et al., J. Comp. Chem. 24, 1999-2012, 2003)
2	AMBER94 force field (Cornell et al., JACS 117, 5179-5197, 1995)
3	AMBER96 protein, nucleic AMBER94 (Kollman et al., Acc. Chem. Res. 29, 461-469, 1996)
4	AMBER99 protein, nucleic AMBER94 (Wang et al., J. Comp. Chem. 21, 1049-1074, 2000)
5	AMBER99SB protein, nucleic AMBER94 (Hornak et al., Proteins 65, 712-725, 2006)
6	AMBER99SB-ILDN protein, nucleic AMBER94 (Lindorff-Larsen et al., Proteins 78, 1950-58, 2010)
7	AMBERGS force field (Garcia & Sanbonmatsu, PNAS 99, 2782-2787, 2002)
8	CHARMM27 all-atom force field (CHARM22 plus CMAP for proteins)
9	GROMOS96 43a1 force field
10	GROMOS96 43a2 force field (improved alkane dihedrals)
11	GROMOS96 45a3 force field (Schuler JCC 2001 22 1205)
12	GROMOS96 53a5 force field (JCC 2004 vol 25 pag 1656)
13	GROMOS96 53a6 force field (JCC 2004 vol 25 pag 1656)
14	GROMOS96 54a7 force field (Eur. Biophys. J. (2011)
15	OPLS-AA/L all-atom force field (2001 aminoacid dihedrals)

Tabulka A.2: Číselný parametr pro jednotlivé silové pole.

Příloha B

Elektronické přílohy

Součástí příloh této práce jsou následující složky:

Vypočítané struktury (přípona _equal značí složku, kde jsou struktury vypočítané se stejným počtem vodíků jako u experimentálně určených struktur; přípona _not_equal obsahuje struktury, které mají i nemají stejný počet vodíků jako u experimentálně určených struktur; složka raw zahrnuje vypočítané struktury, které neprošly žádnou úpravou):

- Gromacs – *struktury vypočítané nástrojem GROMACS*
  - AMBERG – *struktury vypočítané silovým polem AMBERG*
    - AMBERG_equal
    - AMBERG_not_equal
    - raw
  - GROMOS – *struktury vypočítané silovým polem GROMOS*
    - GROMOS_equal
    - GROMOS_not_equal
    - raw
  - CHARMM – *struktury vypočítané silovým polem CHARMM*
    - CHARMM_equal
    - CHARMM_not_equal
    - raw
- Haad – *struktury vypočítané nástrojem HAAD*
  - Haad_equal
  - Haad_not_equal
  - raw
- Chimera – *struktury vypočítané nástrojem Chimera*
  - Chimera_equal

  - Chimera_not_equal
  - raw

- OpenBabel – *struktury vypočítané nástrojem OpenBabel*
  - OpenBabel_equal
  - OpenBabel_not_equal
  - raw
  - raw_set – *struktury, kde se přesunuly vodíky z konce struktury k jednotlivým residuím*

- PyMOL – *struktury vypočítané nástrojem PyMOL*
  - PyMOL_equal
  - PyMOL_not_equal
  - raw

- Reduce – *struktury vypočítané nástrojem Reduce*
  - Reduce_equal
  - Reduce_not_equal
  - raw

Použitý dataset:

- Original – *neupravené struktury získané z RCSB PDB*
- Molecules – *struktury z Original, kde se odstranily vodíky*

Použité skripty:

- Scripts – *Python skripta pro získání RMSD (align_molecules.py), odstranění vodíků (delete_hydrogens.py), úpravu a párování vodíků (unite_hydrogens.py) a přesun vodíků z konce souboru k residuím (set_hydrogen_name.py, použité u struktur získané nástrojem OpenBabel)*
  - Detail_not_equal – *soubory s detaily ohledně neshod v počtu vodíků*
  - Errors – *soubory ukazující důvody vynechaných struktur u jednotlivých nástrojů*

- Tool_scripts – *skripta použitá pro automatické spouštění jednotlivých nástrojů*

Použité složky k porovnání:

- Parcial_charges – *hodnoty získané z porovnání parciálních atomových nábojů (Statistic.ods) a jednotlivé soubory s těmito hodnotami*

  - EQEq_charges – *složky s vypočítanými hodnotami parciálních atomových nábojů*
  - MGC-OGC - *použitý nástroj k porovnání parciálních atomových nábojů*
  - PDB_to_SDF – *soubory SDF získané přeformátováním PDB souborů*

- RMSD – *všechny hodnoty RMSD (All_rmsd.ods) a složky s vypočítanými RMSD pro jednotlivé nástroje*