Univerzita Karlova
Přírodovědecká fakulta

Fyzikální chemie


Bc. Tadeáš Kalvoda

# Studium konformačního chování krátkých peptidových fragmentů metodami kvantové chemie

Diplomová práce

Vedoucí práce: doc. Mgr. Lubomír Rulíšek, CSc. DSc.
Konzultant: Mgr. Martin Culka, Ph.D.

Praha, 2020

Charles University

Faculty of Science

Physical Chemistry


Bc. Tadeáš Kalvoda

# Conformational Behaviour of Small Peptide Fragments Studied by the Quantum Chemical Methods

Master Thesis

Supervisor: doc. Mgr. Lubomír Rulíšek, CSc. DSc.

Consultant: Mgr. Martin Culka, Ph.D.

Prague, 2020

## Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a veškerou použitou literaturu jsem citoval v seznamu použité literatury na konci této práce. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze dne 2. 4. 2020 ..... Tadeáš Kalvoda

# Poděkování

Na prvním místě děkuji doc. Mgr. Lubomíru Rulíškovi, CSc. DSc. za odborné vedení této práce, formulování jejích cílů a záměrů a postupnému měnění těchto cílů a záměrů tváří v tvář neúprosné realitě, za laskavý a vřelý přístup, každodenní nabídky kávy a vtipné historky při posezení s touto kávou, žádná však nebyla opakována více než pětkrát. Děkuji Mgr. Martinovi Culkovi, Ph.D. za 1 538 624 rad, 563 479 oprav, 421 792 připomínek, 2 obědy a 1 jízdu na kole za účelem konzultace (všechny uvedené hodnoty se pohybují v mezích 2% chyby). Děkuji Mgr. Ondřeji Guttenovi Ph.D. za stejný počet poznámek o tom, že to nebude nikdy fungovat, což snížilo můj naivní dětinský idealismus na přijatelnou úroveň. Děkuji Mgr. Danielu Bímovi, Ph.D, za objevení fontů Garamond a CMU Serif, kterými je psána tato práce. Děkuji Mgr. Adamu Jarošovi za zářný příklad toho, jaké martyrium mě čeká v případě, že úspěšně obhájím tuto práci. Děkuji Mgr. Michalu Strakovi, Ph.D, za ještě více odstrašující příklad toho, kolik starostí a organizace mě čeká, když obhájím i tu následující práci. Děkuji Lucii Tučkové za její přítomnost v této skupině, díky čemuž jsem nebyl nejmladší a nejmenší. Děkuji Mgr. Monice Staś, Ph.D. a Mgr. Agnieszce Stańczak za lekce ze základů polštiny, získání schopnosti číst kvantovou mechaniku v polštině a zjištění, že polština je v řadě ohledů archaická čeština.

Děkuji všem výše zmíněným za obohacení o ideu frustrometeru, totiž měřiče aktuální hodnoty frustrace, který se stal pevnou součástí mého každodenního života.

Děkuji výpočetnímu centru IT4I za poskytnutou výpočetní kapacitu.

Dále děkuji Kateřině Novotné za řadu příjemných a uvolňujících obědů v nejrůznějších částech budovy, které mně vždy dodaly energii a elán, nutný k mnohdy vyčerpávající práci. Děkuji Karolíně Kordačové za to, že mi svým příkladem neustále se rozvíjející osobnosti dodávala inspiraci a další motivaci nejen k této práci, ale životnímu růstu a rozvoji obecně. Děkuji Martině Mikulů za velkou pomoc ohledně správného porozumění fyzice proteinů a biofyzikální chemii, která

tvoří značnou část teoretické části této práce. Děkuji Mgr. Alexandře Berendové, Ph.D. za filologické a lingvistické konzultace k této práci.

Konečně děkuji všem ostatním lidem, kteří mě podporovali během těžkých dnů, a všem molekulám kofeinu, které mě podporovaly během těžkých večerů.

*„The scientist does not study nature because it is useful; he studies it because he delights in it, and he delights in it because it is beautiful.“*

Henri Poincaré

# Abstrakt

Do jaké míry konformační preference ukrytá v základních stavebních blocích proteinů určuje jejich trojrozměrnou strukturu? Rozsáhlé kvantové-chemické výpočty spojené s moderními solvatačními metodami představují jedinečnou sadu nástrojů k objasnění klíčových faktorů biomolekulární struktury *ab initio*. Na modelových systémech představujících krátké peptidové fragmenty byl provedeno úplné konformační vzorkování (sampling). Získané výsledky ukazují, jak tyto konformační preference mohou spoluurčovat tvorbu prostorové struktury proteinů. Zároveň poskytují nesmírně cenná data pro nalezení optimálního algoritmu, který účinně dosáhne pokrytí (ideálně všech) nízkoenergetických konformerů delších a delších peptidových fragmentů.

Klíčová slova: Konformační prostor, struktura proteinů, peptidy, solvatační metody, Ramachandranův diagram, DFT-D3 metody

# Abstract

To what extent conformational preference of short peptide sequences within proteins determine their three-dimensional structure? Large-scale quantum chemical calculations coupled with modern solvation methods represent unique set of tools to elucidate key determinants of the biomolecular structure *ab initio*. Full conformational sampling was performed on model systems representing short peptide fragments. The computed data reveal some of the underlying physico-chemical principles determining the spatial structure of proteins, and provide very important data for finding and tuning the optimal algorithm that may provide a full coverage of (ideally all) low-energy conformers.

Keywords: Conformational space, peptide fragments, protein structure, solvation methods, Ramachandran plot, DFT-D3 methods

# Obsah

# Motivace

Pochopení podstaty toho, jak konformační prostor aminokyselin či peptidových reziduí ovlivňuje tvorbu třírozměrné struktury proteinů, je jedním z důležitých cílů současné biochemie a strukturní biologie. Zkoumání konformačních preferencí (či pnutí) v proteinech a jejich ligandech může představovat nový a výpočetně sledovatelný způsob, jak významně prohloubit naše porozumění skládání (sbalování, *angl.* folding) proteinů a interakcím mezi proteinem a ligandem. Dosažení tohoto cíle by umožnilo významný pokrok – například, byť nepřímo - v oblasti navrhování a vývoje léčiv či v navrhování specifických enzymových katalyzátorů, což by značně zjednodušilo chemickou syntézu v průmyslu.

Cílem této diplomové práce je popsat, pokud možno úplně, konformační prostor modelových dipeptidů a energetickou distribuci výsledných unikátních (neredundantních) konformerů. To poskytne obrovské množství cenných dat, které poté bude možné analyzovat například metodami strojového a hlubokého učení. Na základě těchto výsledků by bylo teoreticky možné porozumět trendům a pravidlům, která určují prostorovou strukturu celého složeného proteinu a energii konečné proteinové struktury. Dále bude možné odvodit pravidla pro počet unikátních konformerů jednotlivých dipeptidů a jejich energetické distribuce, nebo vliv jednotlivých postranních řetězců, či alespoň typu řetězce (nepolární, polární, nabitý, ...) na konformační energii. S výsledky bude možné efektivně zhodnotit existující metody mapování konformačního prostoru porovnáním jejich výsledků se získaným datovým souborem.

Nakonec je v plánu prozkoumat možnosti kombinování znalostí o konformačním prostoru jednotlivých aminokyselin a dipeptidů, za účelem zjistit, zda je možné provést krok analogický matematické indukci pro předpověď podoby konformačního prostoru větších peptidových fragmentů. Dokážeme-li odvodit

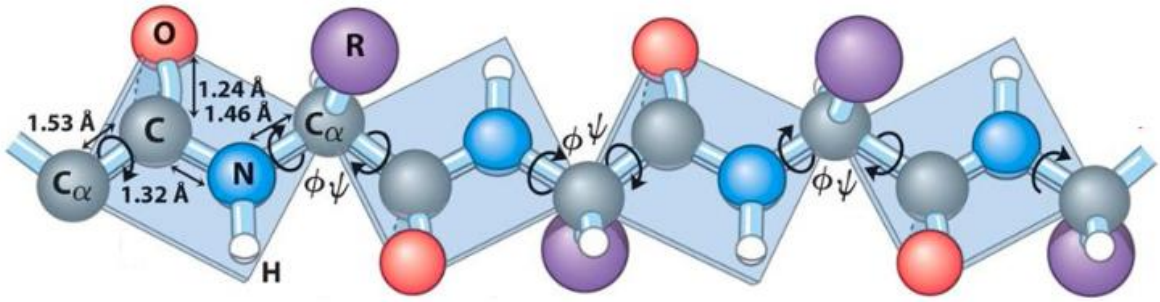
obecná indukční pravidla, jak se pohybovat v inverzní pyramidě konformačních prostorů rostoucího peptidového řetězce (viz obrázek 2.2. v kapitole 2.4.), budeme schopni predikovat skládání proteinů až do rozměrů malých proteinových domén.

# I. Kapitola

# Úvod

## 1.1. Struktura proteinů

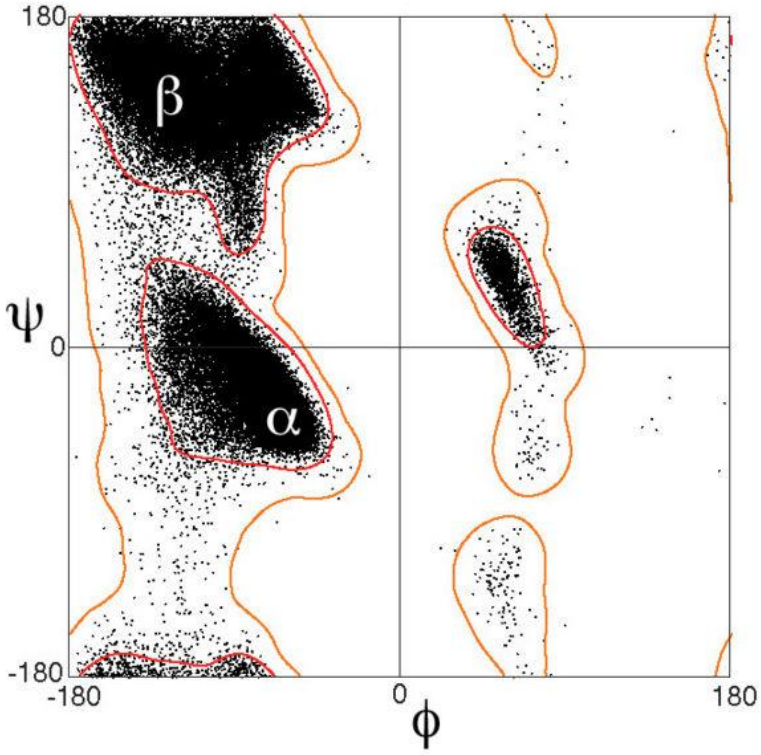
Aminokyseliny jsou molekuly běžně označované jako „základní stavební kameny života a biochemických dějů“. Existuje 20 běžně se vyskytujících, tzv. proteinogenních, aminokyselin, které se liší částí na atomu uhlíku *alfa*, označované jako postranní řetězec, zatímco zbytek (hlavní řetězec) zůstává stejný. Těchto 20 aminokyselin je možno, ať už na základě translace z genetického kódu, nebo uměle ve zkumavce, polymerizovat (principiálně bez omezení) reakcí karboxylové skupiny jedné aminokyseliny s aminovou skupinou druhé. Vzniklá vazba se označuje jako peptidová, tato vazba je kovalentní, její délka 1,32 Å [1] ji řadí mezi jednoduchou a dvojnou vazbu, a ukazuje na její částečně dvojný charakter. Díky němu má tato vazba velmi malou rotační volnost, atomy $C_\alpha$, $C_{karbox}$, O první aminokyseliny a N, H, $C_\alpha$ druhé aminokyseliny se nacházejí v rovině, a tedy peptidová vazba je planární. V principu tak existují její *cis* a *trans* izomery, v praxi však dominuje *trans* forma, vizte obrázek 1.1. [2]. Zbylé dvě vazby v hlavním řetězci jsou jednoduché, poskytují rotační volnost a lze na nich tedy definovat dihedrální úhly běžně označované jako $\psi$ a $\varphi$. Díky tomu lze při znalosti všech párů $\psi$ a $\varphi$ kompletně popsat prostorovou geometrii hlavního řetězce proteinu a při zahrnutí i úhlů postranních řetězců (obvykle značeno $\chi_1$, $\chi_2$, ...) i celého proteinu.




Obrázek 1.1.: Ilustrace dihedrálních úhlů hlavního řetězce $\psi$ a $\varphi$ na příkladu aminokyseliny v peptidovém řetězci (převzato a upraveno z [2]).

Polymerací aminokyselin vznikají kratší polymery, zvané peptidy (typicky od 2 do několika desítek aminokyselin), či delší molekuly, označované jako proteiny (typicky od několika desítek do několika stovek aminokyselin). Protein je jeden z nejdůležitějších typů molekul v živé přírodě, mající celou řadu biologických funkcí, což je možné díky tomu, že postranní řetězce aminokyselin (Dodatek C) [3] se liší svou sterickou náročností, polaritou, nábojem, reaktivitou atd., proto je možné díky velké variaci mnoha aminokyselin syntetizovat specifický protein se zcela specifickou funkcí. Funkci proteinu tedy určuje sekvence (pořadí) aminokyselin (označovaná jako primární struktura proteinu). Každá primární struktura má unikátní tendenci ke sbalování do vyšších strukturních kategorií. Sekundární strukturou se rozumí prostorová struktura krátkých peptidů v rámci proteinového řetězce, přičemž se rozlišují dva základní typy označované jako α-helix a β-list, vizte dále). Existují nicméně i části proteinového řetězce bez definované sekundární struktury, označované jako nestrukturovaný řetězec, anglicky termínem „random coil“, dále „.loops“ či „turns“. V rámci proteinu se tak vyskytují jak segmenty s nějakým typem sekundární struktury, tak i úseky bez takového typu, které jsou spojnicemi mezi nimi. Například může tedy nejprve být úsek α-helixu, poté úsek nestrukturovaného řetězce, následuje β-list a tak dále. Uspořádání sekundárních struktur a jejich (mnohdy

nestrukturovaných) spojnic v prostoru (terciální struktura) se běžně označuje termínem konformace proteinu. Některé konformace jsou stabilní a přirozeně se vyskytující, jiné jsou v podstatě nedosažitelné. Konformace proteinu se často statisticky vyhodnocuje z pohledu jednotlivých aminokyselin, což lze ilustrovat grafickým znázorněním kombinací úhlů $\psi$ a $\varphi$ nalezených v proteinu, běžně označovaném jako Ramachandranův diagram po svém tvůrci, Gopalasamudramu Narayananu Ramachandranovi [4], vizte obrázek 1.2.

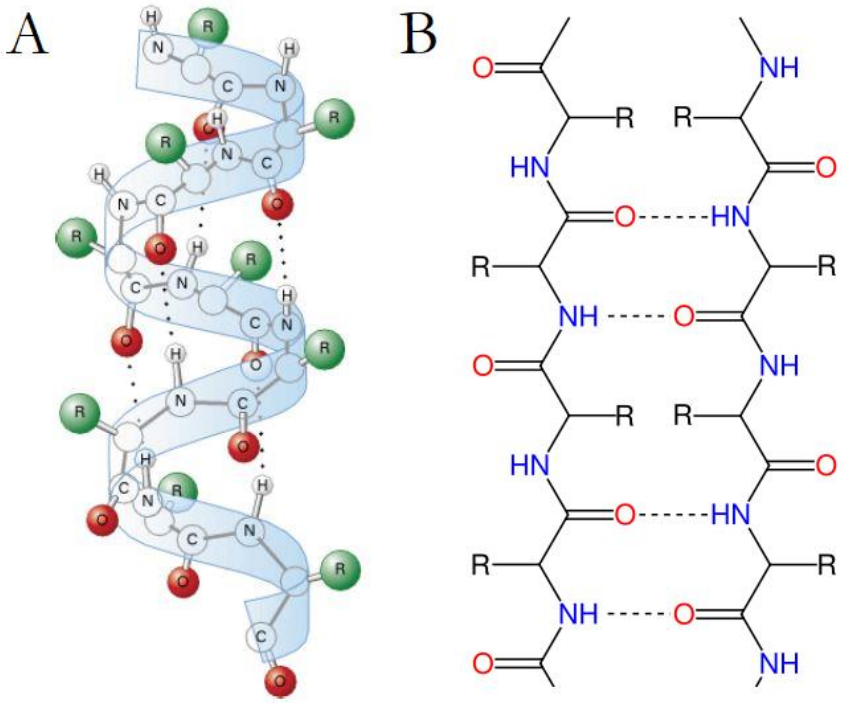



Obrázek 1.2.: Příklad Ramachandranova diagramu pro 100 000 peptidových reziduí z krystalografických dat. Červená linie označuje oblast, ve které se nachází 98 % všech bodů, oranžová pak 99,95 % všech bodů. Znaky α,β zde představují oblasti α-helixu a β-listu, vizte níže. Oblast v pravé části náleží dalším, méně obvyklým strukturám.

Tento diagram ukazuje, že zdaleka ne všechny kombinace $\psi$ a $\varphi$ se běžně vyskytují v proteinových strukturách, a ve většině případů jsou populované jen jisté domény. Tyto domény přísluší sekundárním strukturním typům. První

z nich je α-helix, což je tvar pravotočivé šroubovice, který je stabilizován vodíkovými vazbami mezi aminokyselinami $n$ a $n$+3, vizte obrázek 1.3. (A). Dalším je pak β-list, kde je hlavní peptidový řetězec téměř lineární, a struktura je stabilizována vodíkovými vazbami mezi jednotlivými řetězci, vizte obrázek 1.3. (B). Kromě těchto dvou nejvíce rozšířených typů existují ještě další typy sekundární struktury, jako je levotočivý helix (oblast okolo [60, 60] v obrázku 1.2.) β-otáčka, kolagenová struktura a další.

Konformační uspořádání celého proteinu je tedy jeho důležitá charakteristika, mající přímý vliv na jeho biologickou funkci.




Obrázek 1.3.: Znázornění struktury α-helixu (A), resp. β-listu (B). Převzato a upraveno z [5], [6].

## 1.2. Proteinový folding

### 1.2.1. Sbalování proteinů a faktory ovlivňující výslednou proteinovou strukturu

Protein je v buňce syntetizován postupným připojováním aminokyselin na základě sekvence přeložené z DNA a již během tohoto procesu nabývá konformace odpovídající biologicky aktivnímu (nativnímu) stavu [7]. Tato nativní konformace je dosažena procesem, běžně označovaným jako sbalování proteinů nebo proteinový folding. Jedná se o komplexní a velmi citlivý proces, a nabytí správné nativní struktury je klíč k funkci proteinu. Chyba během proteinového foldingu vede často k špatné či nežádoucí funkci daného proteinu a změně biochemických cest v organismu, což je důvodem řady onemocnění, jako je Alzheimerova či Huntingtonova choroba (chyba u fibrilárních amyloidních proteinů) či cystická fibróza (špatný folding transmembránového regulátorového proteinu) [8],[9].

Mezi nejvýznamnější faktory, ovlivňující proces proteinového foldingu, a tedy i podobu výsledné nativní konformace, patří:

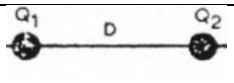
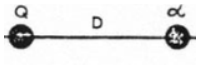
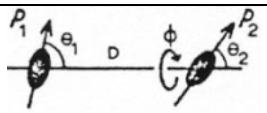
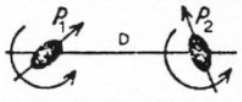
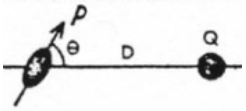
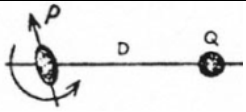
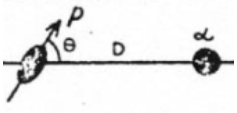
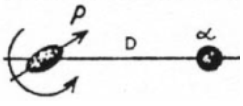
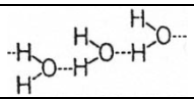
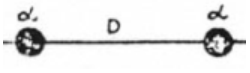
Lokální sterická náročnost, tedy pnutí mezi velmi blízkými atomy či skupinami, což je dobře známý faktor ovlivňující celou organickou chemii a téměř všechny biochemické procesy. Sterické efekty vedou k tomu, že určitá rotace vazeb v peptidu by vedla ke zvýšení energie a tedy k termodynamicky nevýhodnému ději, a proto k této změně proteinové konformace nedojde. Může však být vykompenzována výraznější změnou v jiné části proteinu, vedoucí k lokálnímu zvýšení energie díky sterickému efektu, ale globálně nižší konformační energii. Sterický efekt je krátkodosahový, ovlivňující jen nejbližší okolí dané skupiny, na kterou v rámci proteinu působí.

Chemické nevazebné interakce pokrývají celou škálu možného vzájemného ovlivňování jednotlivých částí proteinů. Liší se svým dosahem i silou, kterou k výsledné konformační energii přispívají, a lze je považovat za nejdůležitější faktor formování nativní proteinové struktury. Jejich stručný přehled ukazuje tabulka 1.1., detailnější rozebrání pak tabulka 1.2.

Tabulka 1.1.: Rozdělení nevazebných interakcí do kategorií podle typu a srovnání s vazebnou kovalentní interakcí [3].

| Interakce | Energie (síla) [kcal·mol$^{-1}$] | Poznámka |
|---|---|---|
| Kovalentní | 50 - 240 | vazebná |
| Elektrostatická | 2,4 - 7,2 | Dlouhodosahová |
| Vodíková vazba | 1,2 - 4,8 | Specifický úhel |
| Van der Waalsova int. | 0,2 - 2,4 | Disperzní síly |
| Hydrofobní interakce | 0 - 2,4 | Nesplňuje definici síly |

Tabulka 1.2.: Přehled jednotlivých typů nevazebných interakcí a jejich energie (nákresy převzaty z [3]).

| | Typ interakce | Příspěvek k interakční energii |
|---|---|---|
| Náboj - náboj |  | $U = \frac{Q_1 Q_2}{4\pi\varepsilon_0 D}$ |
| Náboj - neutr. atom |  | $U = \frac{-Q^2\alpha}{4\pi\varepsilon_0 D^4}$ |
| Dipól - dipól |  Fixované dipóly (nízká T) | $U = \frac{-\rho_1\,\rho_2}{4\pi\varepsilon_0 D^3}(2\cos\theta_1\cos\theta_2 - sin\,\theta_1 sin\,\theta_2\cos\phi$ |
| |  Rotující dipóly (vysoká T) | $U = \frac{-\rho_1^2\,\rho_2^2}{6\pi\varepsilon_0 kTD^6}$ |
| Náboj - dipól |  fixovaný | $U = \frac{-Q\,\rho\,\cos\phi}{4\pi\varepsilon_0 D^2}$ |
| |  rotující | $U = \frac{-Q^2\,\rho^2}{12\pi\varepsilon_0 kTD^4}$ |
| Dipól - neutr. atom |  fixovaný | $U = \frac{-\rho^2\alpha(1+3\cos^2\theta)}{\theta\pi\varepsilon_0 D^6}$ |
| |  rotující | $U = \frac{-\rho^2\alpha}{4\pi\varepsilon_0 D^6}$ |
| Vodíková vazba |  | $U = -\frac{Konstanta}{D^2}$ |
| Van der Waals |  | $U = \frac{-\frac{3}{4}hf_e\alpha^2}{(4\pi\varepsilon_0)^2 D^6}$ |

Nejsilnější z těchto interakcí jsou interakce s nábojem, v případě nabitého řetězce popsané známým Coulombovým zákonem, (případně Gaussovým zákonem pro intenzitu elektrického pole) které ze zmíněných typů mají nevyšší dosah, jejich pokles je úměrný druhé mocnině vzdálenosti, nicméně jsou to hlavní síly, uplatňující se při fixaci proteinového řetězce při dosažení vhodné konformace.

Dalšími faktory pak jsou pro nenabité species elektrický dipólový moment a polarizabilita, Dva (i parciální) náboje, $-q$ a $+q$, ve vzdálenost $r$ od sebe tvoří dipólový moment $p$:

$$\boldsymbol{p} = q\boldsymbol{r} \tag{1}$$

Dipólový moment je charakteristika polárních molekul. Nepolární molekuly mohou taktéž získat dipólový moment $P_{ind}$, zvaný indukovaný, který je úměrný intenzitě elektrického pole $E$ působící na molekulu, kde konstantou úměrnosti je polarizabilita α:

$$\boldsymbol{P_{ind}} = \alpha\boldsymbol{E} \tag{2}$$

Polarizabilita je experimentálně i výpočetně dostupná veličina a dává informaci o distorzi molekuly, na kterou působí elektrické pole. Protože proteiny obecně mají výrazný elektrický dipól (samotná aminokyselinová jednotka má dipólový moment $0{,}72 \cdot e$) [10], jsou tyto faktory běžné a ovlivňují výslednou proteinovou strukturu. Například v uspořádání α-helixu jsou dipólové momenty jednotlivých peptidových jednotek orientovány téměř paralelně s osou helixu, což vytváří pole o síle téměř stejné, jakou má pole pozitivního náboje na aminové či karboxylové skupině aminokyselin. Nicméně narozdíl od skutečných nábojů na proteinech, které jsou vždy stíněny jinými náboji, či solvatovány (v případě

fyziky proteinu je vždy nutné uvažovat solvatační vrstvu), což jejich výsledný účinek snižuje, fiktivní náboj dipólového momentu si zachovává svůj plný efekt.

Dalším důležitým faktem ovlivňujícím fyziku a konformační chování proteinu jsou iontové vazby, a to buď s ionty v solventu, nebo mezi nabitými skupinami postranních řetězců aminokyselin (sem patří tzv. solné můstky, vizte dále). Pokud se (v porovnání s dalšími zmíněnými interakcemi velmi silné) iontové vazby nacházejí v prostředí o vysoké dielektrické permitivitě (např. ve vodě, $\varepsilon_r = 80$), snižuje to jejich sílu na úroveň vodíkových vazeb nebo i níže. Nicméně, v případě proteinu je situace odlišná – díky amidovým skupinám má permitivita elektrického pole uvnitř proteinu hodnotu okolo 4 [3], což znamená, že iontové vazby se stávají výrazně silnějšími. Případné makroskopické náboje uvnitř proteinu jsou tak vyvolané spíše asymetrickou elektronovou distribucí na iontových vazbách.

Vodíkové vazby jsou jev, který lze popsat jako nevazebnou interakci elektrostatické povahy s residuálním charakterem vazby kovalentní, kdy odhalené kladně nabité jádro vodíku interaguje s blízkým, elektronově bohatým partnerem o vysoké elektronegativitě, přičemž je ještě nutné správné natočení vazeb zajišťující efektivní překryv orbitalů (90 % N-H·····O vazeb leží v rozmezí 140-180°, pro O-H·····O pak 90-160°) [11].

Přesný výpočet vlastností vodíkových vazeb je složitý úkol, protože zde existuje pět příspěvků o podobné síle (elektrostatická energie, výměnná repulze, polarizační energie, kovalentní a Van der Waalsovský příspěvek), nicméně vodíkové vazby jsou zásadní pro řadu vlastností důležitých pro biochemické systémy; tyto vazby potřebují ke svému vzniku elektronegativní donor vodíkové vazby, na kterém je vodíkový atom navázaný, a akceptorový atom blízko tomu vodíkovému ve směru vazby vodíku na donor. Energie těchto vazeb drasticky klesá s rostoucí vzdáleností, nelinearitou (přibližně o 10 % na 20° odchylky) a snižující se elektronegativitou, pročež v praxi se vyskytují vodíkové vazby pouze

v souvislosti s atomy F, O a N [3]. Všechny tyto podmínky jsou v případě proteinů poměrně snadno splněny, díky čemuž jsou vodíkové vazby důležitým prvkem, který se při stabilizaci proteinové struktury běžně vyskytuje a je zcela zásadní pro vznik obou hlavních sekundárních proteinových struktur, α-helixu a β-listu. Navíc v případě vodíkových vazeb může docházet k přenosu protonu, což je zcela kruciální a základní předpoklad pro množství biochemických reakcí [12], které by jinak vůbec neprobíhaly. Můžeme tedy říci, že vodíkové vazby jsou pro dosažení stabilní konformace proteinu i jeho biologickou funkci nepostradatelné.

Energie vodíkových vazeb je nejčastěji v rozmezí 2-10 $kJ \cdot mol^{-1}$ [3], což nicméně neodpovídá energii, kterou vazba přispívá ke stabilizaci proteinové struktury. V reálném případě se totiž během foldingu většiny nevazebných interakcí účastní tím či oním způsobem molekuly vody. Během foldingu jsou původní vodíkové vazby roztrhány, nicméně až 80 % karbonylových skupin na hlavním řetězci už další vodíkovou vazbu nevytvoří [13]. To samo o sobě zvyšuje enthalpii a má destabilizující vliv, nicméně díky tomu dochází k nárůstu neuspořádanosti soustavy a tedy ke zvyšování entropie, které převáží a proces je tedy termodynamicky výhodný. Příspěvek vodíkových vazeb ke stabilizaci je větší, pokud se nachází v nepolárním prostředí uvnitř proteinu než na jeho okraji.

Podobným faktorem ovlivňujícím konformaci proteinů je tvorba solných můstků, což jsou nekovalentní interakce mezi kladně nabitými postranními řetězci aminokyselin v proteinu (Lys, Arg, případně při vyšším pH i His) a jejich záporně nabitými protějšky (Asp, Glu), případně i s tyrosinem nebo serinem. Jejich význam je vyšší zejména uvnitř proteinu, kde nejsou narušovány interakcemi s molekulami vody a jejich síla je zhruba čtyřikrát větší [14], a mohou tedy stabilizovat jinak entropicky méně příznivé konformace proteinů. Jejich počet je však malý a jejich vliv je tak omezen.

Konečně Van der Waalsovy síly lze chápat jako statistické vytvoření indukovaného dipólu díky náhodné koherenci nahodilých dipólů jednotlivých atomů. Výsledkem je přitažlivá síla, úměrná polarizabilitě a ionizační energii. Pro malé molekuly je zpravidla interakce slabší, v řádu jednotek $kJ \cdot mol^{-1}$, ale pro velké makromolekuly může výsledný efekt za vhodných podmínek dosáhnout i desítek $kJ \cdot mol^{-1}$ [15][16], stále se však jedná o krátkodobý efekt, který je v porovnání s dříve zmíněnými příspěvky slabší.

Termodynamické aspekty proteinového foldingu jsou výchozím bodem při jeho zkoumání, včetně souvislosti s rovnováhou a dynamickými aspekty, protože konkrétně změny entalpie a entropie v průběhu biochemických reakcí patří k nejdůležitějším klíčům k porozumění reakčních mechanismu.

Enthalpie $H$ je veličina, udávající energii, která se ve formě tepla uvolní nebo pohltí při jednom obratu chemické reakce. Je definovaná jako:

$$H = U + pV \tag{3}$$

Kde $U$ je vnitřní energie systému (celková kinetická i potenciální energie všech částic) $p$ je tlak a $V$ objem. Tyto veličiny jsou stavové a proto je i enthalpie stavovou veličinou, která je navíc za izobarických podmínek rovna reakčnímu teplu (obecně nestavové veličině).

Entropie $S$ je pak také stavová veličina, chápaná jako hustota obsazení energetických stavů a zároveň je i kritériem uspořádanosti soustavy, což souvisí i s hustotou obsazení energetických stavů, protože nejuspořádanější stav bude ten, kdy všechny částice jsou v základním energetickém stavu. Tak lze systém realizovat jen jedním způsobem, což se dá popsat veličinou zvanou termodynamická pravděpodobnost $P$, která udává počet možných realizací daného uspořádání systému. V tomto případě bude její hodnota rovna jedné, a pohybuje se obecně v intervalu $\langle 1, \infty)$. Naproti tomu jakýkoliv jiný, méně

uspořádaný stav systému lze realizovat více způsoby a bude tedy vykazovat větší hodnotu termodynamické pravděpodobnosti. Entropie se pomocí termodynamické pravděpodobnosti definuje jako:

$$S = k \cdot \ln (P) \tag{4}$$

Kde $k$ (někdy označována $k_B$) je tzv. Boltzmannova konstanta. Protože logickým kritériem samovolného děje je, že bude probíhat do dosažení pravděpodobnějšího stavu, je růst neuspořádanosti izolovaného systému samovolným dějem, a tedy zvyšování se entropie izolovaného systému je ukazatel samovolnosti probíhajícího děje.

V praxi ovšem biochemické a biofyzikální systémy nejsou izolované, nýbrž uzavřené (tedy vyměňující práci a teplo s okolím, ale bez látkového toku). Z toho důvodu pro správný popis samovolnosti dějů v živé přírodě je nutné zahrnout do úvahy i enthalpii. Vhodným kritériem pro uzavřený systém je tak (za izobaricko-izotermických podmínek, což je však v biochemických reakcích běžně splněný požadavek) Gibbsova energie $G$:

$$G = H - TS \tag{5}$$

Kde $T$ je termodynamická teplota. Gibbsova energie je tedy kritériem samovolnosti zahrnující v sobě tepelné aspekty i aspekty uspořádání a za izotermických podmínek bude samovolný děj takový, pro který je změna Gibbsovy energie záporná:

$$\Delta G = \Delta H - T\Delta S \ < 0 \tag{6}$$

Změnu Gibbsovy energie pak udává teplota a hodnoty enthalpie a entropie. Na změnu Gibbsovy energie se pak můžeme podívat jako na funkci teploty, s následujícími výsledky:

- Pro $\Delta H > 0$, $\Delta S > 0$ je při vysoké teplotě děj samovolný
- Pro $\Delta H > 0$, $\Delta S < 0$ je děj nesamovolný při jakékoliv teplotě
- Pro $\Delta H < 0$, $\Delta S > 0$ je děj samovolný při jakékoliv teplotě
- Pro $\Delta H < 0$, $\Delta S < 0$ je při vysoké teplotě děj samovolný

Z druhého úhlu pohledu je pak samovolnost děje dána kombinací toho, jestli změny entalpie a entropie budou kladné či záporné (a jaká bude velikost těchto veličin) V praxi se stává, že i v případě, že změny obou veličin jdou z hlediska samovolnosti děje proti sobě, výrazný příspěvek jedné vykompenzuje druhou a tedy je výsledná samovolnost děj dána tímto převažujícím a určujícím příspěvkem (v angličtině označováno ,,entropy-driven“ a ,,enthalpy-driven“)

Proteinový folding je logicky samovolný děj, v jehož průběhu tedy klesá Gibbsova energie a nativní struktura je pak struktura o minimální Gibbsově energii. Efektů snižujících Gibbsovu energii je celá řada. Jedním případem jsou například výše zmíněné vodíkové vazby či solné můstky, Druhým významným faktorem je tzv. hydrofobní efekt (někdy nesprávně nazýván hydrofobní interakce, ale nejedná se v tomto případě o žádnou sílu, která by interakci způsobovala).

Molekuly vody mají tendenci obklopovat protein klecovým efektem kolem nepolárních míst a tím snižovat entropii. V takovém případě je systém ve stavu vysokého uspořádání, nicméně po sbalení proteinu je tento uspořádaný stav narušen, molekuly vody se mohou vrátit do původního neuspořádaného stavu, a díky tomu se zvýší entropie a tedy klesne Gibbsova energie celého systému, což vede k termodynamické stabilizaci proteinu. Hydrofobní efekt je tedy hnaný

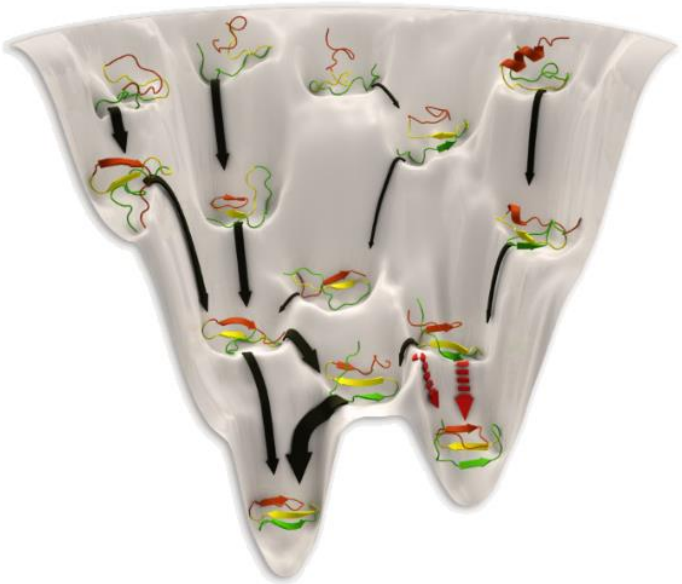
zvyšující se entropií a je obecně považován za nejvýraznější efekt pro formování zejména globulárních proteinových struktur.

Vzhledem k tomu, že entropie i enthalpie je rozdílně závislá na teplotě, je hydrofobní efekt za určité teploty nejsilnější, a slábne nad a pod touto teplotou. Vymizení hydrofobního efektu se snižující se teplotou je pak hlavní příčinou tepelné denaturace proteinů, kde je proteinu zbavován nativní konformace jeho postupným ochlazováním. Síla hydrofobního efektu byla experimentálně měřena a v současné době existují i její výpočetní modely.

### 1.2.2. Teorie proteinového foldingu

Levinthalův paradox je myšlenkový experiment či spor, formulovaný molekulárním biologem Cyrusem Levinthalem v roce 1969 [17], [18]. Jeho podstata spočívá v tom, že jakýkoliv běžný protein, sestávající se ze stovek aminokyselin, má odhadem $10^{150}$ stupňů volnosti – každá aminokyselina má totiž nejméně tři (*trans* a dvě *gauche*) stabilní konformace při rotaci dihedrálního úhlu $\psi$ nebo $\varphi$, což by znamenalo, že jen analytickým měněním jednotlivých dihedrálních úhlů s tím, že by taková změna proběhla za milisekundu, by protein nedosáhl nativní konformace za dobu trvání vesmíru. Přesto menší proteiny jsou schopné dosáhnout nativní struktury v řádu desítek milisekund. Tento rozpor je vysvětlován například tím, že proteinový řetězec nejprve nabyde globulárního tvaru, a teprve poté se zformují elementy sekundární struktury [19], anebo tím, že folding je urychlen lokálními doménovými interakcemi, které zafixují části proteinu a následně působí jako výchozí opěrné body pro další fázi foldingu. Pro tyto segmenty o už výsledné konformaci, které jsou nakonec součástí celkové nativní struktury existuje v angličtině výraz „foldons“ [20]. Tyto metastabilní tranzitní stavy proteinového foldingu byly předpovězeny a následně i experimentálně nalezeny [21], z čehož

vyplynula představa tzv. foldingového trychtýře – totiž že závislost energie proteinu během foldingu má trychtýřovitý tvar, jehož stěny nejsou hladké, ale vyskytuje se na nich řada mělkých minim, která odpovídají právě metastabilním tranzitním stavům v průběhu fodlingu.



Obrázek 1.4.: Schéma foldingového trychtýře s naznačenými metastabilními strukturami. Převzato a upraveno z [22].

Tato představa je spjata s ideou tzv. hydrofobního kolapsu, podle které je stabilizace dosažena díky rychlému shlukování hydrofobních postranních řetězců aminokyselin, ze kterých je protein tvořen, do jakýchsi „semi-micel“ což zvyšuje entropii okolní vody, čímž klesá celková energie. Navíc nabité a polární postranní řetězce nejsou hydrofobními tolik stíněny, dochází k jejich snadnější interakci s molekulami vody a zřejmě i rozrušení solných můstků uvnitř proteinu. Hloubka „foldingového trychtýře“ pak je dána konformační energií, tedy rozdílem mezi lineární (původní) konformací proteinu a jeho nativní formou.

Zároveň nativní proteinová struktura nemusí nutně představovat globální energetické minimum, pokud je toto minimum obtížně kineticky dosažitelné nebo není dosažitelné vůbec.

Floryho hypotéza izolovaného páru je aproximativní představa, říkající, že téměř veškeré lokální sterické bránění na proteinu v nenativní konformaci nedosahuje dál, než na nejbližší sousední aminokyselinové reziduum, tedy že dihedrální úhly $\psi$,$\varphi$ je stericky nezávislé na svých sousedech. Tato hypotéza byla v poslední době podrobena intenzivnímu testování, kdy byly hustěji populované oblasti Ramachandranova diagramu konformačně vzorkovány a bylo zjištěno, že platí pro region β-listu, ale selhává pro všechny ostatní regiony, což je očekávatelné zjištění, protože ve struktuře β-listu jsou páry skutečně izolované [23]. Selhání této hypotézy i pro krátké peptidy ($n$=7) zpochybňuje platnost „random coil" představy, protože populace nenativních konformací je díky tomu výrazně bohatší, než odpovídá „random coil" představě. Nicméně novější práce [24] ukazují, že efekt je minimální už pro větší oligopeptidy, a tedy „random coil" hypotéza může být přesto považována za správnou. Tato práce podrobuje zkoumání platnost Floryho hypotézy zkoumáním vzájemného ovlivňování konformačních preferencí dvou sousedních aminokyselin v modelu dipeptidu s chráněnými konci, kdy N-koncová aminoskupina byla ochráněna acetylovou čepičkou, C-koncová karboxylová skupina N-methylovou čepičkou.

### 1.2.3. Metody studia proteinového foldingu

Obecně lze definovat dva základní přístupy ke studiu proteinového foldingu – experimentální a výpočetní. Experimentální studium proteinového foldingu je starší a lze ho opět dělit na zkoumání vlastního procesu foldingu, a techniky popisu výsledné struktury.

Pro model výsledné struktury je běžnou metodou krystalová strukturní analýza. Soustředěním rentgenového záření na krystalovou mřížku, kde jeho paprsky

podléhají difrakci, následkem čehož je výstupem měření tzv. difrakční obrazec. Jeho následnou analýzou (spojenou s obtíží vyřešení fázového problému) pak je možné zkonstruovat model nativní struktury. Kromě toho se používá NMR spektroskopie, a v poslední době nachází výrazné využití metoda kryogenní elektronové mikroskopie, zvaná Cryo-EM [25]. Samotný proces foldingu pak lze studovat například použitím fluorescenční spektroskopie [26], kde se využívá vysoký kvantový výtěžek fluoresence u Tyr a Trp (spojený i s tím, že vzhledem k jejich nepolárnímu charakteru jsou často v hydrofobních oblastech uvnitř proteinu). Při změně konformace a jejich vystavení hydrofilnímu prostředí se fluorescence snižuje. Tato technika tedy může být použita pro detekci různých přechodových stavů. Dále se používá např. cirkulární dichroismus, kde se využívá interakce chirálních prvků sekundárních struktur s cirkulárně polarizovaným světlem. Úroveň absorpce pak ukazuje na stupeň sbalení proteinu. Kromě těchto technik existuje množství dalších, jako je různé využití NMR, časově rozlišení laserová spektroskopie a další [27].

Teoretické metody studia proteinového foldingu, resp. predikce nativní struktury pak lze rozdělit do čtyř kategorií [28]. První jsou metody typu *ab initio* (v tomto kontextu výraz *ab initio* nesouvisí s označením kvantově-chemické třídy metod), tj. bez „externí“ informace z proteinových databází, tedy bez dalších informací nad rámec aminokyselinové sekvence studovaného proteinu - jedinou informací mimo vlastní strukturu jsou případně údaje pro parametrizaci použitého silového pole (např. AMBER, CHARMM, GROMACS [29], a jiné.). Tyto metody tedy hledají minimum použité funkce popisující volnou (Gibbsovu) energii. V případě přílišné výpočetní náročnosti je užitečná tzv. geometrická reprezentace, kdy protein namísto souboru všech atomů reprezentuje nějaký zjednodušený model, jako například neuvažování vodíků neschopných vytvořit vodíkovou vazbu, nebo implicitní rozpouštědlo.

Druhá možnost je užít spolu s ab-initio přístupem informace z proteinových databází, kdy se obvykle porovnávají fragmenty studovaného proteinu s fragmenty v již popsaných případech nativních struktur, protože nově objevené typy nativních struktur („nové foldy“) se obvykle sestávají z prvků již popsaných případů [30]. Díky tomu může být nativní struktura „sestavena“ z těchto nalezených podobných fragmentů. Při sestavování je pak brán zřetel na požadavek minimální energie. Do této skupiny patří např. metoda I-TASSER [31] či ROSETTA [32] nebo FRAGFOLD [33].

Třetím typem jsou pak metody, které se zakládají na předpokladu, že struktura je konzervována více než sekvence, díky čemuž by proteiny bez podobné sekvence mohly podléhat podobnému typu foldingu. Dle rozsáhlých studií [34] se totiž 10 běžných typů nativní struktury objevuje v circa 50 procentech proteinů se známou strukturou [35]. Cílem je tedy najít a vybrat správný už známý strukturní model a správně zaměnit jeho části sekvence. Sem patří např. metody GENTHREADER či HHpred.

Posledním typem pak jsou metody srovnávacího modelování, kde je cílová sekvence aminokyselin v proteinu srovnána s jiným proteinem o už známé nativní struktuře a pokud je sekvence podobná, tak jsou informace o známém proteinu použity pro modeling nativní struktury zkoumaného proteinu. Do této skupiny patří SWISS-MODEL [36] či PyMOD [37].

Lze tedy říci, že teoretické studium proteinového foldingu je extrémně náročný úkol, protože se jedná o velmi složitý problém ovlivněný mnoha proměnnými, a kde jsou metody využívající databázové informace ovlivněny jejich principiální neúplností, a naproti tomu metody typu *ab initio* zase vysokou dimenzionalitou konformačního prostoru a tedy i extrémní výpočetní náročností. Z toho důvodu je toto téma stále předmětem intenzivního vývoje, s vidinou nalezení obecně použitelné, ideálně *ab initio* metody, vycházející z malých fragmentů proteinu či

aminokyselin a schopné správně a přesně predikovat proteinovou strukturu, což je i jedním z cílů této práce.

II. Kapitola

# Výpočetní chemie

## 2.1. Teoretická chemie

Následující část o teoretickém základu výpočetních metod užívaných v chemii a jejich použité v rámci této práce. Přestože se pojmy teoretická a výpočetní chemie do značné míry překrývají, lze vymezit teoretickou chemii jako obor zabývající se matematickým a fyzikálním popisem chemických jevů. Výpočetní chemie zahrnuje především implementaci a automatizaci matematických metod a modelů pro aplikace na chemické problémy.

### 2.1.1. Základní pojmy

Schrödingerova rovnice je základní stavební kámen celé kvantové mechaniky a výpočetních metod. Tato rovnice popisuje, ve své stacionární podobě, vztah mezi vlnovou funkcí daného systému (například molekuly) a její energií, vizte rovnici 7:

$$\hat{H}(r)\psi(r) = E(r)\psi(r) \tag{7}$$

Kde $\hat{H}(r)$ je časově nezávislý Hamiltonův operátor, $\psi(r)$ časově nezávislá vlnová funkce popisující daný systém (atom, molekulu aj.) a $E(r)$ energie tohoto systému. V praxi lze analyticky vyřešit Schrödingerovu rovnici pouze pro atom vodíku (přesněji pro atomy vodíkového typu, např. $He^{+}$). Z tohoto důvodu jsou metody výpočetní chemie poskytující řešení Schrödingerovy rovnice pro molekulové systémy vždy aproximativní.

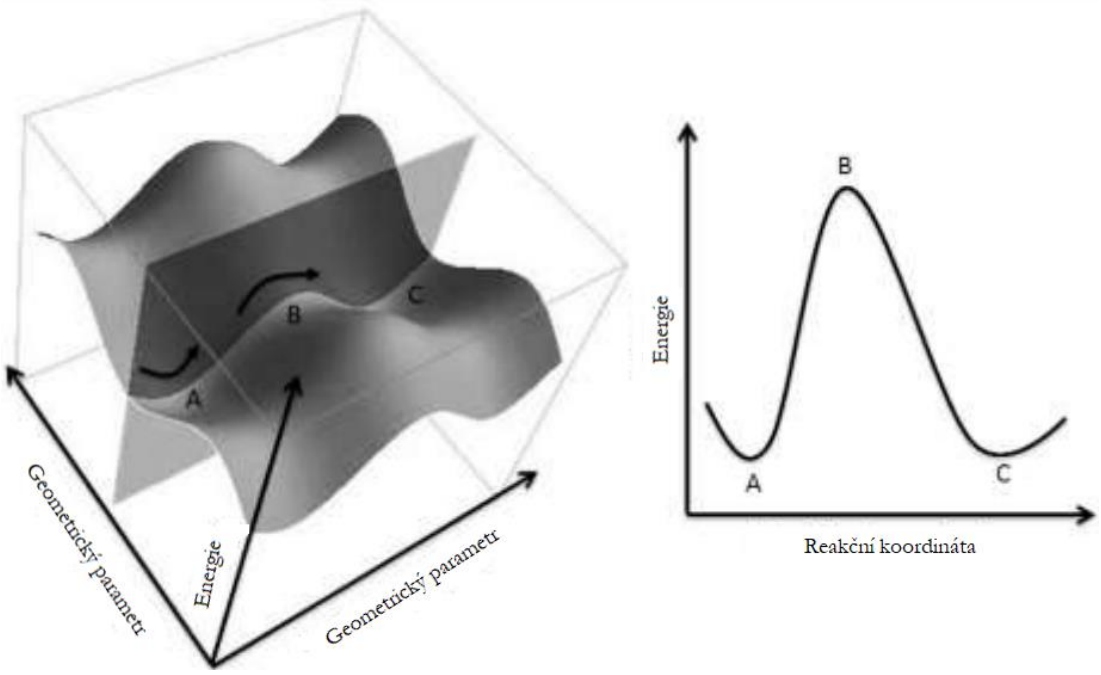
Bornova-Oppenheimerova aproximace (BOA) je nejzákladnější a nejdůležitější aproximace v kvantové chemii, která výrazně zjednodušuje řešení kvantově

mechanických systémů, které by jinak bylo možno řešit složitě nebo nebylo možno řešit vůbec. Základ této aproximace spočívá v tom, že jádra atomů jsou řádově $10^3$ krát těžší než elektrony, proto jsou jejich rychlosti v porovnání s rychlostmi elektronů mnohem menší, a tedy elektronický a jaderný pohyb v molekulách lze považovat za nezávislé na sobě (separovatelné). Díky tomu je možné nahlížet na konfiguraci jader jako na fixní, z pohledu pohybu elektronů. Proto elektronová vlnová funkce závisí na jaderných pozicích, ale pouze parametricky, jak ukazuje rovnice 8:

$$\psi_{tot}(R,r) \approx \psi_{nuc}(r)\psi_{el}^{R}(r) \tag{8}$$

Kde $\psi_{tot}(R,r)$ je celková vlnová funkce závislá na souřadnicích jader $R$ a elektronů $r$, $\psi_{nuc}(r)$ vlnová funkce jader a $\psi_{el}^{R}(r)$ vlnová funkce elektronů. Nevýhoda této separace je zanedbání spřažení mezi jaderným a elektronovým pohybem, které při některých chemických procesech zanedbat nelze. Nicméně, pro většinu chemických problémů Bornova-Oppenheimerova aproximace funguje výborně a její použití vede ke standardním kvantovým metodám (QM).

Hyperplocha potenciální energie je matematický vztah mezi energií molekuly (nebo obecně studovaného systému) a jejími jadernými souřadnicemi. Tento vztah je možný díky Bornově-Oppenheimerově aproximaci. Hyperplochu potenciální energie lze například získat postupným vyřešením elektronové Schrödingerovy rovnice pro známou fixní konfiguraci jader. Předpokládejme, že systém má $N$ jader. Pak lze definovat 3 souřadnice x, y, z pro každý atom, což poskytuje $3N$ stupňů volnosti. Když odstraníme ty z nich, které popisují tři translace těžiště celé soustavy - ve směrech x, y, z a dále tři rotace – podle osy x, y, z, je výsledkem počet nezávislých souřadnic $3N$-6 (speciálně $3N$-5 pro lineární molekuly, kde jedna z rotací poskytuje fyzikálně nerozlišitelné konfigurace). Příklad hyperplochy je uveden na obrázku 2.1.




Obrázek 2.1.: Schematické znázornění hyperplochy potenciální energie pro případ dvou (vlevo) resp. jednoho (vpravo) rozměru [38].

### 2.1.2. Metody výpočetní chemie

Jedním z cílů výpočetní chemie je popsat hyperplochu potenciální energie co možná nejpřesněji, a zároveň za co možná nejmenší výpočetní námahu. Všechny přístupy jsou ve své podstatě aproximacemi výše zmíněné (exaktní) Schrödingerovy rovnice. Současná výpočetní chemie disponuje velkou škálou různých metod, vycházejících jak z variačního, tak poruchového přístupu, stejně jako i kombinací kvantově mechanického, klasického a empirického přístupu. Každá z těchto metod má své silné stránky a nedostatky a jejich znalost je důležitá pro volby vhodného nástroje pro daný výpočetní problém.

Metody molekulové mechaniky (MM) jsou založeny na představách klasické fyziky, čili neřadí se mezi metody kvantové chemie. Tento typ metod ignoruje pohyb elektronů a hyperplochu potenciální energie se snaží aproximovat pouze jako funkci souřadnic jader, s řadou empirických nebo semiempirických

parametrů, takzvaného silového pole (angl. force field, FF). Nejčastější přístup je funkce sestávající se ze součtu několika členů, jmenovitě odchylky vazebné délky od rovnovážné délky $l_{i,0}$ pro každý atom $i$ z celkem $N$ atomů, vazebného úhlu $\theta_{i,0}$, příspěvku deformace dihedrálního úhlu $\omega$ a nevazebných, např. Coulombických a van der Waalsových interakcí:

$$\boldsymbol{U(r^N) = \sum_{vazby} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{vaz.\ úhly} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{dihed.\ úhly} \frac{k_i}{2}(1 + \cos(n\omega - \gamma))}$$

$$+ \sum_{dihed.\ úhly} \frac{k_i}{2}(1 + \cos(n\omega - \gamma)) + \sum_{i=1} \sum_{j=i+1} \left( 4\varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \quad (9)$$

Kde $U(r^N)$ je energie závislá na $N$ jaderných souřadnicích $r$, $l_i, l_{i,0}$ je obecná resp. rovnovážná vazebná délka, $\theta_i$, $\theta_{i,0}$ obecný resp. rovnovážný vazebný úhel, $k_i$ silové, úhlové či torzní konstanty, $\sigma_{ij}$ a $\varepsilon_{ij}$ konstanty z Lennard-Jonesova potenciálu 12-6, $q_i, q_j$ náboje atomů $i$ a $j$ a $\varepsilon_0$ permitivita vakua a $n, \gamma$ konstanty kosinového rozvoje torzního potenciálu.

První dva příspěvky vycházejí z Hookova zákona harmonického oscilátor, třetí člen je jedním z možných vyjádření energetického příspěvku torzního potenciálu, poslední člen je kombinací Lennard-Jonesova potenciálu 12-6 a Coulombova zákona. Všechny tyto vztahy jsou založené na klasické fyzice.

Fakt, že tento přístup je aplikovatelný, je důsledkem řady předpokladů – jako je BOA, bez které by nebylo možno energii jako funkci jader vůbec vyjádřit, představa harmonické aproximace, atd.

Podstatná vlastnost MM přístupu (resp. silového pole) je jeho přenositelnost (transferabilita), kdy vyladění parametrů pro menší množství problémů dovoluje rozsáhlejší použití. Pro zcela nové molekuly je nicméně nutné najít správné hodnoty parametrů (tedy silových konstant vazeb, úhlů, torzního potenciálu

atd.), což není vždy jednoduché. Běžně užívanými silovými poli jsou např. AMBER či CHARMM. [29], [39], [40].

Jedním z příkladů užití MM metod je například modelování enzymové kinetiky, kdy je většina enzymu modelována MM metodikou a pouze pro aktivní místo se substrátem, vyžadující větší preciznost a přesnost, je užita nějaká QM metoda. Tento přístup se běžně označuje QM/MM modelování.

MM force field lze velmi dobře použít na studium velkých proteinů [41]. Nicméně se ukazuje, že MM selhává v přesném popisu *lokálních* interakcí a repulzí v proteinu [42]. Z tohoto důvodu nebyly v této práce MM metody výrazněji použity, přestože (a protože) předmětem výzkumu jsou krátké proteinové fragmenty, respektive jejich energie a vlivy lokálních interakcí na proteinový folding, a byla raději dána přednost DFT přístupu, který je výpočetně náročnější, ale lze od něj očekávat významně přesnější výsledky.

*Ab initio* metody jsou metody, založené na řešení Schrödingerovy rovnice, a název *ab initio* (od počátku, bez dalších znalostí) odkazuje na fakt, že při výpočtech postačuje znalost fyzikálních konstant, nikoliv informace o řešeném výpočetním problému nebo experimentálně změřená data. Při tom je využívána celá řada různě přesných aproximací, které výpočty zjednodušují a činí je realizovatelné ve smysluplném čase.

Obvyklým výchozím bodem těchto metod je Hartreeho-Fockova (HF) metoda, která je dnes málo využívaná [43], ale tvoří základní stavební kámen *ab initio* metod. V Hartreeho-Fockově metodě je elektronová interakce popisována jako interakce každého elektronu se zprůměrovaným polem ostatních elektronů. Zásadním problémem však je, že je zanedbán příspěvek okamžité mezielektronové repulze, a toto zanedbání (v souladu s variačním principem) vypočítanou energii oproti správné hodnotě zvyšuje (obvykle cca o jedno procento). Tento rozdíl se nazývá *korelační energie* a je vždy záporný.

Pro lepší přesnější výpočet je nutno užít pokročilejší, tzv. post-HF metody. Mezi tyto patří například Møllerova–Plessetova (MP) metoda, která korelační energii přidává pomocí poruchové teorie. V závislosti na řádu poruchové teorie (první řád dává jen HF energii) se varianty této metody nazývají MP2 (korekce z poruchové teorie druhého řádu) MP3, MP4 atd. Tyto metody se dříve používaly např. pro výpočty rychlostních konstant [44].

Dalším možným přístupem je Metoda konfigurační interakce (CI) která používá variační princip a skutečnost, že přesná vlnová funkce se dá vyjádřit pomocí všech excitovaných stavů (tedy jednou, dvojitě atd. excitovaným) jako jejich lineární kombinace. Při uvážení excitací ze všech obsazených do všech neobsazených (virtuálních) orbitalů dostáváme tzv. full-CI metodu (FCI) o extrémní přesnosti, ale velké výpočetní náročnosti. V praxi je možná spíše CISD metoda, kde se uvažují jen jedno- a dvojnásobné excitace. I tak je ale problematická pro větší systémy, její přesnost klesá s velikostí systému.

Konečně lze dobrého popisu (zahrnutí) korelační energie dosáhnout metodou spřažených klastrů (CC, coupled cluster). Podstatou této metody je upravení CI rozvoje podle zapojení jednotlivých excitovaných stavů do vlnové funkce. Výhoda tkví v tom, že oproti CI stejného řádu (např CISD a CCSD, jakožto metoda spřažených klastrů pro jedno- a dvouelektronové excitace) jsou hodnoty korelační energie přesnější. Zvláštní variantou je dnes hojně rozšířená CCSD(T) metoda, která trojité excitace zahrnuje v podobě poruchové teorie, a která je pokládána za „zlatý standard“ výpočetní chemie.

Kromě těchto přístupů existuje celá řada dalších metod, jako např. mulitereferenční metody CASSCF, MRCI, a další.

Řešení Schrödingerovy rovnice jedním z těchto přístupů (při aplikování Bornovy-Oppenheimerovy aproximace) vede k zmapování hyperplochy potenciální energie jakožto závislosti energie, získané řešením elektronické

Schrödingerovy rovnice, a souřadnic jader. To je však v mnoha případech velmi výpočetně i časově náročné proto byly pro účely získání závislosti energie systému na souřadnicích atomů vyvinuty i jiné přístupy, jako jsou metody založené na teorii hustotního funkcionálu, vizte dále.

Metody založené na teorii hustotního funkcionálu (DFT) jsou v současné době zdaleka nejrozšířenější metody, které dosahují podobné výpočetní (ne)náročnosti jako HF metody, jsou ovšem mnohem přesnější, protože zahrnují elektronovou korelaci. Základní rozdíl totiž spočívá v tom, že jejich ústřední bod není vlnová funkce, nýbrž elektronová hustota, jakožto funkce tří prostorových souřadnic. Matematicky je zjednodušeně funkcionál operátor, přiřazující určité funkci hodnotu (číslo). Variací funkcionálu je pak míněna změna výsledné hodnoty se změnou funkce. Pro minimum funkcionálu analogicky platí, že variace je nulová.

Elektronová hustota má na rozdíl od vlnové funkce fyzikální význam pravděpodobnosti výskytu elektronu, je maximální v poloze jader, její integrál dává počet elektronů a její derivace souvisí s nábojovým číslem.

Tyto veličiny ale zároveň jednoznačně definují molekulový Hamiltonián, což znamená, že energie je funkcionálem elektronové hustoty. Na této myšlence jsou založeny i Hohenbergovy–Kohnovy teorémy.

Podle prvního Hohenbergrova-Kohnova teorému je elektronová energie jednoznačným funkcionálem elektronové hustoty. Díky tomu je možné, při znalosti funkcionálu a elektronové hustoty zjistit energii bez použití vlnové funkce. Přesný tvar takového funkcionálu není známý, o elektronové hustotě však nicméně mluví druhý Hohenbergův–Kohnův teorém, obdoba variačního principu. Ten říká, že pro daný externí potenciál, kterému přísluší daná (přesná) elektronová hustota, bude platit, že energie získaná dosazením jiné elektronové hustoty do přesného funkcionálu bude vždy větší. Správnou elektronovou hustotu tedy lze získat variačním principem, tedy minimalizací, resp.

zpřesňováním, zkusmé elektronové hustoty [45]. Moderní DFT metody jsou založeny na formulaci přibližných hustotních funkcionálů, což umožnila Kohnova-Shamova metoda (vedoucí k tzv. Kohnovým-Shamovým rovnicím). Kohnovy-Shamovy rovnice jsou založené na modelu neinteragujících elektronů, čímž se jednodušeji získá funkcionál kinetické energie, coulombická část přesného funkcionálu, a pro zbytek, běžně označovaný jako výměnně-korelační funkcionál se hledají vhodné funkční formy, často (semi)empirickým přístupem.

Novější funkcionály využívají gradient elektronové hustoty, (metody zobecněného gradientu, GGA). Takových funkcionálů je celá řada, jmenovitě například B88 [46], BLYP [47], P86 [48], PBE [49], či BP86 [50] využívaný pro výpočty v této práci. Dalším vylepšením je pak zahrnutí druhé derivace elektronové hustoty, což využívá rodina meta-GGA funkcionálů (B95, TPSS)

Další třídu tvoří hybridní funkcionály, kde je část výměnné energie pochází přímo z výpočtu metodou Hartreeho-Focka. Příkladem takového funkcionálu je zřejmě historicky nejpopulárnější funkcionál B3LYP, který je zřejmě i nejpoužívanější kvantově-mechanickou metodou vůbec [43].

Moderní metody DFT navíc obsahují i korekci na Londonovy disperzní síly, kterou Kohnova-Shamova teorie predikuje značně nepřesně [51]. Obecně disperzní energetický příspěvek závisí na šesté mocnině vzdálenosti:

$$E_{disp} = \frac{C_6}{R^6} \tag{10}$$

Kde $C_6$ je disperzní koeficient šestého řádu. Klasické DFT funkcionály dobře modelují příspěvek pro malé vzdálenosti, pro větší však selhávají, A proto bylo vynaložené velké úsilí pro řešení tohoto problému, které vyústilo např. v semilokální funkcionály, vylepšené Van der Waalsovské DFA, optimalizované jednoelektronové potenciály aj. Nicméně větší pokrok nastal až ve 21. století,

kdy byly vyvinuty mimo jiné metody DFT-D [52]. Základ korekcí DFT-D spočívá v uvážení všech možných párů atomů v chemickém systému a jejich příspěvků podle Londonovy teorie. Jejich součet je pak disperzní korekce:

$$E_{disp}^{DFT-D} = -\frac{s_6}{2}\sum_{A\neq B}\frac{C_6^{AB}}{R_{AB}^6} f_{damp}^{DFT-D}(R_{AB}) \tag{11}$$

Kde $s_6$ je globální škálovací parametr šestého řádu, $R_{AB}$ vzdálenost atomů $A$ a $B$, $C_6^{AB}$ disperzní koeficient šestého řádu získaný jako geometrický průměr koeficientů pro dané atomy, a $f_{damp}^{DFT-D}$ je tlumící faktor závislý na $R_{AB}$ jehož funkce je zabránění vícenásobnému zahrnutí disperze. Nicméně i tento přístup má problémy už při d3 a d4 prvcích. Tyto nedostatky odstraňuje korekce D3 [52]. Verzí korekce D3 existuje více, například v DFT-D3(BJ) je výraz pro disperzní korekci upravený a tlumící faktor pochází z práce Beckeho a Johnsona [53]:

$$E_{disp}^{DFT-D3(BJ)} = -\frac{1}{2}\sum_{A\neq B}\sum_{n=6,8} s_n \frac{C_n^{AB}}{R_{AB}^n \left[f_{damp}^{DFT-D3(BJ)}(R_{BJ}^{AB})\right]^n} \tag{12}$$

Kde $s_n$, $C_n^{AB}$ mají analogický význam jako v předchozím případě, a tlumící faktor je konečná funkce:

$$f_{damp}^{DFT-D3(BJ)}\left(R_{BJ}^{AB}\right) = \alpha_1 R_{BJ}^{AB} + \alpha_2 \tag{13}$$

Parametry $\alpha_1, \alpha_2$ jsou měnitelné a kontrolují disperzi pro krátké i delší vzdálenosti. $R_{AB}$ pak už je definován jako odmocnina poměru disperzních koeficientů osmého a šestého řádu, které jsou získány z hodnot atomového náboje a dipólového a kvadrupólového momentu.

Analogickým způsobem je řešena i korekce na disperzi tříatomárního příspěvku, její matematický popis je však složitější. Podrobnější informace lze najít například v [54],[55]).

Ve výsledku jsou disperzně korigované DFT-D3 mimořádným zlepšením DFT výpočtů [56], a měly zásadní dopad na výpočetní chemii, zejména proto, že korekce je odvozena pouze z geometrie řešeného systému, a to za zcela zanedbatelný výpočetní čas. Díky tomu, že disperzní koeficienty nejsou závislé na konkrétním systému, je DFT-D3 nejen přesnější, ale i flexibilnější a umožňuje aplikaci pro naprostou většinu prvků periodické tabulky.

Lze tedy říci, že DFT metody zejména po rozšíření o disperzní korekci D3, patří svou rozsáhlostí a univerzálností k páteři současné výpočetní chemie a mnohokrát prokázaly svou užitečnost. Zejména z důvodů uvedených výše bylo spojení funkcionálu BP86 s DFT-D3 použito na hlavní výpočty v této diplomové práci.

Semiempirické metody jsou metody, které využívají základů *ab initio* metod, protože vychází z elektronové Schrödingerovy rovnice, ale pouze s použitím efektivního hamiltoniánu. Při řešení využívají různé empirické parametry, například parametrizují či rovnou zanedbávají různé typy integrálů, vnitřní elektrony zahrnují jako součást efektivního hamiltoniánu, apod. Jejich smyslem je tak snížit výpočetní náročnost na rozumnější míru, při co největšímu se přiblížení přesnosti *ab initio* metod.

Nejjednodušší semiempirickou metodou je rozšířená (*angl.* extended) Hückelova metoda, v rámci které se při řešení sekulárních rovnic například zanedbávají překryvové integrály AO na různých atomech, a další integrály nahradíme různými parametry ze spektroskopických či termochemických dat. V některých případech takto lze dobře odhadnout např. excitační energii. V současné době se

Hückelova metoda už nepoužívá, nicméně byla důležitým průkopníkem ve vývoji semiempirického přístupu v kvantové chemii.

Moderní semiempirické metody používají obecně sofistikovanější parametrizaci, umožňující širší použití s dobrými výsledky a jsou užívané ve spojení s celou řadou populárních programů. Rozšířenými semiemprickými metodami jsou například AM1, založená na aproximaci nulového překryvu různých orbitalů s korekcí na repulzi [57], nebo PM3 [58] kde je použit základ metody AM1, ale s automatickou, ,,ab initio" parametrizací, či pokročilejší SAM1 [59] a další. Jiným typem jsou semiempirické metody založené na DFT. Příkladem takové metody je GFN2-xTB, dále označovaná jen jako xTB [60],[61]. Tato metoda je široce aplikovatelná, obsahuje komplexnější elektrostatické a výměnně-korelační příspěvky k hamiltoniánu. Velkou devizou je také zahrnutí self-konzistentní korekce na disperzi D4, díky čemuž je možné efektivně zahrnout efekty elektronové struktury. Používané parametry jsou spolehlivé pro všechny prvky až do radonu, a poskytují dobré modely struktur, vibračních frekvencí a nekovalentních interakcí.

GNF2-xTB byla od počátku (2016) specificky vyvíjena jako metoda pro organické a biochemické systémy v řádu tisíců atomů. Oproti předchozí verzi GFN-xTB poskytuje výrazně vylepšený popis nekovalentních a elektrostatických interakcí. Jeví se jako velmi vhodná pro studium proteinových interakcí a konformačního chování proteinů. Protože výše zmíněné typy studií jsou součástí této práce (vizte podkapitolu 2.4.), byla tato metoda v rámci této práce využívána pro geometrickou optimalizaci (podrobnější popis využití je uveden v kapitole 3.1.2).

## 2.2. Metody konformačního samplingu a predikce proteinové struktury

Mapování konformačního prostoru biomolekul, také jinak zvané konformační sampling, je významný směr současné biochemie díky úzkému propojení reálně dosažitelných a populovaných konformací a funkcí biomolekul.[62] Správný konformační sampling může napomoci přesným predikcím řady důležitých aspektů proteinové dynamiky, jako jsou strukturní, kinetické a termodynamické veličiny, a trajektorie sbalování proteinu (proteinového foldingu). Protože biomolekuly se v živých organismech zpravidla nachází ve stavu roztoku, což výrazně mění jejich fyzikální chemii v porovnání s molekulami v plynné fázi, je termodynamicky správný přístup je hledat minima Gibbsovy energie celého systému [63], tedy proteinu včetně explicitně či implicitně zahrnutého solventu (obvykle vody). Nejjednodušším typem konformačního samplingu je náhodné hledání (random search methods), jehož podstatou je generování nových konformerů náhodnou změnou souřadnic či dihedrálních úhlů již získaných a minimalizovaných struktur, což pokračuje obvykle tak dlouho, dokud procesem již nevznikají žádné nové konformery. Tato metoda je sama o sobě snadno implementovatelná, nicméně není jisté, jak velké procento konformačního prostoru uživatel získá, a navíc většinu struktur vygeneruje (zbytečně) vícekrát [41].

Druhým základním typem samplingu je systematické hledání. V tomto případě se výchozí struktury volí analyticky tak, aby pokryly celý konformační prostor rovnoměrně, např. ve formě čtvercové mříže. Po jejich následné optimalizaci tyto struktury zkonvergují do nejbližšího energetického minima. Podobně jako u náhodného samplingu je tak většina výsledných konformerů redundantních, nicméně oproti náhodnému samplingu lze do velké míry ovlivnit, jak moc úplný konformační prostor uživatel získá, volbou hustoty pokrytí konformačního prostoru výchozími strukturami. Tato větší jistota byla i přes vysokou výpočetní

náročnost a redundanci hlavním důvodem, proč byl tento typ samplingu zvolen pro hlavní náplň práce.

Dalším možným přístupem je metoda postupné výstavby, během které jsou vždy nalezeny lokální minima krátkých fragmentů. Tyto fragmenty se pak spojí, opět jsou energeticky minimalizovány, a tak dále, až do celkové struktury, jejíž energetická minima se hledají. Podobným přístupem je deformační sampling, kde je díky vyřešení difuzní rovnice hyperplocha potenciální energie v hladší podobě, což umožňuje snadnější nalezení hlubších minim odstraněním těch mělčích [64].

Podobným přístupem je metoda distanční geometrie [4],[5], jejíž podstatou je náhodné generování geometrických matic. Tyto matice vždy obsahují vzdálenosti mezi atomy (takže hlavní diagonála je vždy nulová). Protože vzdálenosti atomů v molekule jsou spřažené a platí pro ně řada omezení (například pro lomenou molekulu ABC není možné, aby vzdálenost atomů A a C byla větší než součet AB a BC, atomy v cyklických molekulách od sebe nemohou být dále než je určitá hranice atd.), zdaleka ne všechny matice je třeba převádět na skutečné konformery a minimalizovat je, a nereálnost potenciálního konformeru lze odhalit už ve formě matice. Minimalizuje se pak jen málo skutečně vytvořených konformerů, opět však není nijak zaručeno, že nebude docházet ke generaci stejných konformerů vícekrát.

Pro potřeby konformačního samplingu pak lze použít i metody molekulové dynamiky či Monte Carlo – díky jejich základní charakteristice, totiž přípustnosti kroku směrem k vyšší energii [66], je možné jimi dosáhnout jinými metodami obtížně dosažitelných minim.

Metody Molekulové dynamiky (MD) jsou metody sloužící k modelování realistického časového vývoje studovaného systému [66]. Tento časový vývoj trajektorie systému lze spočítat tak, že určíme celkovou sílu působící na každou

částici, jako záporně vzaté parciální derivace celkového potenciálu (který nám poskytne použité silové pole) podle prostorových souřadnic daného atomu. Samotný časový vývoj pak získáme řešením soustavy Newtonových pohybových rovnic pro všechny částice:

$$m_i \frac{d^2\vec{r}_i}{dt^2} = -\sum_{i \neq j}^{N} \frac{\partial U_{ij}(|\vec{r}_i - \vec{r}_j|)}{\partial(|\vec{r}_i - \vec{r}_j|)} \tag{14}$$

Kde $m_i$ je hmotnost částice $i$ a $U$ označuje zvolený potenciál. Kromě toho je samozřejmě potřeba znát počáteční podmínku – konfiguraci v čase $t = 0$. Rovnice se řeší numericky, zpravidla tak, že spočítáme vývoj po krátký časový úsek, na kterém předpokládáme, že se působící síly nemění. Z tohoto nového bodu pak postupujeme stejně.

Jednou z mnoha metod molekulové dynamiky je tzv. „horká“ molekulová dynamika (hot-MD) při které se systém simuluje za nerealisticky vysokých teplot (nerealistických v tom smyslu, že např. studovaná molekula by takové podmínky nevydržela) díky čemuž dosahuje simulovaný systém vysokých hodnot energie, pročež dokáže překonat energetické bariéry, které by jinak byly překonatelné obtížně nebo vůbec.

V úzkém vztahu k MD samplingu je sampling Monte Carlo, založený na přijetí kroku k vyšší energii na základě tzv. Metropolisova kritéria [67], souvisejícího s Boltzmannovým rozdělením. Krok ze stavu 1 do stavu 2 je v případě zvýšení energie přijat, pokud pravděpodobnost přijetí $p$, definovaná pomocí energií obou stavů $E_1$ a $E_2$ jako:

$$p = min\left[1, \quad \frac{\exp\left(-E_1/_{kT}\right)}{\exp\left(-E_2/_{kT}\right)}\right] \tag{15}$$

je větší nebo rovna náhodně vybranému číslu z intervalu od nuly do jedné. Znaky $k,T$ zde klasicky mají význam Boltzmannovy konstanty a termodynamické teploty. Monte Carlo patří k nejvyužívanějším metodám samplingu, zejména pro případ kapalin či roztoků [45] a používá jej například i Rosseta [32].

Konformační samplingy typu Monte Carlo či molekulové dynamiky patří k nejrozšířenějším typům samplingu [45], zejména jeho vysokoteplotní varianta – díky tomu dosahují konfomery vysokých hodnot kinetické energie a jsou schopny přejít v podstatě jakoukoliv bariéru. Tento typ samplingu je v této práci užit pod názvem PlainMD, vizte dále.

Konkrétních metod založených na výše uvedených přístupech je celá řada, každá vyvíjená s ohledem na určité specifické využití. Jedním příkladem je large-scale low-mode samplingový algoritmus (LLMOD) [68], vyvinutý pro makromolekuly. Principem LLMOD algoritmu je, že se v každém kroku vybere určitý konformer z už známého setu (v prvním kroku vychází ze vstupní struktury), který se pak výrazně poruší podél směru jednoho z nízkofrekvenčních normálních vibračních módů. Takto vytvořený konformer je poté energeticky minimalizován a přidán k sadě už získaných konformerů. Energetická minimalizace však probíhá bez výpočtu Hessovy matice pomocí metody ARPACK, která na ni pouze nepřímo závisí. Tento LLMOD algoritmus se osvědčil pro konformační sampling makrocyklických sloučenin [69].

## 2.3. Použitý software a algoritmy

V rámci této práce byl použit jako hlavní nástroj pro provádění kvantově-chemických výpočtů software Turbomole, jehož původní verze se datuje do roku 1989. Jedná se o program s implementovanou velkou řadou současných QM i DFT metod, jejich různých variant a jeho výstupní data jsou podporována

velkou řadou dalších programů, a je využitelný pro celou řadu aplikací, jako je teoretický popis katalýzy, organické chemie, spektroskopických parametrů i biochemie.

Důvodem volby Turbomolu pro tuto práci je fakt, že se osvědčil při paralelizaci velkého množství výpočtů, jako jsou například optimalizace statisíců oligopeptidových fragmentů, což byla hlavní výpočetní náplň této práce. Velmi efektivně je v Turbomolu implementována tzv. „resolution-of-the-identity“, někdy též nazývaná „density-fitting“, aproximace DFT, která v praxi vede k řádovému urychlení DFT výpočtů.

Jak již bylo zmíněno, studium konformačního chování peptidových řetězců pozbývá smyslu bez modelování v roztoku, tedy bez zahrnutí efektů rozpouštědla. Rozpouštědlo může být modelováno tak, že jsou kolem zkoumané molekuly (v tomto smyslu označované jako rozpouštěná látka) simulovány i jednotlivé molekuly rozpouštědla, a to buď pomocí QM nebo MM metod. Takový model se pak označuje za explicitní solvatační model, a jakkoliv je tento přístup velmi přesný, v praxi je jeho výpočetní náročnost obrovská. Proto se v řadě případů používá tzv. implicitní modely solvatace, které simulují vliv rozpouštědla, resp. jeho elektrostatické interakce tak, že vytváří kolem molekuly kavitu o určité hodnotě relativní permitivity $\varepsilon_{\mathrm{r}}$, a rozměr této kavity je odvozen z Van der Waalsovských poloměrů. Jednodušší implicitní modely popisují kavitu pomocí elektrického náboje, Kirkwoodův-Onsagerův model pak pomocí dipólového momentu. Velkým zlomem byla na počátku 80. let formulace metod polarizovaného kontinua (PCM), které v té či oné podobě dodnes tvoří základ všech moderních implicitních solvatačních metod. V této práci byla v rámci programu Turbomole [70] byla užita QM metoda COSMO (conductor-like screening model) a COSMO-RS (conductor-like screening model for realistic solvation). COSMO na rozdíl od dalších implicitních modelů odvozuje polarizační náboje dielektrického kontinua z tzv. aproximace „stíněného vodiče“.

COSMO-RS metoda je schopna poskytnout chemický potenciál v kapalině a další termodynamicky důležitá data, a tedy i solvatační energii [71]. Implementace metody COSMO-RS je dostupná v programu COSMO*therm*.

Kromě výše uvedené kombinace (Turbomole, COSMO*therm*) byly použity programy (moduly) MAESTRO a Macromodel od firmy Schrodinger, Inc. přesněji řečeno jeho samplovací algoritmus MD/LLMOD (molecular dynamics with enhanced sampling along the low-lying vibrational modes) [68].

## 2.4. Cíle práce

Konformačním samplingem (vzorkováním) aminokyselin a dipeptidů se v posledních 20 letech se zabývá dlouhá řada studií. Jedná se o náročný a obtížně dosažitelný cíl, z důvodů jak technických a praktických, jako je značná výpočetní náročnost, tak principiálních, protože neexistuje způsob, jak úplnost konformačního prostoru jakkoliv dokázat. Základní způsoby mapování konformačního prostoru, uvedené výše a podrobněji popsané v [41], byly rozvíjeny a zdokonalovány nejprve na jednodušších systémech, jako je dialanin [72], [73] které na tomto modelovém dipeptidu ukazují, že dostatečný konformační sampling pro dosažení všech hlavních regionů lze na dipeptidech provést více způsoby, jako pomocí MD, tak tzv. metody Leap-Dynamics či dalších. Dále z nich vyplývá, že různé formy aproximací ovlivňují konformační preference různě, například že zahrnutí vlivu náboje posouvá konformační preference více k α-helix regionu, kdežto uvažování implicitního rozpouštědla či jeho pominutí podobný efekt neindukuje. Kromě toho byl dialanin studován i s použitím modelu explicitní vody, k čemuž se další přístupy často odvolávají jako ke srovnávacímu standardu. Dále byly v prostředí implicitního rozpouštědla (vody) konformačně vzorkovány i proteinogenní aminokyseliny [72], [74], [75]; a to jak volné, tak s chráněnými konci methylovou skupinou,

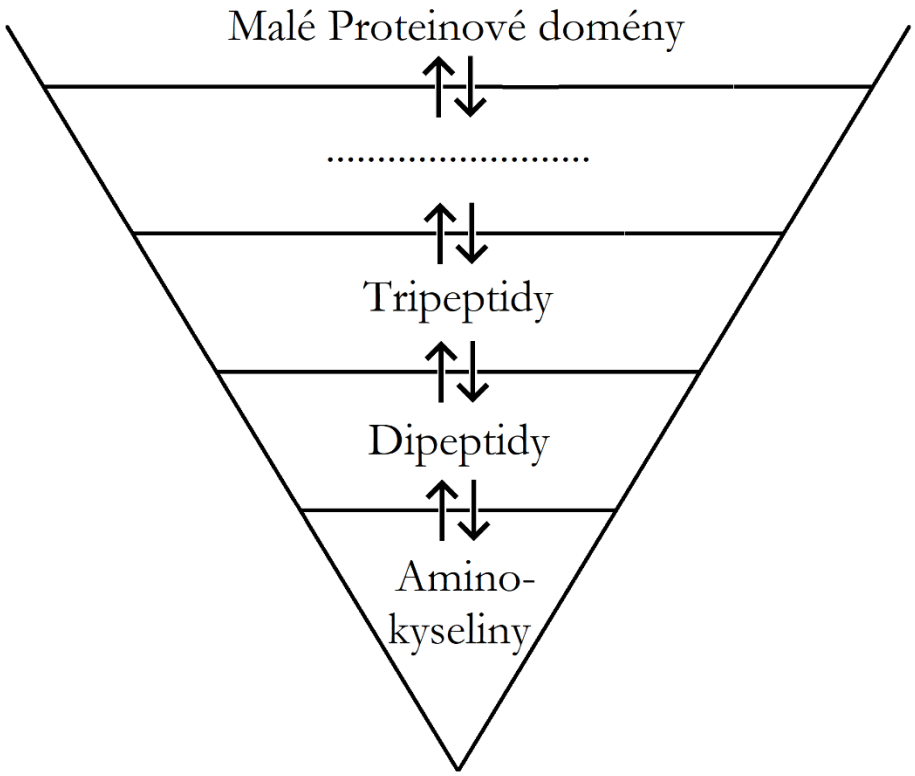
s nabitými i nenabitými postranními řetězci aspartátu, glutamátu, argininu, lysinu a histidinu (obě pozice) a také v komplexech s ionty $Ca^{2+}$ , $Ba^{2+}$ , $Sr^{2+}$ , $Cd^{2+}$ , $Pb^{2+}$ , and $Hg^{2+}$ (tyto komplexy jsou běžné v živé přírodě) při dobré shodě s experimentálními daty. V jiné studii byl vzorkován konformační prostor jednotlivých aminokyselin s konci chráněnými dvěma glyciny (GGXGG, kde „X“ označuje aminokyselinu) a také i nepřirozeně se vyskytující D-aminokyseliny, čímž bylo zjištěno, že konformační prostor je z hlediska změny chirality lichý (souměrný podle počátku). Konečně bylo konformační chování aminokyselin zkoumáno i v prostředí explicitní vody [73]. Dále pak bylo zkoumáno více systémů, jako například dipeptidy obsahující fenylalanin, ovšem v plynné fázi a s nenabitým postranním řetězcem, což ukázalo jistá pravidla pro typické hodnoty dihedrálních úhlů druhého postranního řetězce daná rigiditou fenylalninové části [76]. Konečně byly na vybraných aminokyselinách a dipeptidech v omezené míře zkoumány možnosti indukčních kroků, tedy zda je možné odvodit konformační prostor dipeptidů z aminokyselin [77], případně tripeptidů z dipeptidů [78]. Byla nalezena metoda, která *do určité míry* predikuje konformační prostor dipeptidů ze znalosti konformačního prostoru aminokyselin, použitím představy tzv. důležitých konformerů a „skládáním“ dipeptidů z těchto konformerů (prostým nahrazením, kde byly opět analogicky definované důležité konformery nalezeny. Metoda byla vyvinuta na 13 dipeptidech a s úspěchem použita na dalších 8. Následně byla její analogie používána i na přechod mezi konformačními prostory dipetidů a tripetidů. Nejednalo se však ani zdaleka o úplný konformační prostor, a navíc výpočty byly provedeny v plynné fázi. Na druhou stranu však teoretická infračervená spektra získaných konformerů poskytla dobrou shodu s experimentálními daty. Celkově však lze konstatovat, že bylo až dosud podrobeno extenzivnímu konformačnímu samplingu jen málo dipeptidů (např. TG [79]), a ještě méně delších peptidů či peptidů se složitějšími postranními řetězci [77].

Nicméně, velmi málo pozornosti bylo věnováno souhře *všech* faktorů určujících konformační chování peptidů v roztoku či v kontextu proteinu: extenzivnímu vzorkování, vlivu náboje postranního řetězce, větší variabilitě postranních řetězců, ochránění koncových skupin (které v rámci proteinu jsou součástí vazeb) zahrnutí vlivu rozpouštědla, ideálně pokročilejším solvatačním modelem atd. Větší aproximace v jakékoliv z těchto oblastí (například absence extenzivnějšího samplingu, či výpočty prováděné pouze v plynné fázi) nutně vedou k ovlivnění získaného konformačního prostoru. Za zmínku také stojí, že mnoho podrobnějších studií v tomto oboru se zabývá pouze velmi jednoduchými dipetidy, nejčastěji dialaninem.

Cílem této práce je použitím teoretických metod kvantové a výpočetní chemie zmapovat úplný konformační prostor aminokyselin a vybraných reprezentativních modelových dipeptidů (oboje s chráněnými konci) a to ve vodném prostředí modelovaným implicitním solvatačním modelem. Výsledný soubor dat nebude pouhou podmnožinou či aproximací, nýbrž limitně úplným konformačním prostorem studovaných systémů, u kterého je obtížné si představit, že by jej bylo možné získat jakýmkoliv běžným konformačním algoritmem. Získání takového souboru bude dosažitelné pouze náročnou a extenzivní lidskou i výpočetní prací, zahrnující geometrickou optimalizaci a výpočty energie statisíců až milionů molekul (kvantově-mechanickými metodami), dobře navrženou paralelizací výpočtů, stejně jako pokročilého skriptování.

Následně pak bude získaný soubor dat podroben zevrubné analýze, přičemž budou hledány jakákoliv pravidla a trendy, která konformační chování krátkých peptidických fragmentů určují či ovlivňují (např. vliv sousední aminokyseliny na konformační prostor), a bude proveden pokus o zobecnění těchto pravidel. Zároveň budou stejné dipeptidy podrobeny i vzorkování pomocí samplovacích algoritmů, za účelem otestování jejich kvality.

Konečně budou prozkoumány možnosti, jak lze konformační trendy v dipeptidech uchopit a udělat pomocí nich indukční kroky na větší oligopeptidy jak naznačují práce [77], [78], tj. jak z konformačního chování dipeptidů udělat krok k popsání konformačního chování tripeptidů, až do rozměru malých proteinových domén, vizte obrázek 2.2., je-li takový krok možný.




Obrázek 2.2.: Schematické znázornění inverzní pyramidy konformačních prostorů a myšlenky indukčních kroků mezi nimi.

Odvození podobných indukčních kroků by vedlo k výraznému posunu v chápání, jak lokální vlivy určují stabilizaci proteinu, a tedy k lepšímu chápání proteinového foldingu z prvních principů (*ab initio*). To pak může být využito v řadě aplikací, například teoretickém návrhu léčiv.

III. Kapitola

# Výpočetní protokol

## 3.1. Zkoumané systémy

Samotnou výpočetní práci lze rozdělit do čtyř částí:

(i) zmapování konformačního prostoru všech 20 jednotlivých aminokyselin,

(ii) zmapování konformačního prostoru 17 modelových dipeptidů metodikou vyvinutou na případu aminokyselin,

(iii) zmapování konformačního prostoru vybraných dipeptidických fragmentů z proteinové databanky,

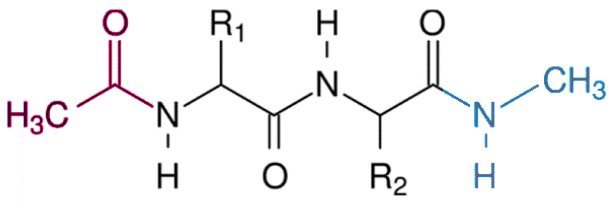
(iv) zmapování konformačního prostoru týchž dipeptidů za použití konformačních samplovacích algoritmů, jmenovitě MD/LLMOD od firmy Schrödinger, Inc. a metodou horké molekulové dynamiky označované v rámci této práce PlainMD.

Výsledky z první a druhé části byly zároveň použity při optimalizaci výpočetního protokolu třetí, resp. čtvrté části.

### 3.1.1. Zmapování konformačního prostoru aminokyselin

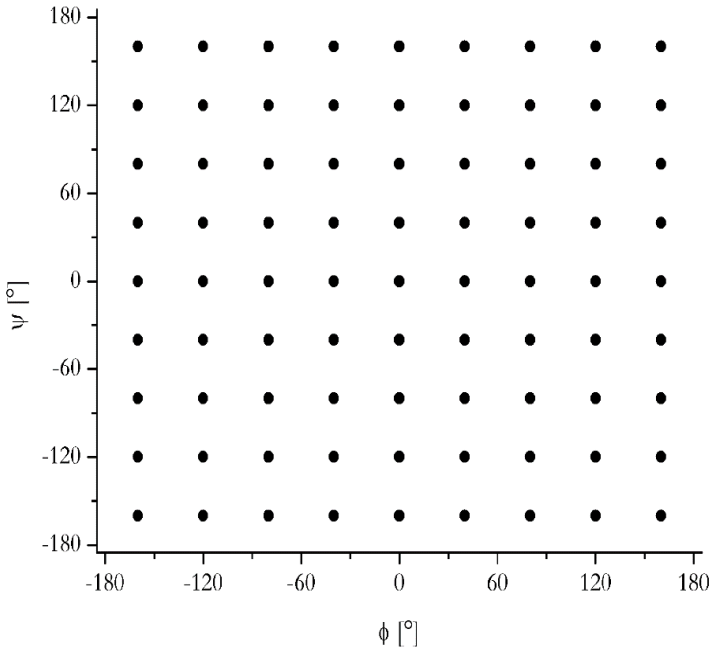
Jako modelový systém pro všechny studované oligopeptidy byl zvolen systém s chráněnými konci, tedy acetylovaný na *N* konci, a *N*-metylovaný na *C* konci, vizte obrázek 3.1. Zvolený model tak zahrnuje nejbližší okolí dané aminokyseliny v rámci proteinového řetězce, a je věrnější reprezentací oligopeptidu jako části

proteinové struktury než samotná volná aminokyselina ve „zwitteriontové“ podobě.



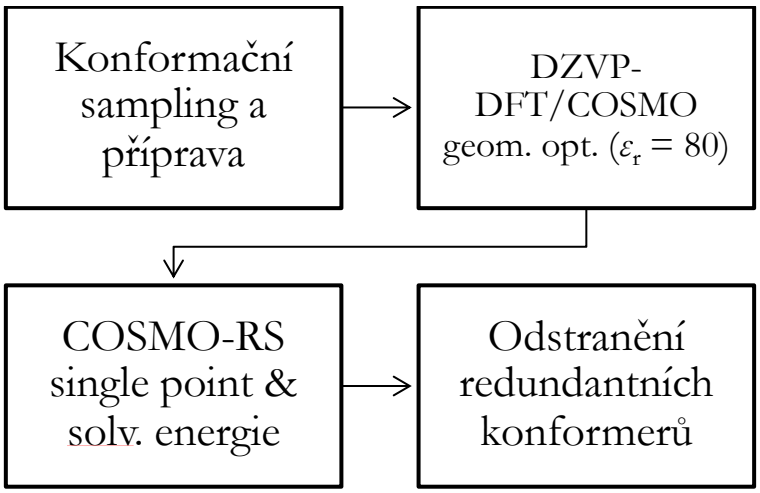

Obrázek 3.1.: Schematické znázornění oligopeptidu s chráněnými konci.

Pro získání či popis úplného konformačního prostoru bylo nutné minimalizovat možnost, že v rámci samplovacího kroku dojde k pominutí určitých konformerů. Proto byl sampling proveden analyticky, a to následovně – všechny dihedrální úhly vazeb mezi těžkými atomy v dané aminokyselině (jmenovitě $\psi$, $\varphi$, $\chi_1$, $\chi_{2,...}$ $\chi_n$) byly rotovány vždy o 40°, čímž bylo získáno $9^N$ výchozích struktur (kde $N$ je počet dihedrálních úhlů). Peptidová vazba je obecně vzhledem k její násobnosti považována za vazbu s nulovou rotační volností, proto byla ponechána fixní. Tím bylo dosaženo rovnoměrného počátečního pokrytí konformačního prostoru, jak ilustruje obrázek 3.2. Tento typ samplingu a výsledky výpočtů po jeho použití budou dále nazývány *systematický* sampling.



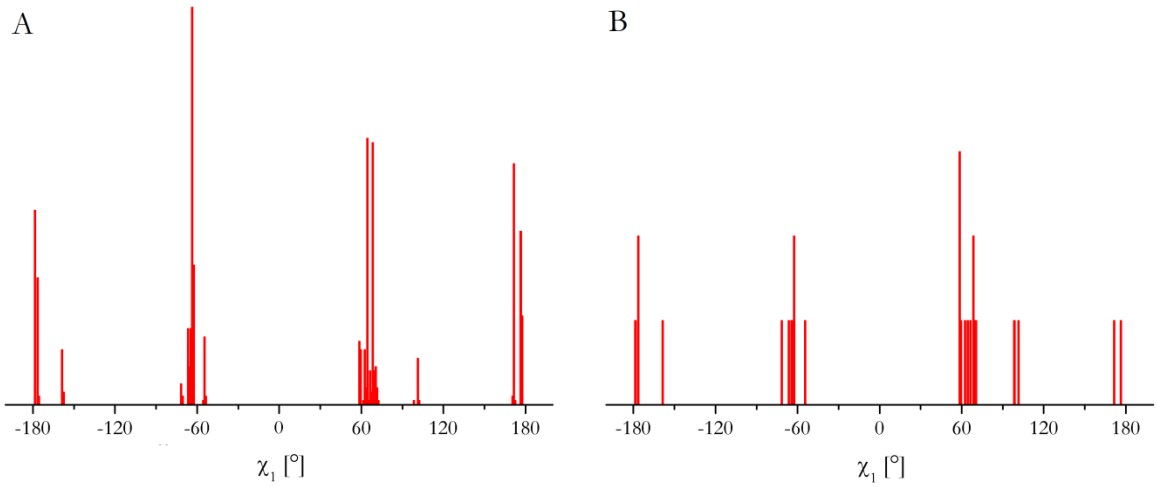

Obrázek 3.2.: Způsob systematického samplingu $N$-rozměrného (zde dvourozměrného) konformačního prostoru na ukázce alaninu.

Následně byly všechny výchozí konformery geometricky zoptimalizovány v programu Turbomole způsobem popsaným v odstavci 3.2.1. a následně byla vypočtena jejich solvatační energie, dle odstavce 3.2.2. Celý výpočetní protokol je znázorněn na obrázku 3.3.




Obrázek 3.3.: Schéma výpočetního protokolu použitého v případě aminokyselin.

Z průběžných výsledků bylo zjištěno, že generace konformačních setů rotací o 40° není nutná pro dihedrální úhly vazby mezi atomy nabývající hybridizace $sp^3$. V těchto případech jsou v naprosté většině populovány pouze antiperiplanární a obě *gauche* konformace vzhledem k této vazbě (vizte obrázek 3.4.), a tedy postačuje rotace těchto vazeb o úhel 120°, čímž se počet výchozích struktur významně redukuje. Navíc i v případě této aproximace byly ve výsledném souboru stejné hodnoty dihedrálního úhlu postranního řetězce jako bez ní, což ji činí zcela legitimní.




Obrázek 3.4.: Ukázka výsedných hodnot úhlu postranního řetězce valinu pro rotaci o 40° před smazáním redundancí (A) a 120° po smazání redundancí (B).

Rotace o 120° však není možná v případě vazby na atom s hybridizací $sp^2$, což se týká aminokyselin Asp, Asn, His, Glu a Gln, kde by toto omezení vedlo k výrazné ztrátě počtu finálních struktur a neúplnosti konformačního prostoru. Je však dostatečná pro Tyr, Phe a Trp, kde sterická náročnost postranního řetězce dovoluje ve většině jen dvě polohy postranního řetězce vzhledem k rotaci této vazby (obě lišící se o 180°) a menšina zbylých dihedrálních úhlů na této vazbě je také tímto postupem získána.

Získané finální konformační sety jsou vysoce redundantní (vizte odstavec 4.3), a tedy byl vyvinut algoritmus pro efektivní výběr unikátních konformerů pro další vyhodnocení. Tento algoritmus je založen na představě, že pouze konformery

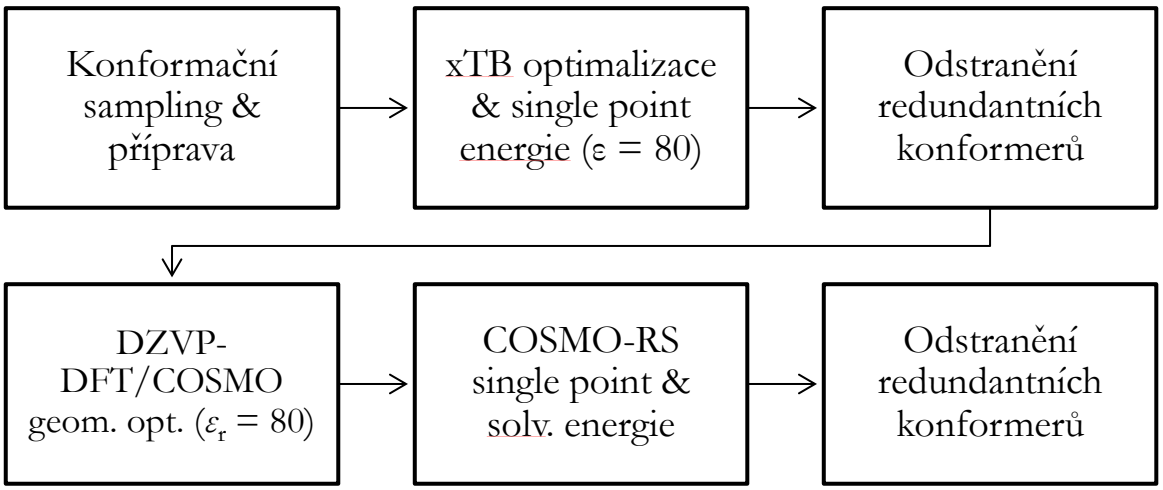
lišící se v alespoň jednom dihedrálním úhlu o více než 20° jsou neredundantní, protože 20° je přibližná vibrační amplituda vazeb mezi těžkými atomy. Celý skript algoritmu, napsaný v prostředí programovacího jazyka python [80],[81], je uveden v Dodatku A k této práci.

### 3.1.2. Zmapování konformačního prostoru 17 modelových dipeptidů

Nejprve byl zmapován úplný konformační prostor 17 modelových dipeptidů, které byly vybrány tak, aby obsahovaly jednak všechny typy postranních řetězců (polární, nepolární, kladně a záporně nabitý a aromatický), dále postranní řetězce stejného typu ale odlišných v délce (např. Asp a Glu) a konečně i permutace aminokyselin v dipeptidu (např. AV a VA). Způsob generace konformačních setů zůstal stejný jako způsob vyvinutý pro případ aminokyselin (vizte výše, odstavec 3.1.1.), s opět rotačně fixovanou *trans*-peptidovou vazbou mezi oběma aminokyselinovými rezidui. Seznam všech dipeptidů včetně počtu výchozích konformerů je uveden v tabulce 3.1.
Následně byl na kratších dipeptidech proveden stejný postup jako v případě aminokyselin.
I přes paralelizaci na 16, resp. 24 procesorech na používaných výpočetních klastrech však čas nutný na zmapování konformačního prostoru byť i kratších dipeptidů (AV, IA) dosahoval neúnosných hodnot v řádu týdnů na jeden systém, což vedlo k nutnosti snížit vhodným způsobem počet konformerů, které je nutné optimalizovat kvantově-chemicky (jakožto rychlost určujícím kroku).
Z toho důvodu byla vlastní optimalizaci konformerů metodou DFT-D3//COSMO předřazena ještě optimalizace semiempirickou metodou xTB (citace, detaily výpočtu vizte kapitolu 3.2.1.), která je oproti optimalizaci DFT-D3//COSMO přibližně 100x rychlejší (jednotky sekund oproti desítkám minut až hodinám) a následné vyřazení redundantních konformerů po optimalizaci v xTB (opět stejným algoritmem, vizte Dodatek A). Až teprve konformery

unikátní po optimalizaci v xTB byly optimalizovány metodou DFT-D3//COSMO, neboť dle referenční studie [56] je samotná metoda xTB pro výpočet konformačních energií výrazně horší než DFT-D3. V našem případě to vedlo je snížení počtu konformerů vstupujících do DFT-D3//COSMO o 90 %, tedy výpočetní náročnost byla de facto snížena o řád. Tento přístup byl otestován na dipeptidech AA a AD a bylo ověřeno, že se jedná o dobrou aproximaci.

Následně pak pokračoval standardní protokol použitý v případě aminokyselin, vizte obrázek 3.5. Takto byly získány konformační prostory 17 modelových dipeptidů a počty výsledných konformerů byly dále použity pro otestování algoritmů LLMOD a PlainMD, vizte dále.




Obrázek 3.5.: Schéma výpočetního protokolu použitého v případě dipeptidů. Detaily jednotlivých kroků vizte kapitolu 3.2.

Tabulka 3.1.: Studované dipeptidy a počty jejich výchozích konformerů.

| Dipeptid | AA | AV | VA | AS | SA |
|---|---|---|---|---|---|
| Počet | 6561 | 19 683 | 19 683 | 19 683 | 19 683 |
| AI | IA | AD | DA | AF | AH |
| 59 049 | 59 049 | 177 147 | 177 147 | 19 683 | 59 049 |
| AN | VV | AQ | IV | AE | AK |
| 177 147 | 59 049 | 531 441 | 177 147 | 531 441 | 531 441 |

Na některých dipetidech byly rovněž sledovány vodíkové vazby. Pro potřeby této studie byla vodíková vazba definovaná jako současné splnění dvou podmínek: 1) Vzdálenost vodíku a akceptoru musela být větší než 2,5 Å a 2) Úhel donor – atom vodíku – akceptor větší než 130°. Vodíkové vazby jsou uvažovány mezi skupinami CO a NH hlavního řetězce a mezi polárními skupinami postranních řetězců a hlavním řetězcem. Pro hledání vodíkových vazeb byl použit skript napsaný s využitím knihovny *Biopython*, užitý již dříve v práci [82].

### 3.1.3. Zmapování konformačního prostoru dipeptidových fragmentů z proteinové databanky

Souběžně s výpočty popsanými v kapitole 3.1.2. byly pro potřeby vyhodnocení šířky energetického okna konformerů sledovaných dipeptidů vyextrahovány jejich struktury (konformace) z databáze Top8000 [82], jmenovitě AA, AV, VA, AD, DA, AS, SA, AF, AK. Tyto dipeptidové fragmenty byly získány tak, že místo $C_\alpha$ uhlíku hypotetické předchozí a následné aminokyseliny v proteinovém řetězci byla substituována methylová skupina, čímž byl dipeptidový fragment v konformaci z reálné proteinové struktury převeden na dipeptid s chráněnými konci, plně odpovídající modelu získanému generací systematickým samplingem.

Následně byly konformery podrobeny stejnému výpočetnímu protokolu jako v případě aminokyselin a kratších dipeptidů.

### 3.1.4. Zmapování konformačního prostoru dipeptidů pomocí samplovacích algoritmů

Jako poslední část byly na zde studovaných dipeptidech (vizte tabulku 3.1.) otestovány dva samplovací algoritmy, jmenovitě LLMOD (Maestro/MacroModel) a PlainMD; pro ověření spolehlivosti a kvality těchto samplovacích algoritmů z důvodu možného použití pro určení konformačních prostorů větších oligopeptidů. Podrobnější popis obou algoritmů lze nalézt v odstavci 2.3.

V případě metody LLMOD byl použit výchozí protokol implementovaný ve verzi z roku 2015, konkrétně force field OPLS2005, solvatační model GB/SA, RMSD 0,75 Å, energetické okno 10 kcal·mol$^{-1}$ pro přijetí daného konformeru. Jediné, co bylo měněno, byl počet očekávaných výsledných struktur, který byl volen s ohledem na v té chvíli už známý počet výsledných struktur po systematickém samplingu a následných výpočtech (vizte tabulku 3.2.). Počty vygenerovaných konformerů byly jen přibližný, protože součást MD/LLMOD je i kontrola a odstranění redundancí ihned po vygenerování konformerů a tedy není možné výchozí počet zvolit přesně. Podrobnější popis detailů LLMOD může být nalezen v [68].

V případě PlainMD bylo použito stejné nastavení, jako ve studii [83]: 5 ns molekulová dynamika při 1000 K v programu NAMD verze 9.2 [84] se silovým polem AMBER ff99SB [85] a implicitním modelem vody GBIS [86], přičemž snímky, ze kterých byly vyextrahovány jednotlivé konformery, byly pořízeny v takových intervalech, aby se počty konformerů blížily počtům neredundantních konformerů získaných systematickým samplingem.

Následně byly konformery podrobeny stejnému výpočetnímu protokolu, jako v případě aminokyselin (vizte obrázek 3.3.; nízký počet konformerů umožnil vynechat aproximativní krok s metodou xTB).

Tabulka 3.2.: Počty výchozích konformerů vygenerovaných LLMOD a PlainMD.

| Dipeptid | AA | AV | VA | AS | SA |
|---|---|---|---|---|---|
| PlainMD | 350 | 350 | 350 | 350 | 350 |
| LLMOD | 329 | 357 | 367 | 387 | 340 |
| AI | IA | AD | DA | AF | AH |
| 1000 | 1000 | 600 | 600 | 400 | 600 |
| 891 | 845 | 478 | 528 | 433 | 503 |
| AN | VV | AQ | IV | AE | AK |
| 1000 | 800 | 3000 | 3000 | 3000 | 9000 |
| 910 | 893 | 3684 | 3489 | 2614 | 8954 |

## 3.2. Praktické aspekty použitých metod

### 3.2.1. Geometrická optimalizace

Geometrická optimalizace metodou GFN2-xTB byla ve všech případech provedena v programu *xtb*, konvergenčními kritérii pro energii a gradient $0{,}5 \cdot 10^{-5}$ a.u., resp. $1 \cdot 10^{-3}$ a.u./α, v implicitním vodném rozpouštědle (při 298,15 K a o permitivitě $\varepsilon_r$ = 80,2) a odpovídajícím nábojem. Všechny ostatní veličiny byly ponechány ve výchozím nastavení (citace manuál XTB)

Geometrická optimalizace DFT-D3 byla provedena pomocí programu Turbomole, v bázi DZVP-DFT, s parametry disperze D3-BJ použité v [87] v prostředí COSMO [71] ($\varepsilon_r$=80), s funkcionálem BP86 [50] a kritériem konvergence pro energii $1{,}0 \cdot 10^{-6}$ a.u.

### 3.2.2. Výpočet energie

Výpočet (vylepšeného odhadu) Gibbsovy volné energie v roztoku se sestával ve všech případech ze dvou postupných kroků. První byl výpočet *single point* energie pomocí programu Turbomole v COSMO – všechny parametry byly stejné jako v případě geometrické optimalizace, až na permitivitu, které byla přiřazena hodnota nekonečno. Druhým krokem byl výpočet solvatační energie pomocí COSMO-RS za použití programu COSMOtherm17 a parametrizace BP-TZVPD-FINE, opět ve vodě jako rozpouštědle. Pro potřeby ospravedlnění aproximace zanedbání frekvenčního příspěvku k energii byl na dvou aminokyselinách spočítán i frekvenční příspěvek k energii pomocí programu *thermo*, [88] tedy standardní termochemická analýza modelem harmonického oscilátoru a tuhého rotoru obohacená o model volného rotoru pro nízko-ležící vibrační frekvence ($< 100$ cm$^{-1}$). Výsledná energie konformeru byla vyjádřena jako součet těchto dvou příspěvků:

$$G_{CELK} = G_{SOLV} + E_{S.P.} \tag{16}$$

Kde $G_{CELK}$ je celková (absolutní) volná energie uvažovaného konformeru, $E_{S.P.}$ *single point* energie vypočtená pomocí COSMO, a $G_{SOLV}$ solvatační energie získaná pomocí COSMO-RS. Pro potřeby vyhodnocení byly následně získány i relativní energie konformerů vůči globálnímu minimu odečtením energie globálního minima od všech konfomerů:

$$G_{CELK,REL} = G_{CELK} + G_{CELK,GMIN} \tag{17}$$

Kde $G_{CELK,REL}$ je celková relativní volná energie diskutovaného konformeru vzhledem ke konformeru v globálním minim daného dipeptidu a $G_{CELK,GMIN}$ je celková absolutní volná energie konformeru v globálním minimu. Relativní

energie konformeru v globálním minimu byla tedy rovna nule, všechny ostatní konformery měly hodnotu vyšší.

IV. Kapitola

# Výsledky a diskuze

Před samotným zahájením mapování konformačního prostoru bylo třeba najít dostatečně přesnou, ale výpočetně únosnou metodiku. Solvatační model COSMO-RS, použitý v této práci, je totiž kalibrován pro funkcionál BP86 a triple-zeta bázový set: def2-TZVPD. Výpočty v takto velké bázi jsou nicméně velice náročné na výpočetní výkon. Pro studii statisíců až miliónů konformerů aminokyselin a dipeptidů provedenou v této práci bylo třeba najít výpočetně levnější alternativu. Z toho důvodu byly konformační sety alaninu a aspartátu získané rotací po 40° optimalizovány nejprve za použití funkcionálu BP86 a DZVP-DFT bázového setu a stejná úroveň byla použita pro následné COSMO-RS single point energie. Konformery byly následně přeoptimalizovány s užitím doporučené a parametrizované kombinace BP86/def2-TZVPD (pro COSMO-RS) a dále byl připočten frekvenční příspěvek k energii spočítaný použitím BP86/DZVP-DFT v plynné fázi. Obě výsledné energie byly porovnány. Pro obě aminokyseliny byla průměrná odchylka relativní energie vzhledem ke globálnímu minimu setu konformerů menší než 0,5 kcal·mol$^{-1}$, což je dostatečné pro tento typ studie a aproximaci to ospravedlňuje. Navíc takovýto výpočet energie je spojený s výraznou úsporou výpočetního času v řádech minut, až desítek minut oproti několika málo sekundám, což odpovídá cca dvousetnásobnému zrychlení.

## 4.1. Velikost konformačního prostoru aminokyselin

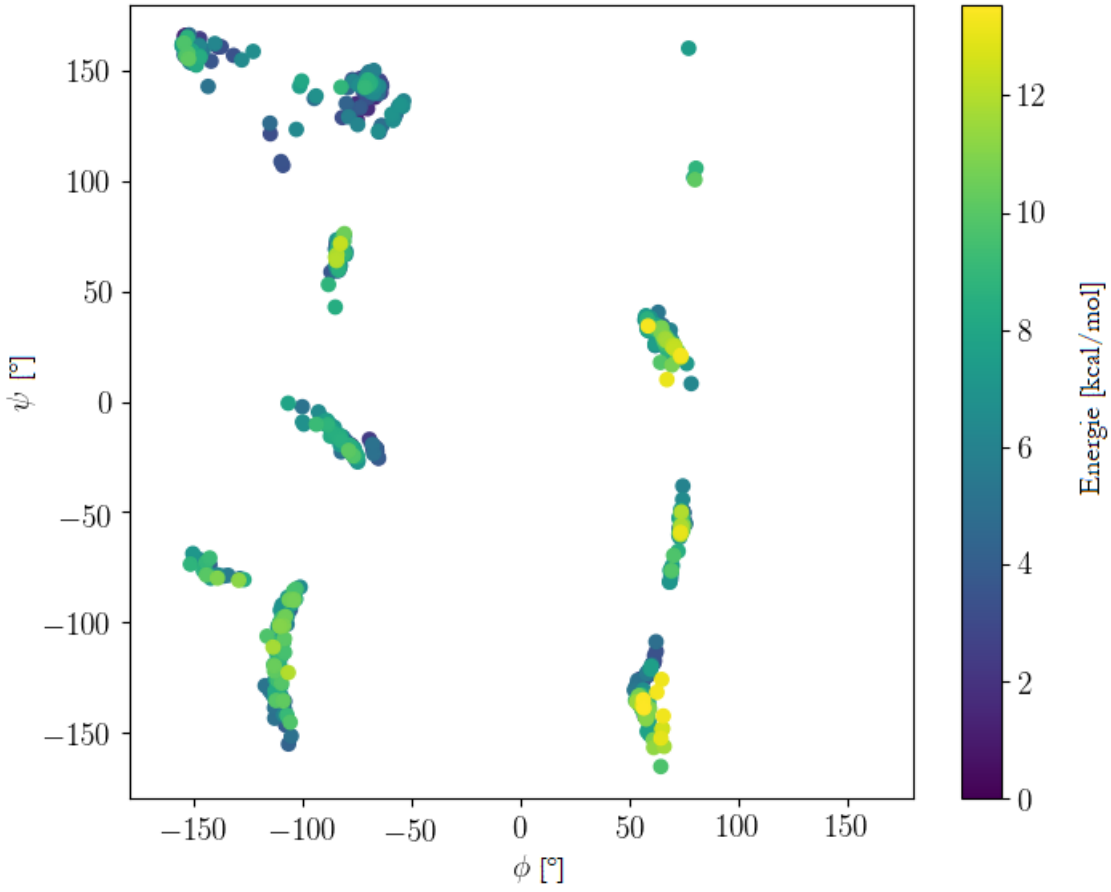
Způsobem popsaným v podkapitole 3.1.1. byly zmapovány konformační preference všech 20 proteinogenních aminokyselin. Počty výsledných unikátních konformerů jsou uvedeny v tabulce 4.1.

Tabulka 4.1.: Počty výsledných unikátních konformerů aminokyselin, počty získaných konformerů ve studii [75] a výchozí počty konformerů.

| Aminokys | Ala | Gly | Pro | Val | Ile | Leu |
|---|---|---|---|---|---|---|
| Unikátní | 7 | 9 | 3 | 26 | 112 | 114 |
| [75] | 5 | 8 | 22 | 23 | 76 | 85 |
| Výchozí | 81 | 81 | 9 | 729 | 2187 | 2187 |
| Phe | Thr | Cys | Ser | Met | Tyr | Trp |
| 37 | 57 | 41 | 74 | 283 | 46 | 82 |
| 25 | 47 | 31 | 59 | 246 | 43 | 57 |
| 486 | 243 | 243 | 243 | 6561 | 486 | 486 |
| Asn | Gln | Asp | Glu | His | Lys | Arg |
| 160 | 481 | 72 | 147 | 58 | 396 | 599 |
| 49 | 134 | 23 | 46 | 57 | 731 | 1218 |
| 2187 | 6561 | 2187 | 6561 | 2187 | 6561 | 19683 |

Protože konformační prostor aminokyselin byl získán již řadou jiných metod, cílem jeho zmapování v této práce je především zkalibrovat metodiku, používanou později na dipeptidy, aby dávala reálné a správné výsledky. Z porovnání výsledků v tabulce 4.1. s konformačními sadami již získanými jinou metodikou [75] vyplývá, že metodou použitou v této práci bylo dosaženo většího konformačního prostoru, než jsou konformační prostory získané v práci [75]. Jedinou výjimku tvoří konformační prostory lysinu, argininu a prolinu, nicméně rozdíl je způsoben volbou kritéria redundance, které je v případě studie [75] založeno na energetické odlišnosti, zatímco v případě této práce na odlišnosti strukturní. Energetické kritérium je totiž splněno i pro případ velmi podobných struktur, lišících se v jednotlivých dihedrálních úhlech jen o jednotky stupňů, tedy dvě struktury, v práci [75] vyhodnoceny jako neredundantní, jsou si navzájem velmi podobné.

Lze tedy shrnout, že výsledky mapování konformačního prostoru jednotlivých „chráněných“ aminokyselin ukazují, že použitá metodika je vhodná pro systematické mapování konformačního prostoru dipeptidů. Ve všech případech byly získány konformery zastupující všechny tři hlavní regiony na Ramachandranově diagramu (α-helix, β-list, další struktury, vizte obrázek 4.1.).




Obrázek 4.1.: Ukázka konformačního prostoru dipeptidu pro případ AI se znázorněním regoinů α-helix, β-list a dalších struktur.

## 4.2. Velikost konformačního prostoru dipeptidů

Pomocí námi vyvinutého výpočetního protokolu popsaného v kapitole 3.1.2. byl stanoven úplný konformační prostor 17 modelových dipeptidů. Dipeptidy byly vybrány tak, aby obsahovaly jednak všechny typy postranních řetězců (polární,

nepolární, kladně a záporně nabitý a aromatický), dále postranní řetězce stejného typu ale odlišných v délce (např. Asp a Glu) a konečně i permutace aminokyselin v dipeptidu (např. AV a VA). Pro histidin byla v této práci pro jednoduchost uvažována pouze kladně nabitá forma s nábojem formálně na ε-dusíku (termodynamicky mírně favorizovanější stav za běžných podmínek). Opět, ve všech případech byly nalezeny konromery ze všech hlavních regionů Ramachandranova diagramu.

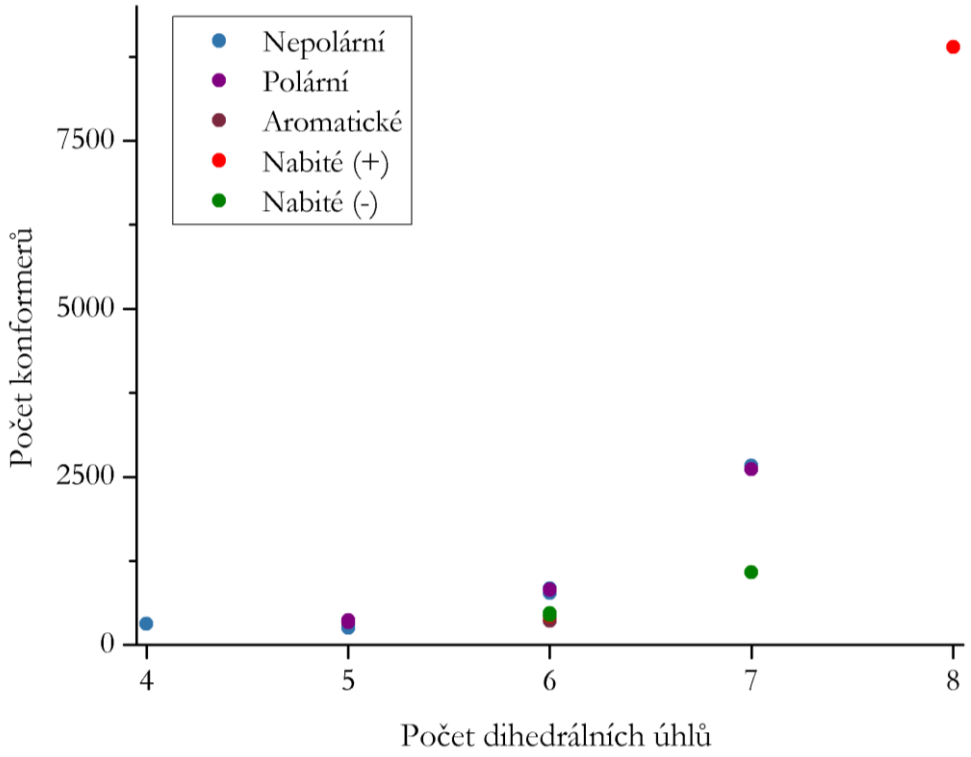
Získané konformační sety unikátních konformerů těchto dipeptidů se pohybují v rozmezí stovek až tisíců, jak shrnuje tabulka 4.2. a ilustruje obrázek 4.1. níže.

Tabulka 4.2.: Počty výsledných unikátních konformerů dipeptidů a procento redundance ve výchozím setu.

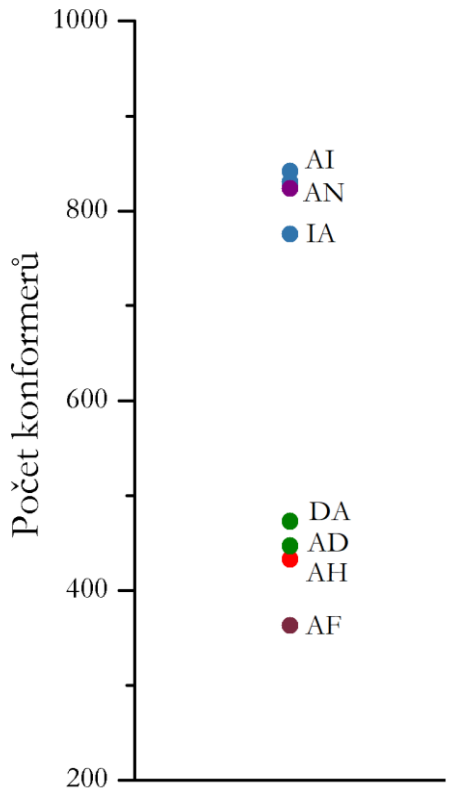
| Dipeptid | AA | AV | VA | AS | SA |
|---|---|---|---|---|---|
| Počet unik. | 312 | 254 | 256 | 367 | 340 |
| Počet vých. | 6561 | 19 683 | 19 683 | 19 683 | 19 683 |
| % | 95 | 98 | 98 | 98 | 98 |
| AI | IA | AD | DA | AF | AH |
| 839 | 828 | 444 | 470 | 360 | 430 |
| 59 049 | 59 049 | 177 147 | 177 147 | 19 683 | 59 049 |
| 98 | 98 | 99 | 99 | 98 | 99 |
| AN | VV | AQ | IV | AE | AK |
| 821 | 773 | 2614 | 2669 | 1083 | 8891 |
| 177 147 | 59 049 | 531 441 | 177 147 | 531 441 | 531 441 |
| 98 | 98 | 99 | 98 | 98 | 98 |

Ve všech případech se ukázalo, že přes 95% (často i 99%) konformerů je redundantních, tudíž vyvstává otázka, zda je možné nalézt efektivnější způsob konformačního samplingu, který povede ke stejnému nebo velmi podobnému výslednému konformačnímu setu pro všechny typy dipeptidů (včetně polárních, nabitých a aromatických). Tato myšlenka je podrobněji rozebrána v kapitole 4.4.

Počty konformerů rostou s počtem dihedrálních úhlů podle očekávání exponenciálně, vizte obrázek 4.2. a 4.3. Dále je z grafu patrný trend nižších počtů konformerů pro dipeptidy s nabitým postranním řetězcem, jako je AE, AD, AH, DA. Protože jsme zároveň ukázali, že konformační prostor jim podobných dipeptidů, jako např. AQ, AF, AN (obsahujících $sp^2$ hybridizaci na příslušném uhlíku zpravidla sousedícím s nabitou skupinou) je větší, je zřejmé, že se jedná o efekt způsobený nábojem a nikoliv hybridizací. Závěrem tedy je, že náboj na postranním řetězci proteinového rezidua významně redukuje jeho konformační prostor, zatímco samotný typ hybridizace podobný efekt neindukuje.




Obrázek 4.2.: Výsledné počty unikátních konformerů tvořící úplný konformační prostor zkoumaných dipeptidů v závislosti na počtu rotovatelných dihedrálních úhlů. Body jsou obarveny na základě vlastností dipeptidů: nepolární (modrá), polární (fialová), aromatické (hnědá), záporně (zelená) a kladně nabité (červená).




Obrázek 4.3.: Přiblížení obrázku 4.1 v oblasti pro 6 dihedrálních úhlů.

Z detailu na obrázku 4.2 je dále vidět efekt záměny různých postranních řetězců o stejném množství dihedrálních úhlů, které dále ukazuje, že konformační prostor očekávatelně snižuje i sterická náročnost postranního řetězce (jak lze vidět srovnáním izoterc-butylu u AI a fenylového rezidua u AF. Stejný efekt také vysvětluje, proč je v případě AA mírně více unikátních konformerů než pro AV či VA – v některých konformacích se dostávají methylová skupina alaninu a izopropylová skupina valinu příliš blízko sobě, a tudíž taková konformace není stabilní, přičemž tento problém odpadá u dvou méně stericky náročných methylových skupin AA. Konformační prostor hlavního řetězce nicméně zůstává větší pro AV/VA než pro AA, což však sterický efekt nevysvětluje a tento problém bude hlouběji řešen v další práci.

## 4.3. Šířka energetických oken dipeptidů a jejich srovnání s energiemi konformerů z proteinových struktur

Energie konformerů, tvořících úplné konformační prostory, které byly získány postupem popsaným v kapitole 3.2, byly vyhodnoceny ve formě histogramů. Protože v některých sadách jsme narazili na (domníváme se nepříliš významné) vysokoenergetické konformery, byla jako reprezentativnější hodnota zvolena energie na 95. percentilu (jinak také devatenáctém vigintilu, běžně značeno $Q_{19/20}$ či $p^{95}$), která lépe vystihuje skutečné energetické okno. Maximální šířka energetického okna spolu s hodnotou 95. percentilu a průměrnou hodnotou pro každý konformer je uvedena v tabulce 4.3.

Tabuka 4.3.: Šířky okna relativních konformační energií ($\Delta E_{\max}$) jejich hodnoty na 95. percentilu (oboje v kcal·mol$^{-1}$) pro sety unikátních konformerů získaných systematickým samplingem.

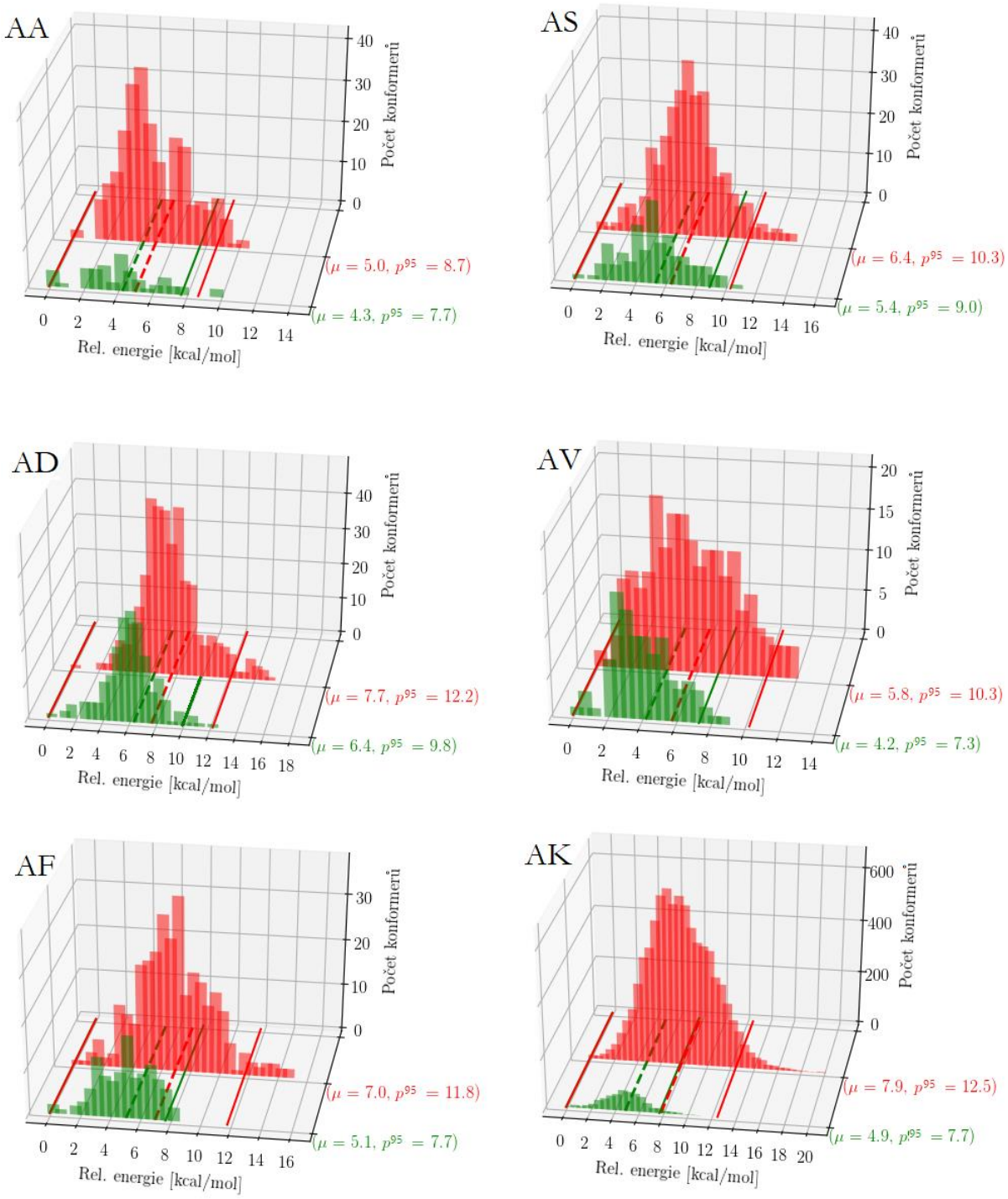
| Dipeptid | AA | AV | VA | AS | SA |
|---|---|---|---|---|---|
| $\Delta E_{\max}$ | 10.4 | 12.0 | 15.5 | 13.4 | 11.9 |
| $Q_{19/20}$ | 8.7 | 10.3 | 12.1 | 10.3 | 9.5 |
| IA | AI | AD | DA | AF | AH |
| 15.6 | 13.5 | 15.1 | 13.5 | 14.6 | 18.1 |
| 12.8 | 11.0 | 12.2 | 10.3 | 11.8 | 11.0 |
| AN | VV | AQ | IV | AE | AK |
| 14.6 | 15.1 | 17.1 | 17.3 | 17.5 | 19.9 |
| 11.4 | 12.2 | 12.7 | 13.9 | 13.2 | 12.5 |

Získané výsledky ukazují, že maximální energie roste s velikostí dipeptidu a tudíž neexistuje nic jako univerzální maximální („cutoff“) energie, která by byla pro všechny nebo alespoň významnou část studovaných dipeptidů stejná.

Následně byly stejným způsobem vyhodnoceny i relativní energie konformerů pro podskupinu 6 reprezentativních dipeptidů (zahrnujících všechny typy postranních řetězců), kde výchozí konformery byly získány z reálných proteinových struktur (databáze Top8000 [82] - výběr z Protein Data Bank) a dále optimalizován výše popsanou metodikou (bez jakéhokoliv omezení), stejně jako v [82]. Získané hodnoty pak byly srovnány se stejnými hodnotami z přístupu systematického samplingu (tabulka 4.3), a srovnání je uvedeno v tabulce 4.4., histogramy pak na obrázku 4.4.

Tabulka 4.4.: Srovnání konformačních energetických oken konformerů vybraných dipeptidů získaných systematickým samplingem (s.s.) a z databáze Top8000.

| Dipeptid | $\Delta E_{\max}$(s.s.) [kcal·mol$^{-1}$] | $Q_{19/20}$(s.s.) [kcal·mol$^{-1}$] | $\Delta E_{\max}$(Top8000) [kcal·mol$^{-1}$] | $Q_{19/20}$(Top8000) [kcal·mol$^{-1}$] |
|---|---|---|---|---|
| AA | 10.4 | 8.7 | 10.1 | 7.7 |
| AV | 12.0 | 10.3 | 8.8 | 7.3 |
| AS | 13.4 | 10.3 | 10.9 | 9.0 |
| AD | 15.1 | 12.2 | 12.5 | 8.9 |
| AF | 14.6 | 11.8 | 13.4 | 8.1 |
| AK | 19.9 | 12.5 | 17.6 | 10.7 |

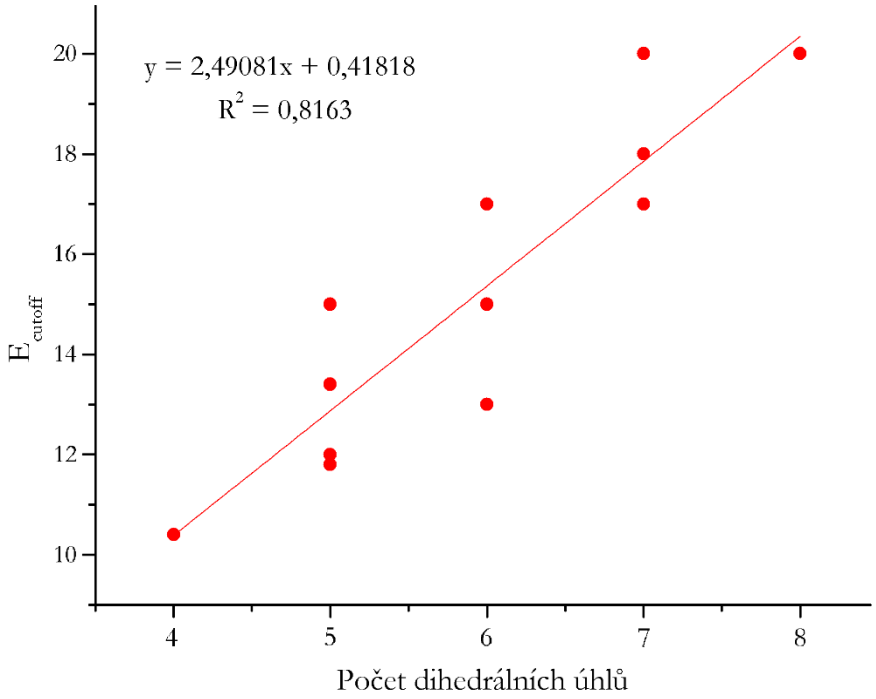



Obrázek 4.4.: Histogramy energií konformerů získaných systematickým samplingem a z Top8000. Pravá plná čára znázorňuje začátek histogramů, levá znázorňuje 95. percentil ($p^{95}$), přerušovaná průměrnou hodnotu.

Tabulka ukazuje, že nekonzistentní šířka energetického okna se vyskytuje i u konformerů z reálných proteinových struktur. Ze srovnání histogramů konformerů získaných systematickým přístupem a konformerů vzorkovaných na základě reálně se vyskytujících dipeptidů je zřejmé, že energetické profily jsou analogické, pouze ve druhém případě je konformerů méně („model prokáceného lesa“). V žádném ze zkoumaných případů nepřesahuje šířka energetického okna 20 kcal·mol$^{-1}$; přičemž po vyloučení 5 % odlehlých hodnot ($Q_{19/20}$) se pohybuje spíše kolem 10 kcal·mol$^{-1}$.

Ze srovnání našich výsledků pro dipeptid AA (DFT-D3 metoda, COSMO-RS solvatace) s obdobnými výsledky z literatury pro tentýž dipeptid (model explicitní vody) [73] je patrné, že histogram je v obou případech podobný, což opět svědčí o relevantnosti získaných výsledků.

Protože se zdá, byť na omezeném vzorku dipeptidů, že šířka energetického okna roste přibližně lineárně s počtem dihedrálních úhlů (vizte obr 4.5.), je možné extrapolovat získaná data pro větší dipeptidy (např. DK, EQ či největší existující dipeptid RR) a určit maximální „*cutoff*“ energii (jakožto hranici, nad kterou by se neměl vyskytnout žádný konformer žádného dipeptidu), jakožto funkci počtu dihedrálních úhlů (alias velikost dipeptidu). Tyto odhadnuté *cutoff* energie mohou být cennou pomůckou pro vyvíjené konformační samplovací algoritmy jako vylučovací kritérium nerealistických konformerů. Provedený fit ukazuje, že odhadnutá maximální energie pro dipeptid o 14 dihedrálních úhlech (RR) je 36,5 kcal·mol$^{-1}$ a tuto energii by za platnosti lineární regrese neměl překročit žádný reálný dipetidický konformer.




Obrázek 4.5.: Korelace šířky konformačního energetického okna dipeptidu s počtem jeho dihedrálních úhlů.

## 4.4. Vliv typu postranního řetězce na energie dipeptidu

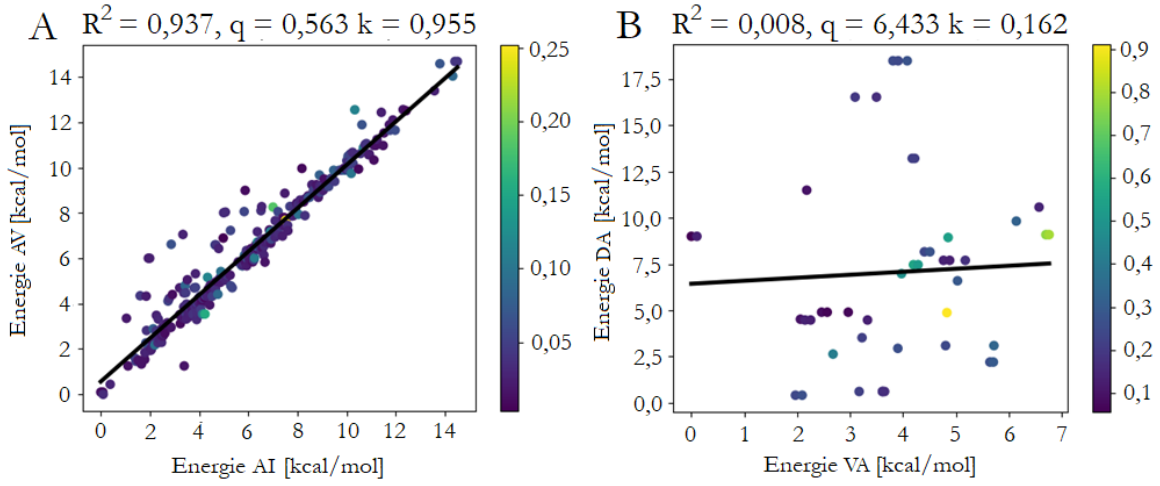
Vzhledem k velmi různé povaze postranních řetězců aminokyselin a tedy jejich předpokládanému jinému chování v rámci dipeptidu vyvstává otázka, nakolik a jakým způsobem záměna postranního řetězce dipeptidu ovlivňuje jeho energii. Za tímto účelem byly porovnány konformační prostory všech dvojic studovaných dipeptidů (pomocí skriptu z dodatku B), a to dle následujího postupu:

U každého konformeru dipeptidu X byly naměřeny čtyři dihedrální úhly hlavního řetězce. Následně byly tyto úhly srovnány se čtyřmi dihedrálními úhly *každého* konformeru dipeptidu Y, a byla spočítána odchylka dihedrálních úhlů ($r$) jako:

$$r = \sqrt{(\psi_1^X - \psi_1^Y)^2 + (\varphi_1^X - \varphi_1^Y)^2 + (\psi_2^X - \psi_2^Y)^2 + (\varphi_2^X - \varphi_2^Y)^2} \qquad (18)$$

Kde $\psi_1$, $\varphi_1$, $\psi_2$, $\varphi_2$ jsou dihedrální úhly první a druhé aminokyseliny (bráno od N-konce) v dipeptidu.

Jako konformer dipeptidu Y *nejpodobnější* ke konformeru dipeptidu X byl pak přirozeně vybrán takový konformer, který měl nejnižší hodnotu odchylky *r*. Tento postup byl zopakován pro všechny konformery dipeptidu X, čímž vznikla množina uspořádaných dvojic $\{X_i,Y_j\}$. Následně byly z této množiny smazány ty dvojice, kde se konformery $X_i$ a $Y_j$ liší o více než 40° v kterémkoliv ze srovnávaných úhlů (jakožto „vynucené" přiřazení dvou konformerů, které si sice jsou z dané množiny nejpodobnější, ale liší se natolik, že nemá praktický smysl je porovnávat). Poté byla sestrojena vážená lineární regrese relativních konformačních energií pro dvojice $\{X_i,Y_j\}$, přičemž jako váha pro tuto regresi byla zvolena převrácená hodnota odchylky *r* a byl spočítán korelační koeficient této regrese (obrázek 4.6.).




Obrázek 4.6.: Ukázky dvou korelací energií konformerů dipeptidů: AI/AV (A), VA/DA (B). Lineární regrese má tvar y = kx + q.

Tento postup byl proveden pro takové dvojice dipeptidů, kde byla zaručena podmínka homomorfního přiřazení (tedy aby byl vždy větší set o více prvcích přiřazován k menšímu setu jako G(M) $\rightarrow$ S(N) kde vždy M $>$ N), protože v opačném případě by docházelo k „vynuceným přiřazením“ z důvodu chybějících prvků a tedy k uměle horšímu korelačnímu koeficientu.

Výsledkem korelační analýzy tedy byly lineární regrese (příklady jsou na obrázku 4.5) a příslušné korelační koeficienty pro každou uvažovanou dvojici. Tyto koeficienty jsou shrnuty v obrázcích 4.7 a 4.8 pro vybrané části (submatice) z „úplné“ matice 17x17.

| | AA | AV | VA | AS | SA | AD | DA | AI | IA |
|---|---|---|---|---|---|---|---|---|---|
| AA | | 0.67 | 0.69 | 0.79 | 0.46 | 0.03 | 0.09 | 0.60 | 0.52 |
| AV | | | 0.50 | 0.41 | 0.36 | 0.10 | 0.01 | 0.94 | 0.33 |
| VA | | | | 0.46 | 0.25 | 0.09 | 0.02 | 0.44 | 0.91 |
| AS | | | | | 0.48 | 0.05 | 0.02 | 0.40 | 0.34 |
| SA | | | | | | 0.00 | 0.05 | 0.31 | 0.24 |
| AD | | | | | | | 0.04 | 0.08 | 0.10 |
| DA | | | | | | | | 0.09 | 0.06 |
| AI | | | | | | | | | 0.27 |
| IA | | | | | | | | | |

Obrázek 4.7.: Matice korelačních koeficientů pro podobnost dipeptidů – submatice odpovídající permutaci aminokyselin.

| | AA | AV | AS | AD | AI | AF | AH | AN | AE | AQ | AK |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AA | | 0.67 | 0.79 | 0.03 | 0.60 | 0,57 | 0.23 | 0.45 | 0.09 | 0.37 | 0.41 |
| AV | | | 0.41 | 0.10 | 0.94 | 0.45 | 0.16 | 0.36 | 0.08 | 0.54 | 0.31 |
| AS | | | | 0.05 | 0.40 | 0.54 | 0.12 | 0.44 | 0.02 | 0.64 | 0.43 |
| AD | | | | | 0.08 | 0.02 | 0.07 | 0.07 | 0.02 | 0.06 | 0.07 |
| AI | | | | | | 0.67 | 0.18 | 0.66 | 0.07 | 0.41 | 0.36 |
| AF | | | | | | | 0.55 | 0.65 | 0.04 | 0.34 | 0.39 |
| AH | | | | | | | | 0.46 | 0.04 | 0.28 | 0.23 |
| AN | | | | | | | | | 0.03 | 0.63 | 0.44 |
| AE | | | | | | | | | | 0.09 | 0.07 |
| AQ | | | | | | | | | | | 0.31 |
| AK | | | | | | | | | | | |

Obrázek 4.8.: Matice korelačních koeficientů pro podobnost dipeptidů – submatice odpovídající záměně jednoho postranního řetězce.

Z matice na obrázku 4.7. je patrné, že korelační koeficient žádné dvojice dipeptidů s permutovanými aminokyselinami (XY vs YX) není blízký jedné, což znamená, že záměna pořadí aminokyselin v dipeptidu se na jeho energii projevuje a tedy že prostá výměna postranních řetězců na dipeptidu má vliv na jeho energii, ale tento vliv nemusí být zásadní (jak ukazují případy AV vs VA, AI vs IA).

Matice na obrázku 4.8. pak ukazuje, že výsledky z práce [77] platí v širším kontextu, tedy že změna substituentů na β-uhlíku je hlavní kritérium pro energetickou podobnost konformerů – při změně postranního řetězce na β-uhlíku se výrazně změní energie při zachování konfigurace hlavního řetězce, a naopak při velmi podobných řetězcích na β-uhlíku je změna malá (například AV a AI, kde je okolí β-uhlíku téměř stejné a změna probíhá až na γ uhlíku) a to i v případě dipeptidů s chráněnými konci a v prostředí vody jakožto rozpouštědla.

Z obou matic je dále vidět, že v případech AD, DA a AH je korelační koeficient blízký nule pro téměř jakoukoliv korelaci, což znamená, že náboj na postranním řetězci dipeptidového fragmentu zásadním způsobem ovlivňuje a mění konformační prostor hlavní řetězce a má výrazný vliv na energii. Z tohoto trendu mírně vybočuje AK s průměrnou hodnotou korelačního koeficientu 0,32, což je patrně způsobeno charakterem postranního řetězce lysinu – nabitá skupina se vyskytuje daleko od zbytku dipeptidu, tudíž se v bezprostředním okolí hlavního řetězce projevuje lysin jako nenabitý fragment.

Aby byla tato hypotéza ověřena, bylo analyzováno, kolik procent z konformerů dipeptidů AD, DA, AH a AK obsahuje vodíkové vazby, a bylo zjištěno, že zatímco u AK vodíkovou vazbu obsahuje 16 % konformerů, v ostatních případech je to více než 30 % (vizte tabulku 4.5), což vysvětluje odlišné chování lysinu popsané výše.

Tabulka 4.5: Procentuální zastoupení konformerů s vodíkovými vazbami.

| Dipeptid | H-vazeb |
|---|---|
| AK | 16 % |
| AH | 31 % |
| AD | 34 % |
| DA | 35 % |

Dále je vidět, že polární nenabité dipeptidy mají nižší hodnoty korelačního koeficientu s dalšími nenabitými dipeptidy, nicméně efekt není tak výrazný, jako v případě náboje. Je tedy zřejmé, že charakter postranního řetězce má vliv na energii konformeru, a potažmo na konformační prostor hlavního řetězce, a tento vliv se uplatňuje u polárních a zejména nabitých specií, a dále že vliv má také záměna dvou postranních řetězců v dipeptidu, ačkoliv tento vliv není tak významný.

## 4.5. Samplingové algoritmy Maestro LLMOD a PlainMD

Protože systematické konformační vzorkování (sampling) vede k získání úplného konformačního prostoru za cenu více než 95 % redundantních výpočtů, byla zkoumána možnost efektivnějšího samplingu, díky kterému by bylo dosaženo stejného nebo velmi podobného výsledku za vynaložení nižší výpočetní námahy.
Pro tento účel byly vybrány dva způsoby samplingu – metoda (horké) molekulové dynamiky (označovaná zde zkráceně PlainMD) a samplingový algoritmus LLMOD, který je součástí softwaru Maestro. Podrobnější popis obou lze nalézt v kapitole 2.2. a způsob provedení v kapitole 3.1.
Výsledné počty neredundantních konformerů (po geometrické optimalizaci DFT protokolem použitým v této práci) shrnuje tabulka 4.6.

Tabulka 4.6.: Počty výsledných unikátních konformerů po geometrické optimalizaci DFT protokolem z výchozích setů generovaných pomocí LLMOD, PlainMD a systematickým samplingem.

| Dipeptid | AA | AV | VA | AS | SA |
|---|---|---|---|---|---|
| PlainMD | 71 | 79 | 72 | 94 | 90 |
| LLMOD | 33 | 32 | 26 | 60 | 69 |
| syst. samp. | 312 | 254 | 256 | 367 | 340 |
| AI | IA | AD | DA | AF | AH |
| 212 | 217 | 105 | 97 | 101 | 124 |
| 116 | 70 | 75 | 55 | 39 | 60 |
| 839 | 828 | 444 | 470 | 360 | 430 |
| AN | VV | AQ | IV | AE | AK |
| 170 | 234 | 812 | 734 | 364 | 1965 |
| 104 | 47 | 289 | 207 | 198 | 407 |
| 821 | 773 | 2614 | 2669 | 1083 | 8891 |

V případě LLMODu je zřejmé, že se v žádném případě nejedná o úplný konformační prostor, ale tento sampling poskytuje v průměru pouze kolem 10 % stabilnějších struktur o menší relativní energii (a tedy i minim na hyperploše potenciální energie), jejichž energetický rozsah je uveden v tabulce 4.7.

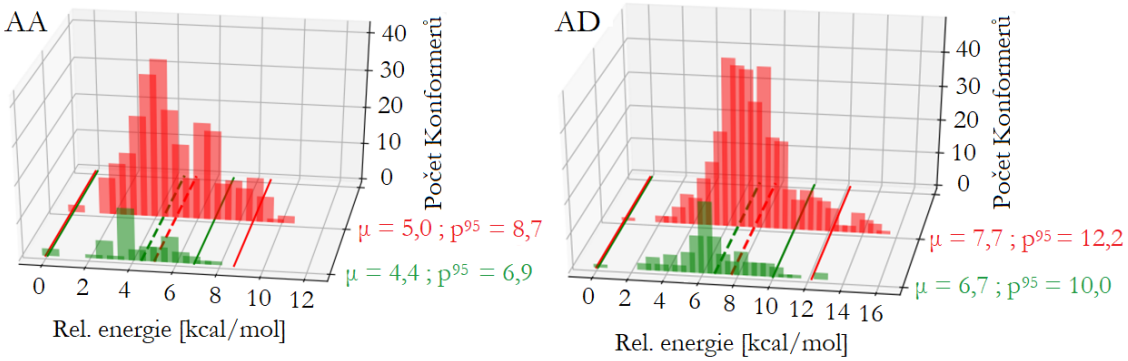
Tabuka 4.7.: Šířky konformačního energetického okna a hodnota na 95. percentilu, (oboje v kcal·mol$^{-1}$) pro sety unikátních konformerů získaných samplingem LLMOD a PlainMD.

| Dipeptid | AA | AV | VA | AS | SA |
|---|---|---|---|---|---|
| LLMOD $\Delta E_{max}$ | 7.5 | 8.7 | 10.6 | 11.0 | 10.3 |
| LLMOD $Q_{19/20}$ | 6.7 | 6.9 | 6.7 | 9.5 | 9.2 |
| MD $\Delta E_{max}$ | 7.7 | 11.1 | 11.9 | 12.7 | 11.4 |
| MD $Q_{19/20}$ | 6.4 | 7.8 | 8.4 | 8.9 | 9.0 |
| IA | AI | AD | DA | AF | AH |
| 10.6 | 10.4 | 12.9 | 10.2 | 12.6 | 11.6 |
| 7.2 | 7.1 | 9.2 | 9.3 | 8.2 | 7.8 |
| 12.1 | 11.7 | 13.4 | 12.7 | 13.8 | 15.2 |
| 8.4 | 8.3 | 10.0 | 10.3 | 9.7 | 9.2 |
| AN | VV | AQ | IV | AE | AK |
| 10.1 | 10.2 | 11.1 | 14.3 | 15.1 | 14.5 |
| 6.3 | 6.8 | 8.0 | 9.8 | 10.9 | 9.7 |
| 12.4 | 10.8 | 12.3 | 15.0 | 17.1 | 15.2 |
| 7.8 | 8.7 | 9.1 | 11.2 | 12.5 | 11.2 |

LLMOD je tedy zcela nevhodný pro dosažení úplného konformačního prostoru, ale lze ho s jistými výhradami použít pro rychlé a výpočetně nenáročné nalezení hlavních regionů konformačního prostoru.

Naproti tomu PlainMD sampling spojený s následnou geometrickou optimalizací poskytuje konformerů více, v průměru 25 % z počtů získaných systematickým samplingem. Energetické charakteristiky takto získaného konformačního

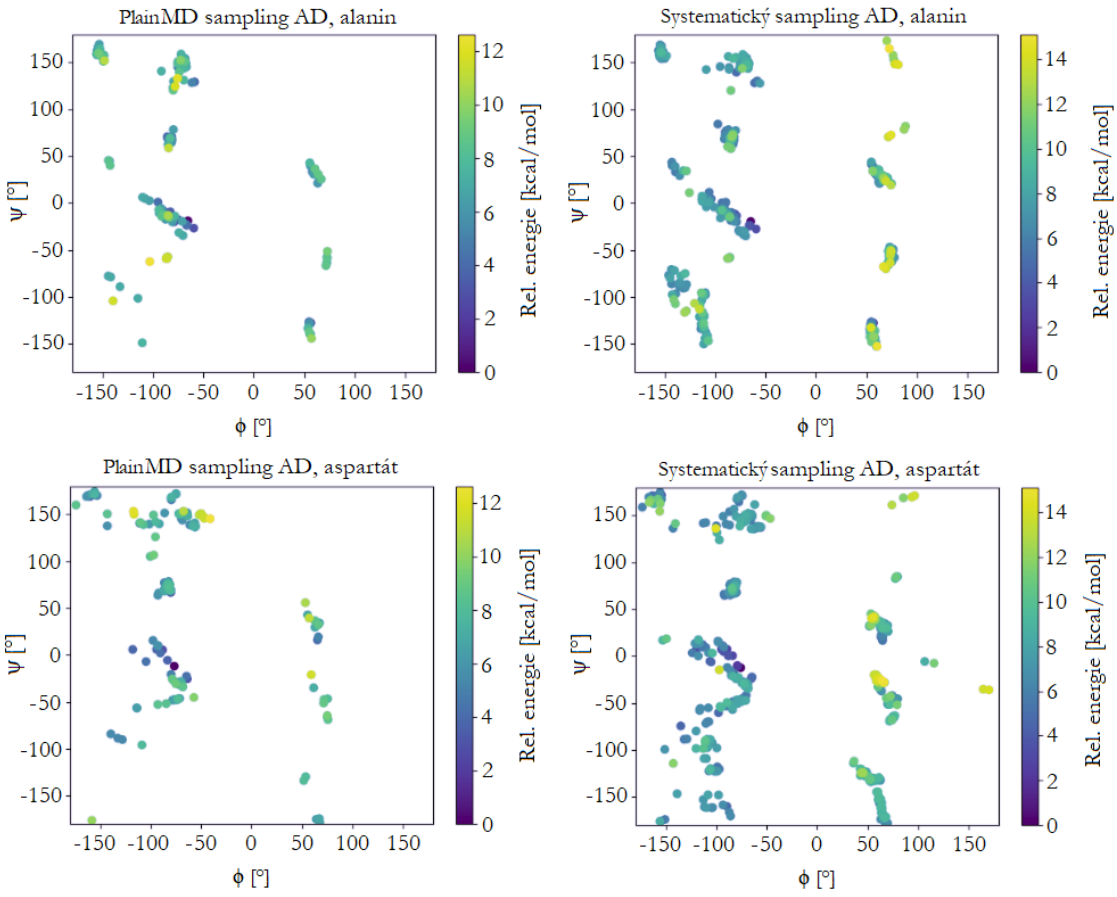
prostoru jsou opět uvedeny v tabulce 4.7. Ze srovnání s energiemi konformerů získaných pomocí LLMOD a PlainMD je patrné, že LLMOD nenachází některé výše energeticky položené konformery, pravděpodobně díky podstatě jeho fungování. Ze srovnání tabulky 4.7. s tabulkou 4.3. je pak zřejmé, že PlainMD sampling je blíže výsledkům náročného systematického samplingu. V rámci vyšetření, jaké jsou charakteristiky takto získaného konformačního prostoru a čím se liší od plného konformačního prostoru, byly srovnány histogramy konformačních energií (zde pro případy AA a AD, vizte obrázek 4.9.). Histogramy jsou podobné, v obou případech bylo nalezeno globální minimum. Průměrná hodnota energie se liší 1 kcal·mol$^{-1}$, což je způsobeno jediným významnějším rozdílem – v histogramu příslušejícímu PlainMD chybí vysokoenergetické konformery, což je patrně způsobeno podstatou metod molekulové dynamiky – při ní může dojít i ke kroku s kladnou (nevýhodnou) změnou energie, nicméně tento krok není pravděpodobný a statisticky k němu tedy dochází více v energeticky nižších oblastech hyperplochy potenciální energie.




Obrázek 4.9.: Srovnání histogramů konformačních energetických oken pro dipeptidy AA a AD pro systematický sampling a při využití PlainMD samplingu. Konformery o relativních energiích cca 1-2 kcal·mol$^{-1}$ v histogramu chybí, protože byly vyhodnoceny jako redundantní s konformerem v globálním minimu.

Ze srovnání Ramachandranových diagramů obou aminokyselin obou dipeptidů z PlainMD a systematického samplingu (obrázek 4.10.) je pak patrné, že všechny hlavní regiony (α-helix, β-list, další regiony) jsou populované, takže PlainMD neobsahuje systematickou chybu (jinou než pominutí vysokoenergetických konformerů zmíněné výše), ale nenalezlo všechny permutace či variace povolených hodnot dihedrálních úhlů. PlainMD sampling lze tedy použít pro nalezení globálního minima, s omezením pak i ke zjištění šířky energetického okna daného oligopeptidu a k zjištění, kde se nacházejí hlavní populované regiony na Ramachandranově diagramu.

Výsledkem tedy je, že ani jeden ze zkoumaných konformačních samplovacích algoritmů neposkytuje ani přibližnou aproximaci úplného konformačního prostoru, oba jsou spolehlivé toliko pro nalezení nejvýraznějších energetických minim.




Obr 4.10.: Srovnání Ramachandranových diagramů dipeptidů AA a AD pro systematický sampling a při využití PlainMD samplingu.

## 4.6. Možnosti indukčních kroků mezi oligopeptidy

Po zmapování úplných konformačních prostorů všech aminokyselin a vybraných dipeptidů (vizte kapitoly 3.1.1. a 3.1.2.) byl proveden pokus o odvození indukčních kroků mezi konformačními prostory aminokyselin a dipeptidů, s cílem předpovědět konformační prostor dipeptidu na základě znalosti konformačních prostorů výchozích aminokyselin, případně jednoduššího (příbuzného) dipeptidu. Pro tento účel byly vybrány aminokyseliny alanin a valin a dipeptidy AA a AV a otázka zněla – je možné pomocí relativně levně získaných setů unikátních konformerů A, V a AA (81, 729 a 6561 výchozích

konformerů) předpovědět konformační prostor AV, jinak získaný relativně draze (19683 výchozích konformerů)? Důvod pro tuto volbu je, že se jedná o (v rámci oligopeptidů) velmi podobné struktury, tudíž jakékoliv indukční kroky lze předpokládat spíše mezi těmito strukturami, než v případě velmi nepodobných oligopeptidů.

Nejprve byla prozkoumána možnost, jestli je možné konformační prostor dipeptidů „složit" z konformačních prostorů odpovídajících aminokyselin – tedy zda konformační prostory AA a AV jsou „direktním součtem" konformačních prostorů dvou A, respektive A a V. Bylo zjištěno, že konformační prostor dipeptidu *nelze* tímto způsobem v žádném případě získat, protože pro některé kombinace výsledných úhlů hlavního řetězce dipeptidů neexistuje ekvivalent v případě aminokyselin – konformační prostory A a V jsou podmnožinou konformačního prostoru AV a konformační prostor A je podmnožinou konformačního prostoru AA, ale zbytek této množiny nelze tímto způsobem najít.

Tento fakt je nahlédnutelný už ze samotných počtů konformerů: V případě alaninu je ve výsledném setu 7 konformerů, tedy počet jejich variací s opakováním (a zároveň počet takto předpokládaných konfomerů AA) je 49, ale skutečný počet konformerů je 312, tedy více než šestkrát tolik. Ve zbylých dvou případech je výsledek podobný. Tento výsledek je očekávatelný, protože v případě aminokyseliny není žádná interakce s dalšími částmi uvažovaného dipeptidu, jehož by byla součástí, a tedy tyto interakce nelze nijak odvodit.

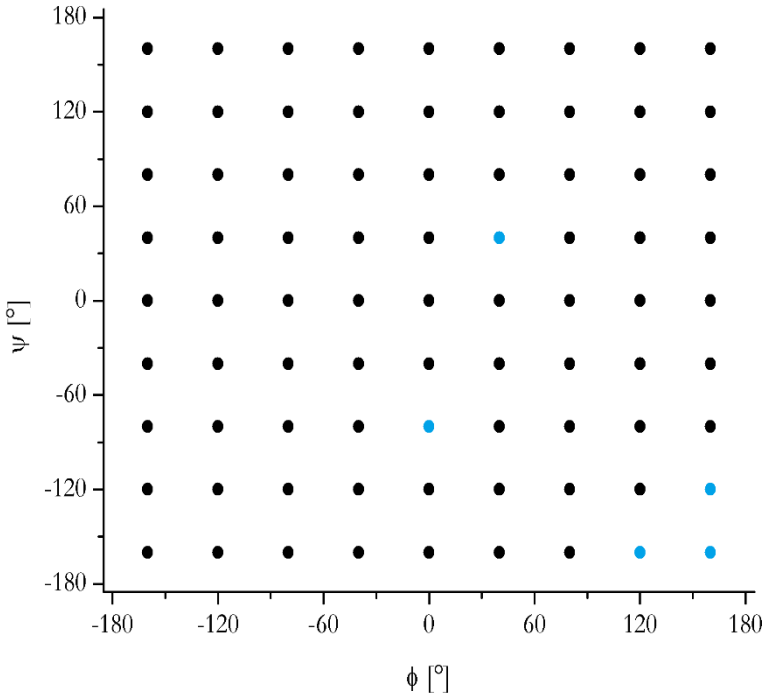
Následně bylo zjištěno, že obdobně konformační prostor hlavního řetězce AV není podmnožinou konformačního prostoru AA a nelze jej tedy získat jakoukoliv metodou systematického výběru správných kombinací dihedrálních úhlů konformerů AA.

Proto byly dále zkoumány možnosti využití výchozích dihedrálních úhlů - tedy zda lze z menších reziduí (např. A a V) vybrat unikátní a výsledné (finální) konformery, nakombinovat jejich původní dihedrální úhly, čímž vytvořit

dipeptid, zoptimalizovat takto vytvořené dipeptidové struktury, a získat tak konformační prostor větších reziduí (např. AV). Nejprve byly pro všechny finální konformery A, V, AA, a AV naměřeny dihedrální úhly hlavního řetězce jejich výchozích struktur (tedy struktur ve stavu předtím, než byly podrobeny výpočetnímu protokolu z kapitol 3.1.1. a 3.1.2.). Následně byly porovnány sety těchto výchozích dihedrálních úhlů mezi aminokyselinou a dipeptidem, který ji obsahuje (A vs AA, V vs AV, A vs AV – v těchto případech se u AV srovnávalo alaninové reziduum s alaninem a valinové s valinem), a také mezi oběma dipeptidy (AA vs AV).

V případech A vs AA, V vs AV a A vs AV bylo vždy zjištěno, že spojením výchozích dihedrálních úhlů jednotlivých aminokyselin nedojde k nalezení všech výchozích dihedrálních úhlů unikátních konformerů dipeptidu, tedy že nalézt tímto způsobem pouze konformery, které po optimalizaci povedou k výslednému setu unikátních konformerů, není možné. Jinými slovy, pokud bychom použili výchozí hodnoty dihedrálních úhlů k tomu, abychom utvořili všechny jejich permutace jako dihedrální úhly dipeptidu, tak výsledné (zoptimalizované) konformery nedávají dostatečný konformační prostor. Závěrem tedy je, že indukční krok, který by z konformačních preferencí aminokyselin odvodil konformační preference dipeptidů nelze jednoduše provést.

V případě dipeptidů AA a AV byly opět srovnávány množiny výchozích dihedrálních úhlů hlavního řetězce, a to pro obě N-terminální alaninová rezidua. Bylo zjištěno, že ačkoli výsledné konformery AV vznikly optimalizací z množiny výchozích konformerů, která obsahuje všech 81 výchozích kombinací dihedrálních úhlů pro alaninové reziduum, pro AA v takové množině 5 kombinací chybí, vizte obrázek 4.11.




Obrázek 4.11.: Chybějící kombinace výchozích dihedrálních úhlů hlavního řetězce pro N-koncový alanin v AA (modře).

Tento výsledek znamená, že konformační prostor AV nevznikl z množiny, která je podmnožinou výchozího konformačního prostoru AA, ale vice versa. Na základě tohoto zjištění tedy není možné provést ani indukční krok mezi dvěma dipeptidy se záměnou postranního řetězce, a to pravděpodobně v žádném případě, protože si lze jen těžko představit, že by tento postup fungoval v případě komplexnější změny, když není možný ani pro postranní řetězce alaninu a valinu. Tento závěr je tedy další ukázka toho, že Floryho hypotéza izolovaného páru obecně neplatí, což je ve shodně s dosud publikovanou literaturou [23], [24].

Závěrem tedy jest, konformační prostor oligopetidových reziduí v proteinu ovlivňují jak dlouhodosahové, tak krátkodosahové interakce, a indukční kroky jak typu $n \rightarrow n+1$, tak typu $n \rightarrow n$` nelze jednoduše provést tak, aby byl konformační prostor spolehlivě úplný.

# Závěr

V rámci této práce jsem se pokusil vyšetřit konformační chování a preference krátkých peptidových fragmentů teoretickým studiem, pomocí výpočetních metod založených na teorii hustotního funkcionálu s disperzní korekcí. K problému jsem přistoupil jako k pokusu o *ab initio* přístup z postupné výstavby z menších fragmentů k fragmentům větším, tedy snažil jsem se o co nejúplnější mapování konformačního prostoru oligopeptidů a nalezení pravidel či trendů pro tyto prostory.

Celkem jsem provedl přes 2 miliony geometrických optimalizací a energetických výpočtů na výkonných výpočetních klastrech o celkové náročnosti přes 1 milionu jádrohodin. Celá práce vyžadovala velmi důkladné a pokročilé skriptování z hlediska přípravy výpočtů i evaluace, aby bylo možné všechny výpočty i celé následné vyhodnocení vůbec dokončit v myslitelném čase.

Výsledky ukazují, že nejvýznamnějším činitelem na ovlivnění jak konformačního prostoru, tak počtu konformerů, který tento konformační prostor definují, je náboj na postranním řetězci oligopeptidu, respektive elektrostatické interakce s ním související. Dalšími důležitými faktory pro konformační prostor jsou elektronegativní atomy (z hlediska tvorby vodíkových vazeb) a hybridizace na uhlíkových atomech postranních řetězců. Samotné konformační sety podobných oligopeptidů spolu souvisí, ale neexistuje jednoduchý způsob, jak z jednoho odvodit podobu jiného. Následkem toho také není možné udělat jednoduchý indukční krok mezi oligopeptidy stejného ani různého řádu.

Získané sady dat jsou přesto cenné, protože nejsou pouhou aproximací, nýbrž téměř úplným konformačním prostorem, a mohou být dále podrobeny podrobnější analýze na pokročilejší úrovni, například využitím strojového nebo hlubokého učení.

Dále tato práce potvrzuje velmi omezenou platnost Floryho hypotézy izolovaného páru, tedy že sousední aminokyseliny v rámci proteinu ovlivňují jeho konformační prostor a jeho energii.

Závěrem lze říci, že bylo dosaženo cílů této diplomové práce, a konstatovat, že lepšího predikování konformačního prostoru peptidových reziduí může být dosaženo pozorováním chování při změnách na nižší úrovni, například tak, že se budou porovnávat změny konformačních prostorů po nahrazení vodíku methylem, methylu karboxylem, a podobně.

# Dodatek A

## Algoritmus pro odstranění redundantních konformerů

Principem tohoto algoritmu je rozdělení N-rozměrného konformačního prostoru (kde N je počet dihedrálních úhlů a každá z N os nabývá hodnot -180° až 180°) na identické buňky o rozměru $(20°)^N$, tedy N-rozměrné buňky o hraně 20°, které budou reprezentovány právě jedním konformerem, a to takovým, který má z konformerů v této buňce nejnižší energii.

Vstupními daty pro tento algoritmus jsou tedy seznam dihedrálních úhlů jednotlivých konformerů seřazených podle stoupající relativní energie (jako první argument) a samostatný soubor těchto energií (jako druhý argument), vizte níže:

```
AV1000 -78.5610854762 -20.3754982038 -154.953153502 171.502155885
166.416727278
AV100 -153.030100734 164.713562821 -70.2589958929 -29.1738474916 -
51.7489786304
AV1001 63.083828065 21.752622515 -84.4465362837 152.476577972 52.5402936967
AV1002 -61.4284491495 128.019364916 64.8832552534 21.1914919491 -
143.69259695
AV1003 -79.0428780363 -20.1278130225 -155.489388469 172.443056138 -
158.824558229
                                    ……
```

```
0
0.1985
0.23664
0.25922
0.26454
   …
```

Kde úhly jsou v pořadí $\psi_1$, $\varphi_1$, $\psi_2$, $\varphi_2$, $\chi_1$, $\chi_2$,... $\chi_n$, tj. první čtyři hodnoty úhlů vždy se týkají hlavního řetězce.

Samotný skript algoritmu je uveden níže:

```
#!/usr/bin/python
# Import používaných knihoven
import numpy as np
from sys       import argv
import string as s
# Definice polí pro data
pole = [[]]
dih = []
```

```
names = []
# Naplnění pole dihedrálními úhly
with open(argv[1]) as file:
        for line in file:
                l = line.strip().split()
                dih.append(l[1:])
                names.append(l[0])
# Definice pole r
pole = dih
r = []
# Zaokrouhlení dohedrálních úhlů na násobky dvaceti
for line in pole:
        lr = []
        for angle in line:
                y = (round((float(angle)/20), 0))*20
                # Nahrazení 180 místo 180,
                if y == -180 :
                        y = 180
                # Naplnění pole r
                lr.append(y)
        r.append(lr)
# Import souboru energií (seřazených od nejnižší) a jeho načtení do pole
energie
energies = []
with open(argv[2]) as file:
    for line in file:
        l = line.strip().split()
        energies.append(float(l[0]))
# Definice pole x, které je následně naplněno nulami o délce pole r
x = []
for j in range(0, len(r)):
        x.append(0)
# Pro každý řádek v poli r (každou kombinaci dihedrálních úhlů) hledání ve
všech řádcích pod ním
for j in range(0, len(r)):
        if x[j] != 1:
                for g in range(j+1, len(r)):
                        # Pokud jsou stejné dihedrální úhly, je nula v řádku
                   g v poli x nahrazena jednotkou aje tak označen redundantní
                   konformer.
                        if r[j] == r[g] :
                                x[g] = 1

# Konečně jsou všechny řádky s danou kombinací úhlů a nejnižší energií
(označené jedničkou) vytištěny
for j in range(0, len(r)):
        if x[j] != 1:
                print names[j],  " ".join(dih[j]), energies[j]
```

# Dodatek B

## Algoritmus pro nalezení nejpodobnějšího dipeptidu

Tento algoritmus hledá pro jednu kombinaci dihedrálních úhlů hlavního řetězce v daném seznamu kombinaci, která je té původní nejpodobnější. Podobnost přitom vyhodnocuje na základě odchylky, jak bylo nastíněno v kapitole 4.3. Nepreferuje žádný úhel, všechny mají stejnou váhu. Zároveň ze separátních souborů vezme i hodnoty energie a vytiskne je spolu s oběma kombinacemi dihedrálních úhlů a hodnotou odchylky, jako součást výstupu.

Vstupními daty pro tento algoritmus jsou tedy seznamy dihedrálních úhlů jednotlivých konformerů seřazených podle stoupající relativní energie a rozdělené na dva řádky, a samostatné soubory energií (jako druhý argument), vizte níže:

| | |
|---|---|
| AI_8008_01 ALA -74.40566 -11.204214<br>AI_8008_01 ILE -102.5449754 8.3965659 61.0415847 -65.3950527 167.1268688<br>AI_2027_01 ALA -74.236405 -11.5465239<br>AI_2027_01 ILE -104.4966159 10.171733 59.8493878 -66.484209787 165.6110758<br>AI_7187_01 ALA -74.939128676 -9.99390050925<br>AI_7187_01 ILE -102.0785712 7.5941869 60.4835944 -65.88095302 166.3693498<br>……. | 0<br>0.45256<br>0.35434<br>0.34532<br>0.63894<br>0.53488<br>… |
| AD_1277_10 ALA -161.152846346 167.802284918<br>AD_1277_10 ASP -151.57338052 175.6971103 -163.812988334 158.568007149<br>AD_3742_11 ALA -125.289380657 13.0016552495<br>AD_3742_11 ASP -156.592535552 175.488006091 -165.702240877 161.16361942<br>AD_8126_12 ALA -117.087919133 9.40778341419<br>AD_8126_12 ASP -157.427555544 175.46254025 -165.521916136 160.507436775<br>……. | 0<br>0.09726<br>0.14526<br>0.31818<br>0.44156<br>0.41845<br>… |

Samotný skript je opět uveden níže:

```
#!/usr/bin/python
# Import používaných knihoven
from sys       import argv
from math import pow, sqrt, sin, cos, radians
```

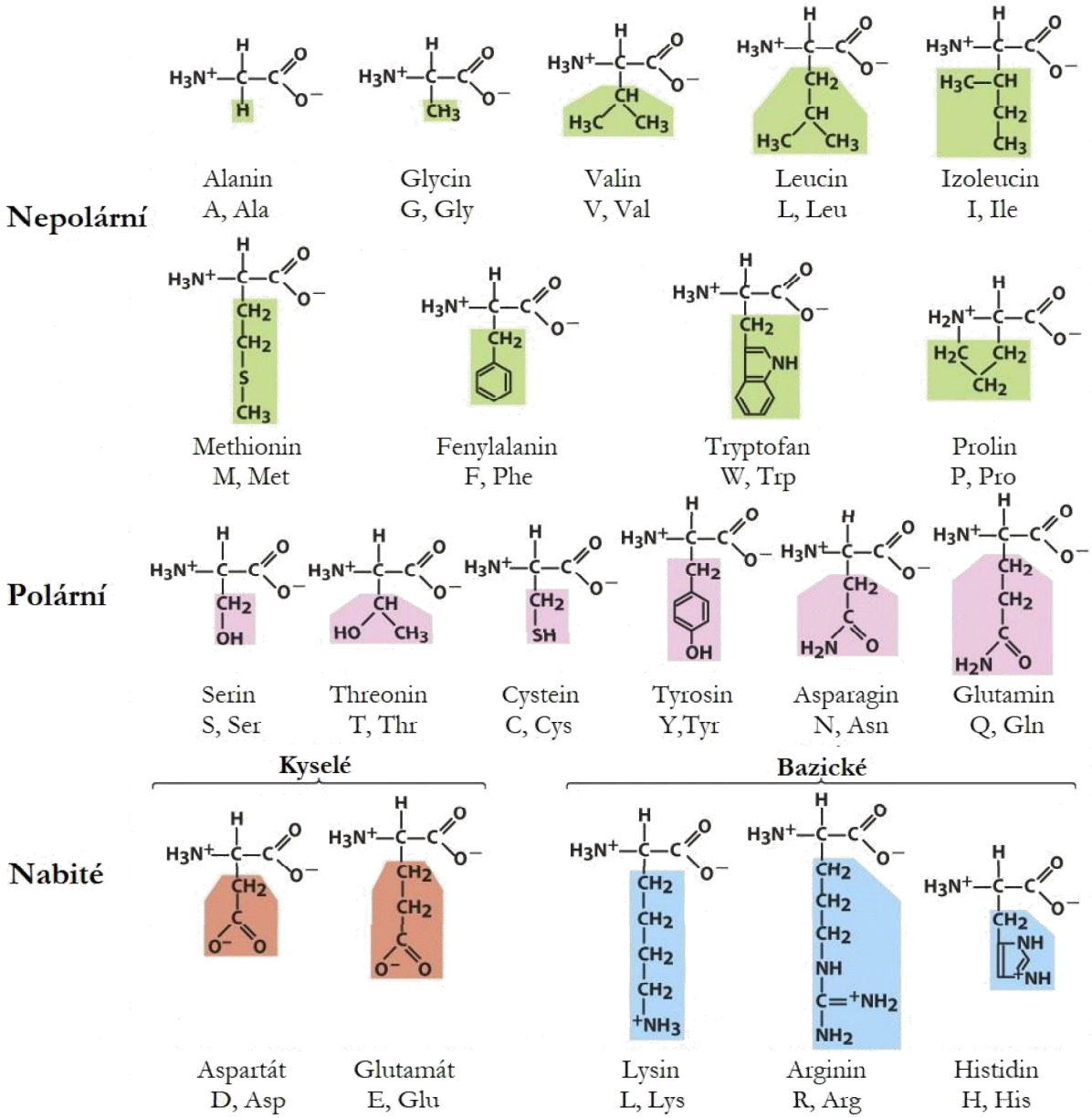
```
# Načtení obou souborů dihedrálních úhlů i energií do příslušných polí
pep2_database1 = []
with open(argv[1]) as file:
    for line in file:
        l = line.strip().split()
        pep2_database1.append(l)
pep2_energies1 = []
with open(argv[2]) as file:
    for line in file:
        l = line.strip().split()
        pep2_energies1.append(l)
pep2_database2 = []
with open(argv[3]) as file:
    for line in file:
        l = line.strip().split()
        pep2_database2.append(l)
pep2_energies2 = []
with open(argv[4]) as file:
    for line in file:
        l = line.strip().split()
        pep2_energies2.append(l)
# Pro první set konformerů jsou definoványproměnné best_dist a best2 for i in
range(0, len(pep2_database1), 2):
    best2 = 0
    best_dist = 999999999

    # Pro druhý set konfromerů je nejprve zrušena periodicita úhlů pomocí fcí
sin a cos, a následně vypočítána odchylka (new dist) pomocí funkce sqrt()
    for j in range(0,len(pep2_database2), 2):
            sum1 = 0
            for k in range (2,4) : sum1 +=
             pow(sin(radians(float(pep2_database2[j][k])))-
             sin(radians(float(pep2_database1[i][k]))),2) +
             pow(cos(radians(float(pep2_database2[j][k])))-
             cos(radians(float(pep2_database1[i][k]))),2)
            for k in range (2,4) : sum1 +=
             pow(sin(radians(float(pep2_database2[j+1][k])))-
             sin(radians(float(pep2_database1[i+1][k]))),2) +
             pow(cos(radians(float(pep2_database2[j+1][k])))-
             cos(radians(float(pep2_database1[i+1][k]))),2)
            new_dist = sqrt(sum1)
            if new_dist<best_dist :
                # Pokud je nalezen podobnější dipeptid, uloží se jeno
označení
                Best2 = j
                # a nová (menší) hodnota odchylky
                best_dist=new_dist
    # Nakonec se všechny kombinace podobných dipeptidů vytisknou:
    print
pepě_database1[i][0],pep2_database1[i][2],pep2_database1[i][3],pep2_database1
[i+1]
```

```
[2],pep2_database1[i+1][3],pep2_energies1[i/2][0],pep2_database2[best2][0],pe
p2_database2[best2][2],pep2_database2[best2][3],pep2_database2[best2+1][2],pe
p2_database2[best2+1][3],pep2_energies2[best2/2][0],float(pep2_energies2[best
2/2][0]),float(pep2_energies1[i/2][0]),best_dist
```

# Dodatek C

Přehled postranních řetězců základních proteinogenních aminokyselin

Převzato a upraveno z [89]



# Seznam zkratek a symbolů

| | |
|---|---|
| $\alpha$ | Polarizabilita |
| $E$ | Intenzita elektrického pole |
| $\varepsilon$ | Dielektrická permitivita |
| $P_{\text{ind}}$ | Indukovaný dipólový moment |
| $p$ | Permanentní dipólový moment |
| $q$ | Náboj částice |
| $\psi$, $\varphi$, $\chi$ | Dihedrální úhly proteinové struktury |
| BOA | Bornova-Oppenheimerova aproximace |
| CC | Spřažené klastry (Coupled Clusters) |
| CI | Konfigurační interakce (Configurational Interaction) |
| COSMO | Conductor-like Screening Model |
| COSMO-RS | Conductor-like Screening Model - Realistic Solvation |
| CryoEM | Kryogenická elektronová mikroskopie |
| DFT | Density Functional Theory (Teorie hustotního funcionálu) |
| FF | Force Field (Silové pole) |
| GB/SA | Generalized Born/Solvent accesible surface area |
| GGA | Generalized Gradient Aprox. (Aproximace zobecněného gradientu) |
| HF | Hartree-Fock |
| LDA | Local Density Approximation (Aproximace lokální hustoty) |
| LLMOD | Large-Scale Low-Mode |
| MC | Monte Carlo |
| MD | Molekulová dynamika |
| MM | Molekulová mechanika |
| NMR | Nukleární magnetická resonance |
| QM/MM | Kvantová mechanika/Molekulová mechanika |

# Použitá literatura


[1] M. Kodíček, O. Valentová, a R. Hynek, *Biochemie: chemický pohled na biologický svět*. VŠCHT Praha, 2015.

[2] "MDAnalysis User Guide." [Online]. Available: https://www.mdanalysis.org/UserGuide/_images/dihedrals.png.

[3] H. Frauenfelder, *The Physics of Proteins; An Introduction to Biological Physics and Molecular Biophysics*. Springer, 2010.

[4] G. N. Ramakrishnan, C. . Ramachandran, "Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units.," *Biophys. J.*, vol. 5, pp. 909–33, 1965.

[5] "Chegg study online textbook." [Online]. Available: https://www.chegg.com/homework-help/questions-and-answers/q5-structure-alpha-helix-r-denotes-amino-acid-sidechains--drawn-direction-n-c-top-bottom-b-q10068487.

[6] "Wikiwand," *Wikipedia reader*. [Online]. Available: https://www.wikiwand.com/en/Beta_sheet.

[7] M. Thommen, W. Holtkamp, a M. V. Rodnina, "Co-translational protein folding: progress and methods," *Current Opinion in Structural Biology*, vol. 42. Elsevier Ltd, pp. 83–89, 01-Feb-2017.

[8] G. G. Glenner a C. W. Wong, "Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein," *Biochem. Biophys. Res. Commun.*, vol. 120, pp. 885–890, 1984.

[9] V. Koppaka a P. Axelsen, "Accelerated Accumulation of Amyloid β Proteins on Oxidatively Damaged Lipid Membranes," *Biochemistry*, vol.

39, pp. 10011–10016, 2000.

[10] W. J. G. Hol, P. T. van Dujinen, a H. J. C. Berendsen, "Alpha-helix dipole and properties of proteins," *Nature*, vol. 273, no. 5662, pp. 443–6, 1978.

[11] E. N. Baker a R. E. Hubbard, "Hydrogen bonding in globular proteins," *Progress in Biophysics and Molecular Biology*, vol. 44, no. 2. Pergamon, pp. 97–179, 01-Jan-1984.

[12] M. L. Paddock, G. Feher, a M. Y. Okamura, "Proton transfer pathways and mechanism in bacterial reaction centers," *FEBS Lett.*, vol. 555, no. 1, pp. 45–50, Nov. 2003.

[13] I. K. McDonald a J. M. Thornton, "Satisfying hydrogen bonding potential in proteins," *J. Mol. Biol.*, vol. 238, no. 5, pp. 777–793, May 1994.

[14] C. Nick Pace, J. Martin Scholtz, a G. R. Grimsley, "Forces stabilizing proteins," *FEBS Letters*, vol. 588, no. 14. Elsevier, pp. 2177–2184, 27-Jun-2014.

[15] V. Adrian Parsegian, *Van der Waals forces: A handbook for biologists, chemists, engineers, and physicists*. Cambridge University Press, 2005.

[16] D. Langbein, "Theory of Van der Waals attraction," Springer, Berlin, Heidelberg, 1974, pp. 1–139.

[17] C. Levinhtal, "How to Fold Graciously," *Mossbauer Spectrosc. Biol. Syst. Proc. a Meet. held Allert. House, Monticello, Illinois*, pp. 22–24, 1969.

[18] C. Levinthal, "Are there pathways for protein folding?," *J. Chim. Phys. Physico-chimie Biol.*, vol. 65, no. 1, p. 44+, 1968.

[19] M. Sadqi, L. J. Lapidus, a V. Munoz, "How Fast Is Protein Hydrophobic Collapse?," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, p. 12117−12122, 2003.

[20] S. W. Englander a L. Mayne, "The Nature of Protein Folding Pathways," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, p. 15873–15880, 2014.

[21] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, a P. G. Wolynes, "Funnels, pathways, and the energy landscape of protein folding: A synthesis," *Proteins Struct. Funct. Genet.*, vol. 21, no. 3, pp. 167–195, Mar. 1995.

[22] "Freie Universitat Berlin, Computational Sciences." [Online]. Available: http://www.compsci.fu-berlin.de/en/index.html.

[23] R. V. Pappu, R. Srinivasan, a G. D. Rose, "The Flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 23, pp. 12565–12570, Nov. 2000.

[24] Y. Z. Ohkubo a C. L. Brooks, "Exploring Flory's isolated-pair hypothesis: Statistical mechanics of helix-coil transitions in polyalanine and the C-peptide from RNase A," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. SUPPL. 2, pp. 13916–13921, Nov. 2003.

[25] E. Callaway, "Revolutionary cryo-EM is taking over structural biology," *Nature*, vol. 578, no. 7794. NLM (Medline), p. 201, 01-Feb-2020.

[26] J. Forman-Kay a H. Sun Chan, "Experimental methods for the study of protein folding," 2018. [Online]. Available: http://arrhenius.med.utoronto.ca/~chan/JBB2026H_Lec5_Oct12_2018.pdf.

[27] V. Muñoz a M. Cerminara, "When fast is better: Protein folding fundamentals and mechanisms from ultrafast approaches," *Biochemical Journal*, vol. 473, no. 17. Portland Press Ltd, pp. 2545–2559, 01-Sep-2016.

[28] M. Dorn, M. e Silva, L. S. Buriol, a L. C. Lamb, "Three-dimensional protein structure prediction: Methods and computational strategies,"

*Comput. Biol. Chem.*, vol. 53, no. B, pp. 251–276, Dec. 2014.

[29] K. Lindorff-Larsen, S. Piana, R. O. Dror, a D. E. Shaw, "How Fast-Folding Proteins Fold," *Science (80-. ).*, vol. 334, no. 6055, pp. 517–520, Oct. 2011.

[30] A. Tramontano, *Protein Structure Prediction*. John Wiley and Sons, Inc., 2006.

[31] A. Roy, A. Kucukural, a Y. Zhang, "I-tasser: a unified platform for automated protein structure and function prediction.," *Nat. Protoc*, vol. 5, pp. 725–738, 2010.

[32] D. Kim, F. DiMaio, R. Yu-Ruei Wang, Y. Song, a D. Baker, "One contact for every twelve residues allows robust and accurate topology-level protein structure modeling.," *Proteins 82*, pp. 208–218, 2014.

[33] D. T. Jones *et al.*, "Prediction of novel and analogous folds using fragment assembly and fold recognition," *Proteins 61*, vol. 7, p. 143, 2005.

[34] C. Floudas, H. Fung, S. McAllister, M. Moennigmann, a R. Rajgaria, "Advances in protein structure prediction and de novo protein design: a review," *Chem. Eng. Sci.*, vol. 61, no. 3, p. 966, 2006.

[35] R. Russell a G. Barton, "Structural features can be unconserved in proteins with similar folds. an analysis of side-chain to side-chain contacts secondary structure and accessibility.," *J. Mol. Biol.*, vol. 224, no. 3, p. 332, 1994.

[36] K. Arnold, L. Bordoli, J. Kopp, a T. Schwede, "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling.," *Bioinformatics*, vol. 22, no. 2, pp. 195–201, Jan. 2006.

[37] E. Bramucci, A. Paiardini, F. Bossa, a S. Pascarella, "PyMod: Sequence

similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL," *BMC Bioinformatics*, vol. 13, no. SUPPL.4, p. S2, Mar. 2012.

[38] N. D. Yılmazer, "Computational Screening of Energy- and Bio-materials," Universität Ulm, 2015.

[39] B. R. Brooks *et al.*, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, Jul. 2009.

[40] D. A. Case *et al.*, "The Amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, no. 16. pp. 1668–1688, Dec-2005.

[41] Leach A. R, *Molecular modelling principles and applications*. Pearson, 2001.

[42] M. Culka a L. Rulíšek, "Interplay between Conformational Strain and Intramolecular Interaction in Protein Structures: Which of Them Is Evolutionarily Conserved?," *J. Phys. Chem. B*, 2020.

[43] P. Slavíček, E. Muchová, D. Hollas, V. Svoboda, a O. Svoboda, *Kvantová chemie: První čtení*. 2019.

[44] A. Merkel, R. Zahradník, a Z. Havlas, "Evaluation of the Rate Constant for the SN2 Reaction CH3F + H- → CH4 + F-in the Gas Phase," *J. Am. Chem. Soc.*, vol. 110, no. 25, pp. 8355–8359, 1988.

[45] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models, 2nd Edition*. Wiley, 2004.

[46] A. D. Becke, "Correlation-energy of an inhomogeneous electron-gas - A coordinate-space model," *J. Chem. Phys.*, vol. 88, no. 2, pp. 1053–1062, Jan. 1988.

[47] C. Lee, W. Yang, a R. G. Parr, "Development of the Colle-Salvetti

correlation-energy formula into a functional of the electron density.," *J. Phys. Chem. B*, vol. 37, pp. 785–789, 1988.

[48] J. P. Perdew, "Density-functional approximation for the correlation energy of the inhomogeneous electron gas," *Phys. Rev. B*, vol. 33, no. 12, pp. 8822–8824, 1986.

[49] J. P. Perdew, K. Burke, a M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.*, vol. 77, no. 18, pp. 3865–3868, 1996.

[50] A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior," *Phys. Rev. A*, vol. 38, no. 6, pp. 3098–3100, 1988.

[51] J. M. Pérez-Jordá a A. D. Becke, "A density-functional study of van der Waals forces: rare gas diatomics," *Chem. Phys. Lett.*, vol. 233, no. 1–2, pp. 134–137, Feb. 1995.

[52] S. Grimme, S. Ehrlich, a L. Goerigk, "Effect of the Damping Function in Dispersion Corrected Density Functional Theory," *J. Comput. Chem.*, vol. 32, no. 7, pp. 1456–1465, May 2011.

[53] A. D. Becke a E. R. Johnson, "A density-functional model of the dispersion interaction," *J. Chem. Phys.*, vol. 123, no. 15, Oct. 2005.

[54] S. Grimme, J. Antony, S. Ehrlich, a H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu," *J. Chem. Phys.*, vol. 132, no. 15, Apr. 2010.

[55] L. Goerigk, "A Comprehensive Overview of the DFT-D3 London-Dispersion Correction," in *Non-covalent interactions in quantum chemistry and physics: theory and applications*, DeLaRoza, AO and DiLabio, GA, Ed. Sara Burgenhartstraad 25, PO BOX 211, 1000 AE

Amsterdam, Netherlands: Elsevie Science BV, 2017, pp. 195–219.

[56] J. Rezac, D. Bim, O. Gutten, a L. Rulisek, "Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set," *J. Chem. Theory Comput.*, vol. 14, no. 3, pp. 1254–1266, Mar. 2018.

[57] M. J. S. Dewar a J. J. P. Stewart, "AM1: A New General Purpose Quantum Mechanical Model," *J. Am. Chem. Soc.*, vol. 107, pp. 3902–3909, 1985.

[58] J. J. P. Stewart, "Optimization of Parameters for Semi-empirical Methods I. Method," *J. Comput. Chem.*, vol. 10, pp. 221–264, 1989.

[59] J. Řezáč a P. Hobza, "Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods," *J. Chem. Theory Comput.*, vol. 8, no. 1, pp. 141–151, Jan. 2012.

[60] C. Bannwarth, S. Ehlert, a S. Grimme, "GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions," *J. Chem. Theory Comput.*, vol. 15, no. 3, pp. 1652–1671, Mar. 2019.

[61] S. Grimme, C. Bannwarth, a P. Shushkov, "A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z=1-86)," *J. Chem. Theory Comput.*, vol. 13, no. 5, pp. 1989–2009, May 2017.

[62] D. Voet a J. G. Voet, *Biochemistry*, 3rd ed. Hoboken: J. Wiley & Sons, 2004.

[63] M. Khalili, A. Liwo, A. Jagielska, a H. A. Scheraga, "Molecular dynamics

with the united-residue model of polypeptide chains. II. Langevin and Berendsen-Bath dynamics and tests on model alpha-helical systems," *J. Phys. Chem. B*, vol. 109, no. 28, pp. 13798–13810, Jul. 2005.

[64] D. J. Wales a H. A. Scheraga, "Review: Chemistry - Global optimization of clusters, crystals, and biomolecules," *Science (80-. ).*, vol. 285, no. 5432, pp. 1368–1372, Aug. 1999.

[65] J. M. Blaney a M. S. Dixon, "Distance geometry in Molecular Modelling," *Rewvews Comput. Chem.*, vol. 5, 1994.

[66] J. Kolafa, *Molekulové modelování a simulace*. VŠCHT Praha, 2015.

[67] N. Metropolis, A. Rosebluth, M. Rosenbluth, A. Teller, a E. Teller, "Equation of State Calculations by Fast Computing Machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[68] I. Kolossvary a G. M. Keseru, "Hessian-free low-mode conformational search for large-scale protein loop optimization: Application to c-jun N-terminal kinase JNK3," *J. Comput. Chem.*, vol. 22, no. 1, pp. 21–30, Jan. 2001.

[69] O. Gutten, D. Bím, J. Řezáč, a L. Rulíšek, "Macrocycle Conformational Sampling by DFT-D3/COSMO-RS Methodology," *J. Chem. Inf. Model.*, vol. 58, no. 1, pp. 48–60, 2018.

[70] "TURBOMOLE V7.2, a Developement of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH." p. od r 2007.

[71] A. Klamt, "The COSMO and COSMO-RS solvation models," *WILEY Interdiscip. Rev. Mol. Sci.*, vol. 8, no. 1, 2018.

[72] D. A. C. Beck, D. O. V. Alonso, D. Inoyama, a V. Daggett, "The intrinsic conformational propensities of the 20 naturally occurring amino acids and

reflection of these propensities in proteins," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 34, pp. 12259–12264, Aug. 2008.

[73] J. Kleinjung, P. Bayley, a F. Fraternali, "Leap-dynamics: Efficient sampling of conformational space of proteins and peptides in solution," *FEBS Lett.*, vol. 470, no. 3, pp. 257–262, Mar. 2000.

[74] C.-L. Towse, G. Hopping, I. Vulovic, a V. Daggett, "Nature versus design: the conformational propensities of d-amino acids and the importance of side chain chirality," *Protein Eng. Des. Sel.*, vol. 27, no. 11, p. 447, 2014.

[75] M. Ropo, M. Schneider, C. Baldauf, a V. Blum, "First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids," *Sci. Data*, vol. 3, Apr. 2015.

[76] S. Mondal, D. S. Chowdhuri, S. Ghosh, A. Misra, a S. Dalai, "Conformational study on dipeptides containing phenylalanine: A DFT approach," *J. Mol. Struct. THEOCHEM*, vol. 810, no. 1–3, pp. 81–89, May 2007.

[77] W. Yu, X. Xu, H. Li, R. Pang, K. Fang, a Z. Lin, "Extensive Conformational Searches of 13 Representative Dipeptides and an Efficient Method for Dipeptide Structure Determinations Based on Amino Acid Conformers," *J. Comput. Chem.*, vol. 30, no. 13, pp. 2105–2121, Oct. 2009.

[78] W. Yu, Z. Wu, H. Chen, X. Liu, A. D. MacKerell Jr., a Z. Lin, "Comprehensive Conformational Studies of Five Tripeptides and a Deduced Method for Efficient Determinations of Peptide Structures," *J. Phys. Chem. B*, vol. 116, no. 7, pp. 2269–2283, Feb. 2012.

[79] L. F. Holroyd a T. Van Mourik, "Tyrosine-glycine revisited: Resolving the discrepancy between theory and experiment," *Chem. Phys. Lett.*, vol. 621, pp. 124–128, Feb. 2015.

[80] "Biopython: freely available Python tools for computational molecular biology and bioinformatics | Bioinformatics | Oxford Academic." [Online]. Available: https://academic.oup.com/bioinformatics/article/25/11/1422/330687. [Accessed: 29-Apr-2020].

[81] "PDB file parser and structure class implemented in Python | Bioinformatics | Oxford Academic." [Online]. Available: https://academic.oup.com/bioinformatics/article/19/17/2308/205793. [Accessed: 29-Apr-2020].

[82] M. Culka, J. Galgonek, J. Vymetal, J. Vondrasek, a L. Rulisek, "Toward Ab Initio Protein Folding: Inherent Secondary Structure Propensity of Short Peptides from the Bioinformatics and Quantum-Chemical Perspective," *J. Phys. Chem. B*, vol. 123, no. 6, pp. 1215–1227, Feb. 2019.

[83] M. Culka a L. Rulíšek, "Factors Stabilizing β-Sheets in Protein Structures from a Quantum-Chemical Perspective," *J. Phys. Chem. B*, vol. 123, no. 30, pp. 6453–6461, Aug. 2019.

[84] J. C. Phillips *et al.*, "Scalable molecular dynamics with NAMD," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1781–1802, Dec. 2005.

[85] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, a C. Simmerling, "Comparison of multiple amber force fields and development of improved protein backbone parameters," *Proteins: Structure, Function and Genetics*, vol. 65, no. 3. pp. 712–725, 15-Nov-2006.

[86] A. Onufriev, D. Bashford, a D. A. Case, "Modification of the generalized born model suitable for macromolecules," *J. Phys. Chem. B*, vol. 104, no. 15, pp. 3712–3720, Apr. 2000.

[87] J. Hostas a J. Rezac, "Accurate DFT-D3 Calculations in a Small Basis

Set," *J. Chem. Theory Comput.*, vol. 13, no. 8, pp. 3575–3585, Aug. 2017.

[88] S. Grimme, "Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory," *Chem. - A Eur. J.*, vol. 18, no. 32, pp. 9955–9964, Aug. 2012.

[89] S. Freeman, *Biological Science*. Pearson, 2003.