

UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO
ODDELEK ZA FIZIKO
MAGISTERSKI PROGRAM 2. STOPNJE FIZIKA
RAČUNALNIŠKA FIZIKA

Simon Perovnik

**GROBO-ZRNATO MODELIRANJE
MONOKLONSKIH PROTITELES IN NJIHOVA
LASTNA NIHANJA**

Magistrsko delo

MENTOR: prof. dr. Miha Ravnik

Ljubljana, 2024

Zahvale

Neizmerno sem hvaležen mojemu mentorju prof. dr. Mihi Ravniku za vso strokovno pomoč in nasvete pri pripravi tega dela. Najini sestanki so mi dali vedno novega zagona in v meni poglabljali veselje do raziskovanja. Predvsem pa hvala za resnično lep zgled mentorstva, ki bo ostajal z mano še dolgo po končanem študiju.

Posebej bi se želel zahvaliti tudi vsem kolegom, s katerimi mi je bilo dano si deliti študijske klopi, za vse pogovore, pomoč in pred-izpitne spodbude v duhu “*ma, bomo že!*”.

Nenazadnje bi se želel iskreno zahvaliti vsej svoji družini, predvsem ženi Eli ter mami in očetu za zgled in brezmejno podporo na moji študijski poti. Hvala bratomu Mateju in Sebastjanu ter sestrama Sonji in Luciji za vse skupaj preživete trenutke.

Grobo-zrnato modeliranje monoklonskih protiteles in njihova lastna nihanja

IZVLEČEK

Grobo-zrnati modeli monoklonskih protiteles so pogosto uporabljeni za *in silico* opis statike in dinamike. V magistrskem delu raziskemo dva pristopa konstrukcije grobo-zrnatih modelov — algoritem strojnega učenja gručenje k-means in metodo bistvene dinamike. Z metodama reproduciramo grobo-zrnate modele, ki so kvalitativno podobni uveljavljenim 3, 6 in 12-delčnim grobo-zrnatim modelom proteinov, s čimer pokažemo smiselnost njune uporabe. V drugem delu naloge lastne nihajne načine, ki smo jih izračunali za uporabo metode bistvene dinamike, koreliramo s hitrostjo agregacije, kvalifikatorjem proteinskih formulacij in z vsoto amplitud nihanj v CDR zankah. Rezultati iskanja linearnih korelacij kažejo na verjetno povezanost lastnih nihajnih načinov in mehanizmov agregacije. V tem kontekstu prepoznamo dva nihajna načina, pri katerih opazimo znatnejše korelacije.

Ključne besede: proteini, modeliranje, grobo-zrnati modeli, gručenje k-means, metoda bistvene dinamike, monoklonska protitelesa, lastni nihajni načini

Coarse-Grained Modelling of Monoclonal Antibodies and Their Vibration Modes

ABSTRACT

Coarse-grained models of monoclonal antibodies are often used for *in silico* description of statics and dynamics. In this thesis, we investigate two approaches to the construction of coarse-grained models — the k-means machine learning algorithm and the essential dynamics method. We reproduce coarse-grained models of proteins that are qualitatively similar to established 3-, 6-, and 12-particle coarse-grained models using these methods, demonstrating the validity of their application. In the second part of the thesis, we correlate the normal modes computed using the essential dynamics method with the aggregation rate, a critical quality attribute of proteins and with the sum of amplitudes in CDR regions. The results of the linear correlation indicate a likely correlation between the normal modes and the aggregation mechanisms. In this context, we identify two oscillatory modes where more significant correlations are observed.

Keywords: proteins, modeling, coarse-grained models, k-means clustering, essential dynamics method, monoclonal antibodies, normal modes

Kazalo

1	Uvod	11
2	Teoretične osnove	13
2.1	Proteini	13
2.2	Monoklonska protitelesa	14
2.2.1	Agregacija protiteles	15
2.3	Grobo-zrnati modeli	16
2.3.1	Grobo-zrnati modeli proteinov	17
2.3.2	Grobo-zrnati modeli monoklonskih protiteles	18
3	Metode	21
3.1	Gručenje k-means	21
3.2	Elastični mrežni model	24
3.3	Analiza lastnih nihajnih načinov	27
3.4	Metoda bistvene dinamike	29
4	Rezultati	33
4.1	Grobo-zrnati modeli	33
4.1.1	GZ modeliranje z metodo k-means	33
4.1.2	GZ modeliranje z metodo bistvene dinamike	38
4.2	Korelacije lastnih nihajnih načinov in hitrosti agregacije	41
4.2.1	Frekvence lastnih načinov	42
4.2.2	Nihanja v variabilnih regijah	46
5	Zaključek	49
6	Literatura	51

1. Uvod

Monoklonska telesa so pomembna raziskovana vrsta proteinov, z neposredno uporabo npr. v biofarmacevtiki [1]. V primerjavi s klasičnimi zdravili imajo zaradi svoje velike specifičnosti pri načinu vezave na druge molekule zelo malo stranskih učinkov, saj se malo molekul veže na nezaželena mesta [2]. Uporablja se jih pri zdravljenju raka, različnih avtoimunih boleznih [3, 4, 5] in nedavno tudi za zdravljenje COVID-19 [6]. Kljub znatenemu razvoju v zadnjih desetletjih pa ostaja še veliko odprtih fizikalnih izzivov v povezavi s stabilnostjo pripravljenih zdravilnih učinkov [7, 8, 9, 10].

V procesu razvoja biofarmacevtikov se v kontekstu nadzora kvalitete in zagotavljanja varnosti uporabe preverja več t.i. kvalifikatorjev proteinskih formulacij. Gre za lastnosti formulacij, kot so hitrost agregacije, viskoznost in stabilnost konformacije proteina [11]. Takšne kvalifikatorje je pogosto težko napovedati, zlasti v primerih formulacij velike gostote, kakršne so običajno potrebne za pripravo zdravil [7, 12, 13]. Vrednosti kvalifikatorjev merimo z eksperimentalnimi metodami, kot so statično in dinamično svetlobno sipanje, UV spektroskopija variabilne dolžine optične poti in različni reološki postopki [13, 14, 15].

Kvalifikatorji so pogojeni z vrsto spremenljivk, kot so pH, temperatura, koncentracija ionov in aminokislinska sekvenca [16]. Ker je težko zapisati eksplisitne zveze, ki bi opisovale obnašanje kvalifikatorjev, se aktivno razvija različne metode napovedovanja vrednosti kvalifikatorjev z uporabo strojnega učenja [17, 18] in s simulacijami molekularne dinamike [19, 20].

Ker so monoklonska protitelesa velike molekule, se simulacije molekularne dinamike pogosto opravlja z uporabo grobo-zrnatih (GZ, ang. coarse-grained) modelov. To so modeli, kjer molekularno strukturo namesto z vsemi atomi predstavimo s poenostavljenou reprezentacijo [21, 22, 23, 24]. Velika prednost takšnih modelov je v tem, da se jih da sklapljati tako med seboj kot z modeli molekul, ki so predstavljeni z vsemi atomi. Tako lahko znotraj ene simulacije predstavimo vsak tip molekule v zanj primerni ločljivosti, glede na fizikalni pojav, ki ga želimo poustvariti v simulaciji [25]. Že uveljavljeni GZ modeli pogosto ne poenostavijo strukture monoklonskih protiteles dovolj, da bi jih lahko uporabili v simulacijah molekularne dinamike, zaradi česar je bilo predlaganih več za te vrste molekul specifičnih GZ modelov, pri katerih monoklonsko protitelo predstavimo z vsaj 3 ali največ 26 gradniki [16, 19, 20, 26].

Prvi cilj magistrskega dela je poiskati smiseln poenostavitev strukture monoklonskih protiteles z uporabo GZ modela, katerega bi se dalo kasneje uporabiti tudi na drugih vrstah molekul. Preizkusili bomo dva pristopa — metodo strojnega učenja k-means, s katero bomo določevali mesta GZ gradnikov z iskanjem predelov v strukturi, kjer so atomi gosto skupaj, in metodo bistvene dinamike. Pri slednji je poenostavljena struktura definirana na tak način, da ohrani dinamične lastnosti, ki jih odraža nepoenostavljen vse-atomski model. Smiselnost rezultatov bomo

POGLAVJE 1. UVOD

preverjali s primerjavo z GZ modeli monoklonskih protiteles, ki se že pojavljajo v literaturi [16, 19, 20, 26]. Drugi cilj te naloge je preveriti, ali lahko iz lastnosti lastnih nihajnih načinov monoklonskih protiteles sklepamo o hitrost agregacije takih molekul. To bomo preverjali z iskanjem močnih linearnih korelacijskih lastnosti in s hitrostjo agregacije.

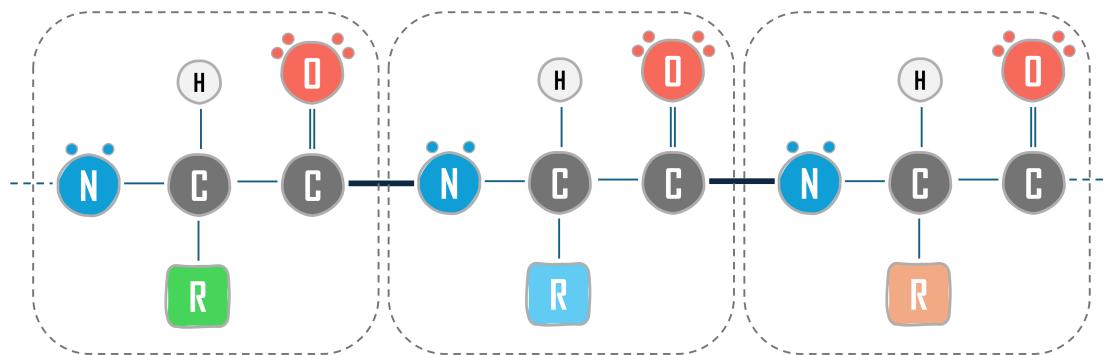
Struktura magistrskega dela je naslednja: poglavje Teoretične osnove začnemo s povzetkom osnovnih značilnosti molekul proteinov in molekul monoklonskih protiteles, kjer vpeljemo hitrost agregacije in pojasnimo, kakšni so mehanizmi agregacije. Sledi pregled GZ modelov, pri katerem najprej osvetlimo nekatere splošne lastnosti takšnih modelov, nato pa podamo nekaj primerov proteinskih GZ modelov. Poglavlje sklenemo s pregledom različnih GZ modelov iz literature, ki se jih uporablja v kontekstu monoklonskih protiteles. V tretjem poglavju predstavimo gručenje k-means, elastični mrežni model, analizo lastnih nihajnih načinov in metodo bistvene dinamike. V četrtem poglavju predstavimo in komentiramo rezultate. V prvem delu pokažemo skonstruirane GZ modele monoklonskega protitelesa, osnovane na metodah k-means in metodi bistvene dinamike, ter jih analiziramo in ovrednotimo glede na referenčne GZ modele. Drugi del poglavja je namenjen iskanju korelacij med lastnostmi lastnih nihajnih načinov molekul monoklonskih protiteles in hitrostjo agregacije. Ob koncu dela podamo zaključek, v katerem povzamemo ključne rezultate tega dela in predlagamo nadaljnje možnosti za raziskavo.

2. Teoretične osnove

Poglavje se začne s splošnim pregledom molekul proteinov in monoklonskih protiteles, ki so podvrsta proteinov. Vpeljanih je nekaj ključnih pojmov in poimenovanj, s poudarkom na molekularni strukturi. Sledita podpoglavlje o agregaciji proteinov ter podpoglavlje o GZ modelih proteinov in monoklonskih protiteles, kjer začnemo s splošnim pregledom, kaj GZ modeli so, zakaj so uporabni in kakšno je njihovo fizikalno ozadje. Nadalje predstavimo nekaj konkretnih GZ modelov proteinov in monoklonskih protiteles, ki se pojavljajo v literaturi.

2.1 Proteini

Proteini so makromolekule, ki znotraj celic opravljajo številne funkcije; sodelujejo pri transportnih procesih, prepisovanju DNK-ja, imunskem odzivu, ohranjanju strukture celične stene in drugih procesih [27]. Sestavljene so iz nekaj deset ali pa tudi več tisoč aminokislin, ki se med seboj povežejo v eno ali več verig. Aminokisline se med sabo povezujejo s peptidnimi vezmi [28], kar je prikazano na Sliki 2.1, kjer je vsaka aminokislina shematsko zaobjeta s sivo prekinjajočo črto. Aminokisline lahko razdelimo na štiri dele — centralni ogljik- α , na katerega je povezan vodikov atom, karboksilno skupino (CO), amino skupino (NH_2) in stransko verigo (označena na Sliki 2.1 z R). Aminokisline se tako razlikujejo med seboj zgolj po kemijski sestavi stranske verige. Najpreprostejšo najdemo pri alaninu (CH_3), največjo pa pri triptofanu ($\text{C}_9\text{H}_8\text{N}$).

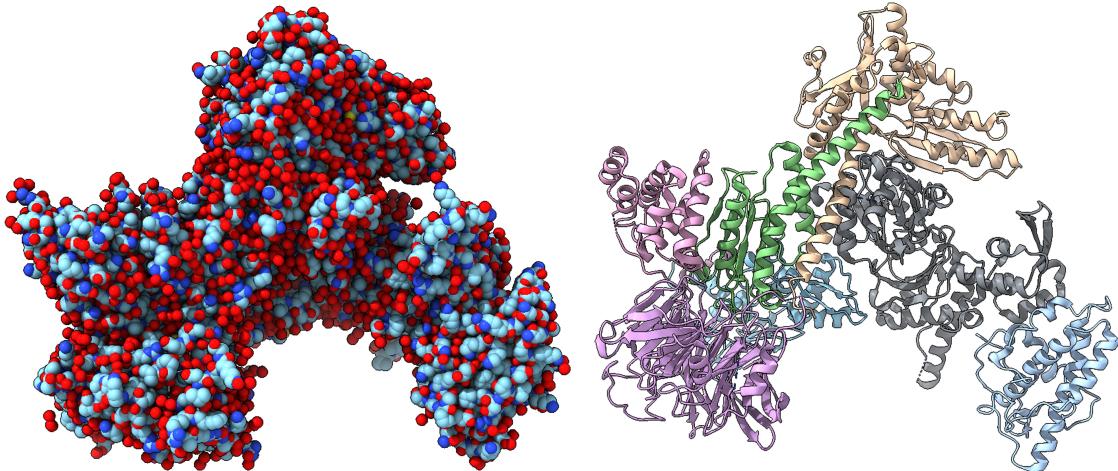


Slika 2.1: Tri aminokisline, povezane s peptidnimi vezmi (označene z odebeljeno črto) med atomom ogljika in dušika. Stranske verige označene z R so različno obarvane, saj se razlikujejo med posameznimi aminokislinami.

Kljub temu da so proteini linearne molekule, ker so sestavljene iz zaporedja aminokislin, je njihova struktura precej bolj zapletena, saj se preko procesa proteinskega zvijanja tvorijo nekovalentne vezi med nezaporednimi aminokislinami [29]. Pogoste strukture, ki se tvorijo v tem procesu, so vijačnice- α in ploščice- β . Končna

POGLAVJE 2. TEORETIČNE OSNOVE

3-dimenzionalna struktura proteina (imenovana tudi nativna konformacija) je natančno določena z aminokislinsko sekvenco in ustreza energijsko minimizirani strukturi. Primer strukture proteina je z dvema vizualizacijama, atomističnim prikazom in s prikazom z vijačnicami- α in ploščicami- β , podan na Sliki 2.2.



Slika 2.2: Vizualizaciji proteinskega kompleksa Arp2/3 (PDB: 1K8K) [30]. V atomističnem prikazu na levi so ogljikovi atomi obarvani svetlo modro, kisikovi rdeče, dušikovi temno modro ter žveplovi rumeno. Zaradi boljše preglednosti atomi vodika niso vključeni v vizualizacijo. Prikaz z vijačnicami- α in ploščicami- β je prikazan na desni, kjer je vsak funkcionalni segment proteina prikazan s svojo barvo. Struktura je vizualizirana s programom ChimeraX [31].

2.2 Monoklonska protitelesa

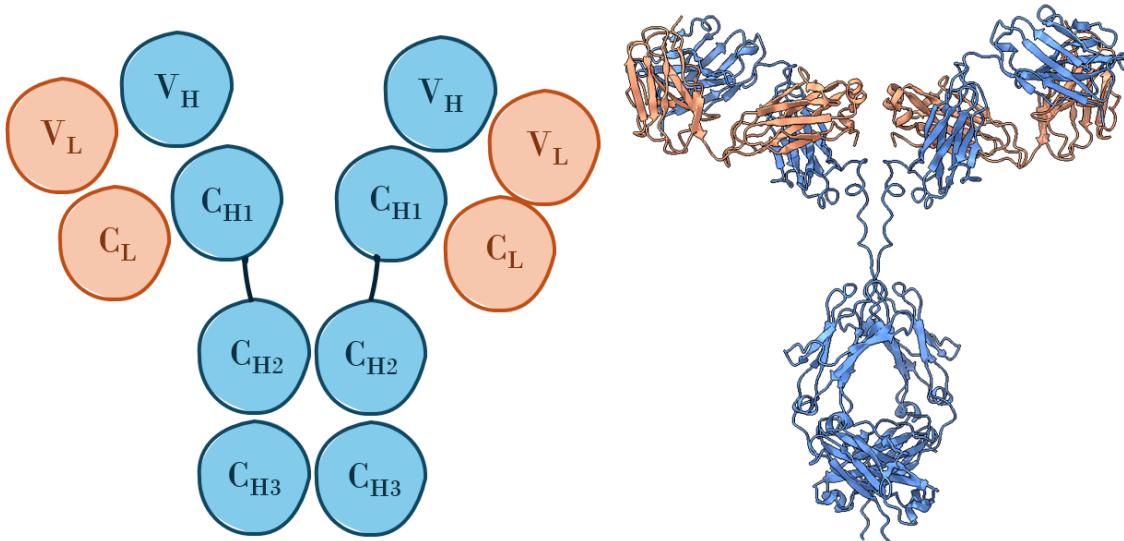
Protitelesa so vrsta proteinov, ki sodelujejo pri imunskejem odzivu organizma. Gre za molekule, ki lahko nase vežejo različne antigene — to so molekule, ki v telesu sprožijo imunski odziv [32]. Po večini so tudi antigeni proteini, ki pa se lahko nahajajo na površini kakšnega tujka, recimo virusa. Ko organizem prepozna tujek v telesu, na podlagi njegove strukture generira (od tod tudi ime *anti-gen*) ustrezna protitelesa, ki z vezavo na epitom antigena poskrbijo, da le-ta ne more več opravljati svoje funkcije in postane neškodljiv telesu [32]. Formulacije protiteles lahko delimo na poliklonske in monoklonske, glede na to, na koliko različnih antigenov se vežejo [33]. Monoklonska protitelesa so tako skupek protiteles, ki se vsa vežejo na isti antigen. Pogosto posamezno molekulo znotraj take skupine imenujemo za monoklonsko protitelo. Dodatno jih lahko razdelimo na 5 razredov glede na njihovo strukturo — M, A, D, E in G. V tej nalogi bomo govorili le o slednjih.

Monoklonska protitelesa so simetrične molekule v obliki črke "Y". Sestavljajo jo štiri proteinske verige, dve krajsi, imenovani tudi lahki, in dve daljši, imenovani težki verigi, ki sta med seboj povezani preko disulfidnih mostičkov. Celotno molekulo lahko razdelimo na 12 funkcionalnih domen, kot je prikazano na Sliki 2.3. Prepoznamo lahko naslednje strukture:

Konstantne regije — predeli težke in lahke verige, označeni s črko C, ki so večinoma enaki za vsa protitelesa. Pripis spodaj definira, kateri verigi pripada in kako je oštrevilčena.

Variabilni regiji — predela težke in lahke verige, kamor se veže antigen, označen s črko V. Je specifičen za vsako monoklonsko protitelo.

CDR zanke (ang. *complementarity-determining regions*) — del variabilnih regij, ki se nahaja na površini proteina in se pripne na antigen. Znotraj vsake variabilne regije se nahajajo tri CDR zanke, kar da skupno 12 zank za celotno protitelo.



Slika 2.3: Shematski prikaz strukture (levo) in prikaz z vijačnicami- α in ploščicami- β (desno) monoklonskega protitelesa Tocilizumab. Z oranžno sta obarvani lahki verigi, z modro pa težki. Desna vizualizacija je bila pripravljena s podatki iz [18] s programom ChimeraX [31].

Ker človeško telo ne zmore proizvesti poljubnega protitelesa, se različna obolenja in bolezni zdravi s sintetično proizvedenimi protitelesi. Monoklonska protitelesa so odličen kandidat za takšna zdravila, saj zaradi svoje monospecifičnosti povzročajo razmeroma malo nezaželenih učinkov.

2.2.1 Agregacija protiteles

V visoko-koncentriranih formulacijah protiteles, kakršne se uporablja za zdravila, pogosto pride do procesa agregacije. S tem izrazom označimo dva različna procesa — reverzibilno samopovezovanje in nereverzibilno (pravo) agregacijo [34]. Pri prvem gre za tvorjenje oligomerov (skupkov proteinov), ki so topni v vodi in jih je mogoče spet ločiti z uporabo pufrov. Pogosto vodi k povečani viskoznosti formulacije. Pri nereverzibilni agregaciji pa se tvorijo oligomeri, ki so lahko topni ali ne, predvsem pa obstaja velika možnost precipitacije (tvorjenja oborine) [35]. Oba procesa predstavljata velik izziv v farmacevtski industriji, saj sta pogojena tako z okoljskimi faktorji, ki jih je težko nadzorovati, kot tudi z lastnostmi samega protitelesa. Slednje se izkaže za posebej težavno, saj je težko vnaprej napovedati, ali bo protitelo hitreje agregiralo. V ta namen se razvija različne modele strojnega učenja [17, 18], ki glede na set eksperimentalnih podatkov napovedujejo hitrost agregacije monoklon-skih protiteles. Eksperimentalno izmerjena hitrost agregacije je določena iz meritev koncentracije neaggregiranih molekul [C_M], za katero se privzame, da sledi dinamiki

$$\frac{d[C_M]}{dt} = -v_{\text{agr}}[C_M]^2, \quad (2.1)$$

ki ima rešitev

$$\frac{1}{[C_M]} = v_{\text{agr}}t + \frac{1}{[C_M]_0}, \quad (2.2)$$

pri kateri $[C_M]_0$ označuje začetno koncentracijo monoklonskih protiteles, v_{agr} pa hitrost agregacije.

Alternativno klasičnim eksperimentalnim meritvam hitrosti agregacije predstavljajo *in silico* eksperimenti molekularne dinamike, kjer začnemo simulacijo z neagregiranimi monoklonskimi protitelesi in spremljamo, kako se pod vplivom različnih dejavnikov tvorijo agregati. Simulacije molekularne dinamike z monoklonskimi protitelesi so pogosto numerično zelo zahtevne, saj so same molekule razmeroma velike in zahtevajo veliko število računskih operacij, kar rešujemo z uporabo GZ modelov.

2.3 Grobo-zrnati modeli

Simulacije molekularne dinamike so pogosto uporabljene za študije biokemijskih procesov [36]. Izkaže se, da je makromolekule, kot so ogljikovi-hidrati, DNK, lipidi in proteini težko simulirati na tak način, saj veliko število atomov vodi tudi k velikemu številu računskih operacij, potrebnih za določanje dinamike. Ta problem lahko naslovimo z uporabo GZ modelov. Glavna ideja GZ modelov je v tem, da je možno strukturo makromolekule poenostaviti, a hkrati še vedno ohraniti nekatere bistvene fizičke lastnosti. Pri tem se lahko poslužimo dveh postopkov — če prepoznamo, kateri predeli molekule ne vplivajo na intermolekularne interakcije, ki jih opazujemo (recimo tvorjenje aggregatov), lahko te dele izpustimo iz simulacije. Podobno lahko prepoznamo tudi predele, ki bi jih bilo sicer potrebno ohraniti, a bi jih lahko predstavili manj detajlno, z manj gradniki. V tem primeru lahko več atomov združimo v t.i. psevdo-atome. Oba načina vodita k istemu cilju, ki je redukcija prostostnih stopenj molekule.

Bistvena lastnost GZ modela je njegova grobost, oziroma kako znatna je poenostavitev strukture. Ker biokemijski procesi potekajo na različnih nivojih — procesni mehanizmi so lahko odvisni od globalnih lastnosti molekule, kot je skupen naboj, ali pa zelo lokalnih, kot je polarnost kakšnega vezavnega mesta — mora biti grobost modela prilagojena konkretnemu biokemijskemu procesu in makromolekulam, ki jih simuliramo. V splošnem vedno iščemo največjo poenostavitev, ki si jo lahko privoščimo, ne da bi izgubili za simulacijo pomembne lastnosti molekule.

Poleg grobosti obstajajo še druge pomembne lastnosti, ki ločijo GZ modele med seboj. Naštejemo lahko nekaj izmed vprašanj, ki jih GZ modeli naslavljajo na različne načine:

- Kako preslikati fizičke lastnosti (kot so naboj, dipolni moment, hidrofobnost, ipd.) iz reprezentacije z vsemi atomi na grobo-zrnato?
- Kako določiti lokacije in velikosti psevdo-atomov?
- Kakšen je primeren fizički opis interakcij med posameznimi psevdo-atomi in med makromolekulami glede na fizičke lastnosti procesa, ki ga modeliramo?

Zadnje vprašanje je naslovljeno v kontekstu določanja t.i. polja sil. Polja sil so sestavljena iz različnih fizikalnih potencialov, ki opisujejo inter-molekularne in intra-molekularne interakcije. So analogni klasičnemu opisu sistema preko sil, ki nanj delujejo, a opisujejo interakcije na atomistični ravni. Prispevki k polju sil so običajno dveh vrst, vezani in nevezani. Prvi vključujejo interakcije, ki vežejo delce skupaj in modelirajo molekulske, kovalentne in druge vezi, drugi pa opisujejo odbojne, sterične (pogosto opisane z Lennard-Jonesovim potencialom) in elektrostaticne učinke. Splošno polje sil je podano z enačbo (2.3), pri kateri je celotni potencial U_{total} razdeljen na vezani U_{vezani} in nevezani U_{nevezani} člen, ki je opisan z Lennard-Jonesovim $U_{\text{Lennard-Jones}}$ in elektrostatičnim $U_{\text{elektrostatični}}$ členom:

$$\begin{aligned} U_{\text{total}} &= U_{\text{bonded}} + U_{\text{non-bonded}}, \\ U_{\text{non-bonded}} &= U_{\text{electrostatic}} + U_{\text{Lennard-Jones}}. \end{aligned} \quad (2.3)$$

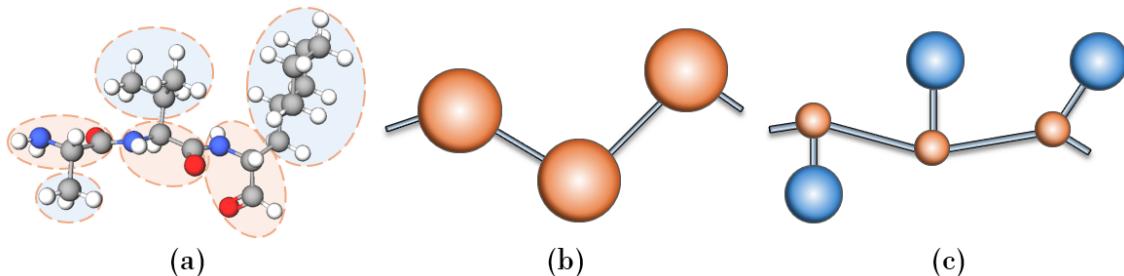
Vsa navedena vprašanja generirajo širok nabor možnosti, kako skonstruirati GZ model. Jasno je, da ni mogoče govoriti o idealnem GZ modelu, saj je primernost njegove uporabe vedno pogojena s konkretnimi molekulami in procesom, ki ga modeliramo. Ker so si strukture makromolekul zelo različne, so modeli navadno ustvarjeni za določen tip makromolekul, kot so ogljikovi hidrati [37], DNK [38], lipidi [39] in proteini [40].

2.3.1 Grobo-zrnati modeli proteinov

Strukturo proteina lahko poenostavimo na različne načine, kar se kaže v številčnosti uveljavljenih GZ modelov. Eden izmed prvih takšnih modelov je bil razvit za potrebe simuliranja proteinskega zvijanja, do danes pa so bili GZ modeli proteinov uspešno uporabljeni tudi pri modeliranju proteinske agregacije, pri preučevanju fleksibilnosti, napovedi strukture proteina ipd. [40].

GZ modeli vedno izhajajo iz referenčne strukture makromolekule. Pogosto so takšne strukture pridobljene s pomočjo eksperimentalnih postopkov, kot so krioelektronsko mikroskopiranje, rentgenska kristalografija in jedrska magnetna resonanca [41, 42, 43]. Pri določanju strukture proteina pa se lahko poslužimo tudi metode strojnega učenja, imenovane AlphaFold, kjer je struktura proteina napovedana direktno iz aminokislinske sekvence proteina z uporabo nevronskih mrež, glede na veliko podatkovno bazo struktur proteinov [44]. Bistvena prednost te metode je v preprostosti uporabe, saj je javno dostopna in jo lahko uporabljamo preko spleta.

GZ modele proteinov lahko razdelimo glede na število gradnikov, s katerimi opisujejo eno aminokislino [45]. Najpreprostejši modeli celotno aminokislino predstavijo z zgolj enim psevdo-atomom, ki je pogosto postavljen v masno središče aminokislinske, fizikalne lastnosti atomov, kot je naboj, pa so preslikane na psevdo-atom kot povprečje ali vsota prispevkov posameznih atomov. Nekoliko bolj detajlni modeli vključijo dodaten psevdo-atom za reprezentacijo stranske verige aminokislinske. Ta dva primera sta prikazana na Sliki 2.4, na sličicah b) in c). GZ modeli večje resolucije vključijo dodatne psevdo-atome za ostale atome znotraj glavne verige aminokislinske, ali pa vključijo dodatne atome za opis stranske verige, pri kateri je število teh atomov prilagojeno razsežnosti stranske verige posamezne aminokislinske. Razvitih pa je bilo tudi nekaj modelov, pri katerih je število psevdo-atomov manjše od števila aminokislins [45]. Takšni modeli so pogosto uporabljeni na večjih proteinih ali proteinskih kompleksih, kjer ostali modeli ne zmanjšajo števila prostostnih stopenj dovolj, da bi jih bilo mogoče simulirati z metodami molekularne dinamike.



Slika 2.4: (a) Vizualizacija treh aminokislin, povezanih znotraj proteina. Atomi stranske verige so prikazani na modri podlagi, atomi glavne verige pa na oranžni. Atomi ogljika so obarvani sivo, atomi kisika rdeče, atomi dušika temno modro in atomi vodika belo. (b) Prikaz GZ modela proteina, kjer je vsaka aminokislina predstavljena z enim psevdo-atomom. (c) GZ model proteina, kjer en psevdo-atom opisuje stransko verigo, drugi pa glavno.

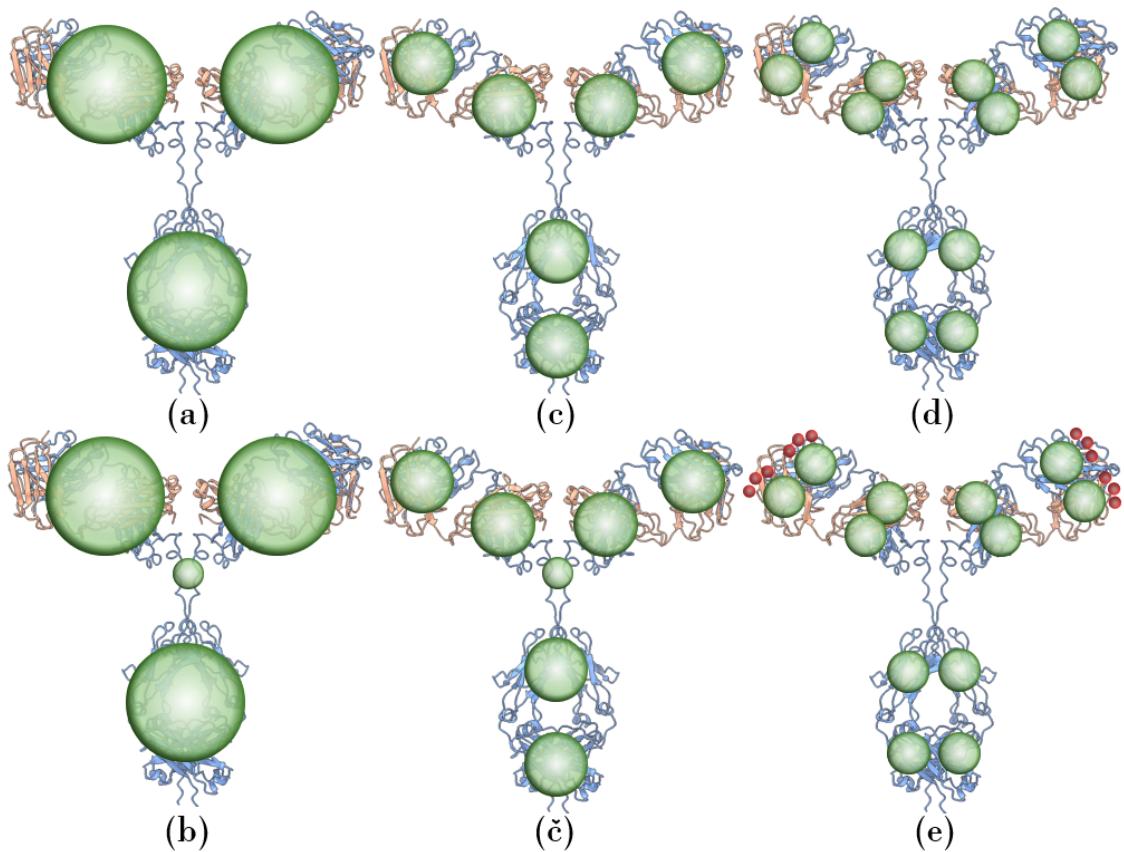
2.3.2 Grobo-zrnati modeli monoklonskih protiteles

Monoklonska protitelesa sodijo med velike proteine, saj navadno vsebujejo več kot tisoč aminokislin, ki skupaj vsebujejo okoli deset tisoč atomov. Zaradi velike razšenosti so simulacije molekularne dinamike možne po večini zgolj z uporabo GZ modelov, pri katerih je protein predstavljen z mnogo manj psevdo-atomi, kot vsebuje aminokislina. Pravzaprav se izkaže, da je način določevanja mest psevdo-atomov v teh primerih povsem drugačen kot pri veliki večini ostalih proteinskih GZ modelov — namesto nivoja aminokislin postane relevanten nivo funkcionalnih domen, ki so prikazana na Sliki 2.3. Tako lahko GZ modele monoklonskih protiteles ločimo na dve skupini: na tiste, kjer je število psevdo-atomov enako ali presega število funkcionalnih domen (to je 12 domen), in na tiste, kjer je število psevdo-atomov manjše od števila funkcionalnih domen.

Najpreprostejši GZ model monoklonskih protiteles vsebuje zgolj 3 psevdo-atome, po enega za vsak ‐krak‐ molekule [16]. Na tak način 4 funkcionalne domene združimo v en gradnik. Bolj podrobni GZ modeli dodajo še en psevdo-atom, ki ga postavijo v tečajno območje (ang. hinge region), s čimer omogočijo prosto gibanje krakov molekule [46]. Razširitev najpreprostejšega modela vključuje 6 psevdo-atomov, po 2 na vsak krak molekule, oziroma enega za opis dveh funkcionalnih domen. V takšnem modelu je sedaj variabilni del molekule predstavljen s svojim psevdo-atomom, kar je pomembna nadgradnja, saj se največ razlik med monoklonskimi protitelesi nahaja prav tam. Podobno obstaja tudi za takšen model razširitev z dodatnim gradnikom v tečajnjem območju [46].

Najbolj razširjeni in uporabljeni so GZ modeli monoklonskih protiteles z 12 psevdo-atomi [19, 20, 47, 48], kjer je vsaka funkcionalna domena predstavljena s svojim psevdo-atomom. Glede na rezultate različnih simulacij se zdi, da je to pogosto najprimernejši opis monoklonskih protiteles. Dodatno razširitev predstavlja dodajanje psevdo-atomov za opis vsake izmed CDR zank in tečajnega območja, kar doprinese dodatnih 14 gradnikov v GZ model, ki torej sestoji iz 26 [19]. Vsi omenjeni GZ modeli so prikazani na Sliki 2.5.

Za razliko od proteinskih GZ modelov, pri katerih so aminokisline dobro definirane strukturne enote in je zato enostavno določiti, kateri atomi se bodo združili v določen psevdo-atom, se izkaže, da funkcionalne domene nimajo tako jasno definiranih razmejitev. Definirane so namreč glede na 3-dimenzionalno strukturo molekule.



Slika 2.5: Referenčni GZ modeli monoklonskih protiteles s 3 (a), 4 (b), 6 (c), 7 (č), 12 (d) in 26 (e) psevdo-atomi. Psevdo-atomi, ki predstavljajo CDR zanke v 26-delčnem modelu, so obarvani rdeče.

Prepoznamo jih lahko kot ločene skupine atomov, ki so med seboj bolj prepletene, a so meje med njimi nejasne. Ker podatki o strukturah monoklonskih protiteles ne vsebujejo informacij o razmejitvah na funkcionalne domene, so GZ modeli pogosto skonstruirani na način, da te razmejitve določi uporabnik glede na lastno presojo ali ob pomoči kakšnega drugega programskega orodja.

3. Metode

V sledečem poglavju se nahaja opis glavnih metod, uporabljenih v magistrskem delu. V prvem podpoglavlju začnemo z opisom metode strojnega učenja, imenovane k-means, kateremu sledijo tri med seboj povezana podpoglavlja — GZ model, ki ga generira metoda bistvene dinamike je namreč osnovan na analizi lastnih nihajnih načinov in elastičnega mrežnega modela. Vse tri metode opišemo in podamo njihova fizikalna ozadja.

3.1 Gručenje k-means

Metodo k-means uvrščamo med modele nenadzorovanega strojnega učenja. Strojno učenje povzema definicijo: “*Računalniški program se pri opravljanju nalog T (ang. tasks) uči iz izkušnje E (ang. experience), glede na izbrano mero uspešnosti P (ang. performance measure), če se uspešnost P pri opravljanju nalog T izboljšuje z izkušnjami E.*” [49] Gre torej za optimizacijske algoritme, ki preko prilaganja modelskih parametrov iščejo minimum uporabljene cenovne funkcije, oziroma maksimalno mero uspešnosti. Da je strojno učenje nenadzorovano, pomeni, da so podatki, posredovani modelu, neoznačeni. Primer takšnega strojnega učenja so modeli gručenja, kjer je cilj, da model znotraj seta podatkov prepozna, kateri elementi so si med seboj podobni, in jih združi v nekaj gruč. Gručenje je dober način za iskanje skritih vzorcev znotraj velikih podatkovnih setov. Na področju fizike so bili modeli gručenja uporabljeni za ločevanje različnih vrst ledu [50], identifikacijo karakterističnih fotonskih nihajnih načinov znotraj nanostruktur [51] in v medicinski fiziki za identifikacijo tumorjev iz spektroskopskih slik [52].

Modele gručenja lahko ločimo glede na to, ali je gručenje popolno (vsak element pripada eni gruči) ali delno (nekateri elementi niso razvrščeni v nobeno gručo). Prav tako lahko ločimo med seboj ekskluzivno ali prekrivajoče gručenje glede na to, največ koliko gručam lahko element pripada. V primeru, ko meje gruč niso jasno definirane, lahko govorimo o mehkem gručenju, kar pomeni, da je pripadnost elementa določeni gruči podana z določeno verjetnostjo. Gručenje k-means v tem oziru uvrščamo med popolna in ekskluzivna gručenja.

Gručenje k-means lahko definiramo na naslednji način: Obravnavajmo set neoznačenih podatkov $\{\mathbf{x}_n\}_{n=1}^N$, kjer je $\mathbf{x}_n \in \mathbb{R}^p$ in je p dimenzija faznega prostora podatkov. Definiramo lahko K gruč s središči $\{\mu_k\}_{k=1}^K$, kjer je tudi $\mu_k \in \mathbb{R}^p$. Središča gruč

$$\mu_k = \frac{1}{|\mathcal{S}_k|} \sum_{\mathbf{x}_n \in \mathcal{S}_k} \mathbf{x}_n, \quad (3.1)$$

ki so definirana kot povprečni elementi gruče (od tod izvira tudi ime metode), natančno razdelijo celoten fazni prostor na K območij. S \mathcal{S}_k smo označili gručo podatkov \mathbf{x}_n , ki pripadajo območju k s središčem μ_k . Metoda k-means ob podanem

POGLAVJE 3. METODE

številu K razdeli set podatkov $\{\mathbf{x}_n\}$ na K gruč, ki so enolično definirane s središči $\{\mu_k\}$, tako da je cenovna funkcija

$$\mathcal{C}(\{\mathbf{x}, \mu\}) = \sum_{k=1}^K \sum_{\mathbf{x}_n \in S_k} (\mathbf{x}_n - \mu_k)^2 \quad (3.2)$$

minimizirana. V fizikalnem smislu je cenovna funkcija hkrati tudi vsota vztrajnostnih momentov, pri katerih je masa vseh točk normirana na 1, središča gruč μ pa ustrezajo njihovim težiščem. Bistvo gručenja k-means se skriva v iterativnem algoritmu, ki ga lahko razdelimo na korake:

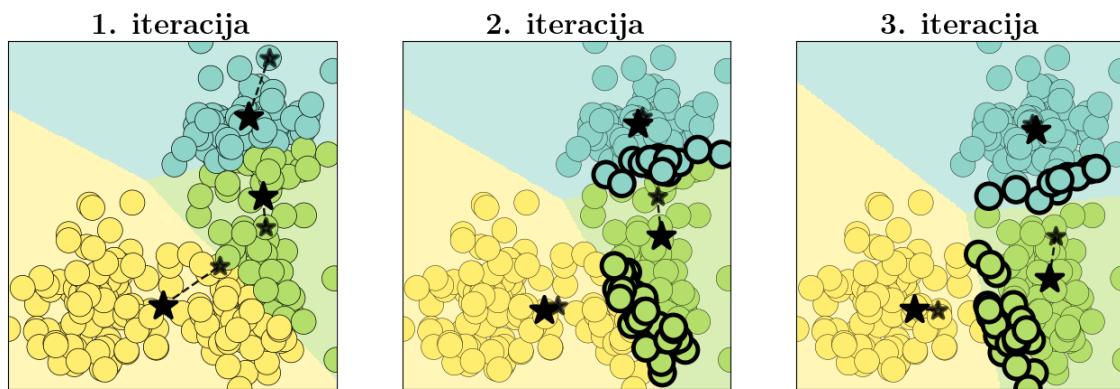
1. Izberemo K in začetna središča μ_k .
2. **Razdelitev v gruče:** vsak element \mathbf{x}_n pripisemo geometrično najbližjemu središču, kar lahko zapišemo kot

$$S_k(t) = \left\{ \mathbf{x}_n : \|\mathbf{x}_n - \mu_k(t)\|^2 \leq \|\mathbf{x}_n - \mu_j(t)\|^2 \quad \forall j, \quad 1 \leq j \leq K \right\}. \quad (3.3)$$

3. **Posodobitev μ :** ponovno izračunamo masna središča gruč $\mu_k(t+1)$ glede na enačbo (3.1).

Koraka 2 in 3 ponavljamo do konvergencije, ko se cenovna funkcija ne spreminja več znatno.

Izkaže se, da tak algoritem vedno konvergira v lokalni minimum, ne zagotavlja pa, da bomo našli globalni minimum, zaradi česar se ga vedno izvede večkrat pri različnih začetnih inicializacijah središč μ_k in uporabi rešitev z najnižjo cenovno funkcijo. Primer nekaj iteracij izvedenega algoritma je prikazan na Sliki 3.1. Skupna časovna zahtevnost algoritma je $\mathcal{O}(pKNi)$, kjer i predstavlja število iteracij.

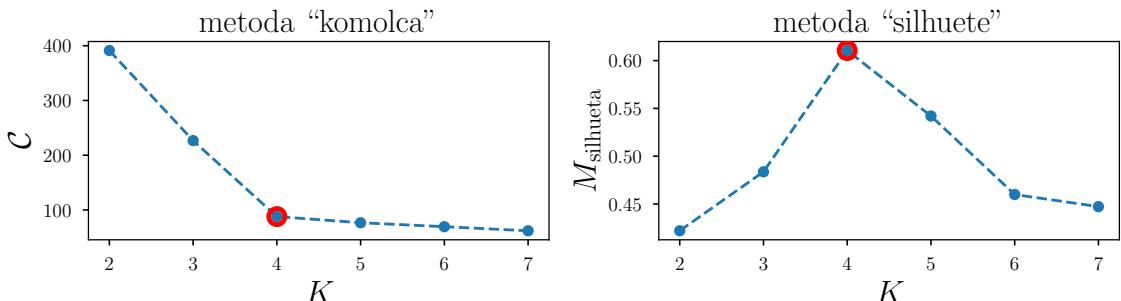


Slika 3.1: Gručenje k-means, izvedeno na primeru treh delno prekrivajočih se gruč. Prikazane so prve tri iteracije algoritma. Z zvezdicami so označena masna središča, pri čemer manjše zvezdice označujejo središče iz prejšnje iteracije. V primeru prve iteracije manjše zvezdice označujejo naključno inicializirana središča. Vsi elementi so razdeljeni v tri gruče. Točke, ki so bile v novi iteraciji pripisane drugi gruči kot prej, so odenbeljene.

Glavni modelski parameter gručenja k-means je število gruč K . Ker gre za nenadzorovano učenje, pri katerem sami pogosto ne poznamo vzorca, ki ga iščemo v podatkih, ni vedno jasno, kakšen je smiseln K . Predvsem v problemih, kjer je fazni prostor več kot 2-dimenzionalen, je težko uvideti, ali je rešitev pri določenem K prava, ali pa se skriva v podatkih še kakšna gruča, ki bi jo lahko ločili od ostalih. V pomoč nam je lahko t.i. "metoda komolca", pri kateri algoritom izvedemo za različne vrednosti K in spremjamamo, kako se zniža vrednost cenovne funkcije. Ko pridemo do preloma, oziroma ko dodatna gruča le malo zmanjša vrednost cenovne funkcije, smo presegli smiseln K . V primerih, ko so gruče različno številčne, je pogost problem, da prelom ni tako znaten in nam metoda ni v pomoč. Primer pomenljivega preloma je prikazan na Sliki 3.2. Poleg metode "komolca" nam je lahko v pomoč tudi t.i. metoda "silhuete", pri kateri algoritom prav tako izvedemo za različna števila gruč K in spremjamamo metriko

$$M_{\text{silhueta}} = \left\langle \langle \|\mathbf{x}_n - \mathbf{x}_j\| \rangle_{\mathbf{x}_j \in S_k}, \langle \|\mathbf{x}_n - \mathbf{x}_l\| \rangle_{\mathbf{x}_l \in S_k} \right\rangle_n, \quad (3.4)$$

kjer sta za vsak element podatkovnega seta $\{\mathbf{x}_n\}$ izračunani dve vrednosti. Prva je povprečna razdalja elementa \mathbf{x}_n do elementov naslednje najbližje gruče $S_{k'}$, druga pa povprečna razdalja elementa \mathbf{x}_n do elementov iste gruče S_k . Razliko teh dveh vrednosti imenujemo silhueta, metrika M_{silhueta} pa torej predstavlja povprečno silhueto. Gre torej za mero, koliko bližje so si elementi znotraj ene gruče glede na razdaljo do naslednje najbližje gruče. Da se pokazati, da je tako definirana metrika vedno na intervalu $[-1, 1]$, kjer višja vrednost pomeni, da so gruče bolj jasno definirane, torej z manj prekrivanja. Primer uporabe metode "silhuete" je prikazan na Sliki 3.2, kjer največjo vrednost metrike zabeležimo pri $K = 4$. Slike lahko vidimo, da nas obe metodi pripeljeta do istega zaključka glede izbire parametra K .



Slika 3.2: Metoda "komolca" in metoda "silhuete" za primer gručenja seta s štirimi gručami. Pri metodi "komolca" spremjamamo spremenjanje vrednosti cenovne funkcije C , pri metodi "silhuete" pa metrike M_{silhueta} glede na različne vrednosti parametra K . Z rdečo je obkrožena vrednost pri $K = 4$, kjer pride do preloma oziroma maksimuma.

Ker algoritem ne zagotavlja konvergenco v globalni minimum, je pomembna izbira načina inicializacije središč $\{\mu_k\}$. Najenostavnnejši način je naključna izbira K elementov iz podatkovnega seta, kar pa lahko vodi do počasne konvergenco ali pa tudi napačnih rezultatov. Najbolj uveljavljena je inicializacija **k-means++**, pri kateri prvo središče določimo z žrebom, nato pa izračunamo za vse ostale elemente razdaljo do najbližjega že določenega središča $D(\mathbf{x}_n)$ in nato žrebamo naslednjo središče po porazdelitvi $D(\mathbf{x}_n)^2$. Postopek ponavljamo, dokler ne določimo vseh središč. Takšna

POGLAVJE 3. METODE

inicializacija je v večini primerov bolj smiselna, saj so začetne točke bolj razporejene po faznem prostoru.

Gručenje k-means je najbolj primerno za uporabo, ko so gruče v faznem prostoru enako velike in sferične oblike. Znano je namreč, da metoda ne uspe vedno priti do smiselnih rezultatov, ko so velikosti gruč različne, ko so gruče različno goste, ali eliptičnih oblik [53]. Podobno kot pri ostalih metodah strojnega učenja je lahko velika dimenzionalnost podatkov problematična, saj je fazni prostor znatno večji in obenem bolj redek, saj so točke v povprečju bolj razmaknjene. Nekaj izmed teh problemov naslavljajo nekatere razširitve metode, kot so k-medoids, "mehki" k-means in bisekcijski k-means. Bisekcijski k-means osnovnemu algoritmu doda hierarhičnost. Povzamemo ga lahko v naslednjih korakih:

1. Izberemo K in vse elemente združimo v eno gručo.
2. Za vse gruče izračunamo vztrajnostni moment.
3. Gručo z največjim vztrajnostnim momentom z algoritmom k-means za $K = 2$ razdelimo na dva dela.

Ponavljamo 2. in 3. korak, dokler ne dosežemo K gruč.

Ko prvič izvedemo 2. in 3. korak, imamo seveda na izbiro zgolj eno gručo (ki vsebuje vse elemente), nato pa sistematično na manjše gruče delimo tiste, ki so bolj raztresene po prostoru. Takšen algoritem da bistveno boljše rezultate za večje K , kjer ima običajni k-means težave, saj pogosto ustvarja neenakomerno velike gruče.

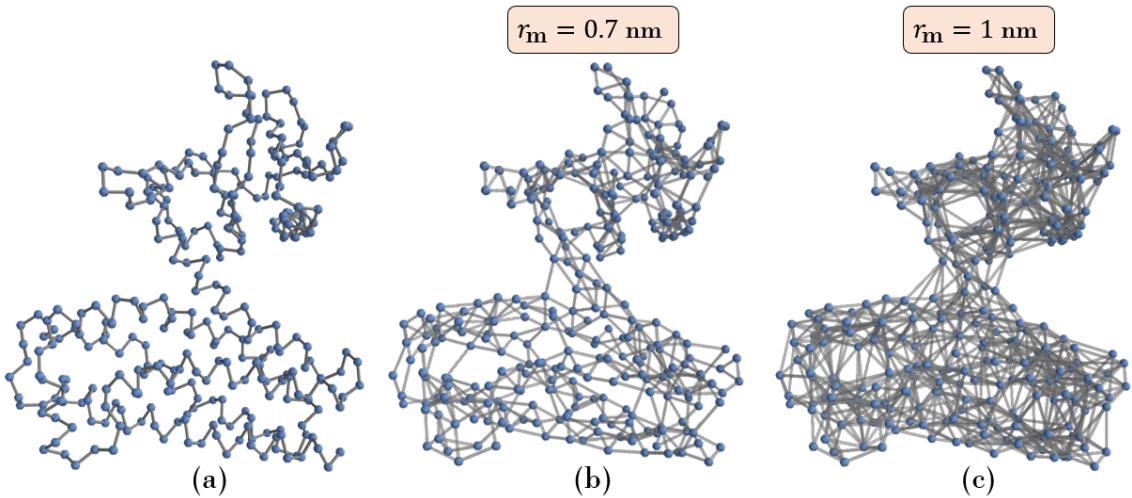
3.2 Elastični mrežni model

Elastični mrežni model je fizikalni model, ki opiše strukturo kompleksnih makromolekul kot set točkastih teles, ki so med seboj povezana z vzemimi [54]. Čeprav je model enostaven, saj predpostavi, da so vse interakcije med atomi kar harmonske, se izkaže, da je z njim mogoče reproducirati osnovne načine gibanja makromolekul [55]. V prvih verzijah modela je bil vsak atom makromolekule modeliran s svojim točkastim telesom, kasneje pa se je izkazalo, da do podobnih načinov gibanja pridemo že z upoštevanjem zgolj izbrane skupine atomov [56]. Rezultati takšnih modelov se dobro ujemajo tako z eksperimenti kot z rezultati simulacij molekularne dinamike [57].

Znotraj elastičnega mrežnega modela atome makromolekule obravnavamo kot točkaste delce, vse interakcije med atomi pa opišemo z linearimi vzemimi. Poljubno strukturo makromolekul tako predstavimo kot mrežo z vzemimi povezanih točkastih teles. Ključen (in v prvih oblikah modela tudi edini) parameter je mejna razdalja r_m , ki določi, kateri gradniki modela so med seboj povezani. Model poveže z vzemimi tiste gradnike, ki so med seboj oddaljeni za največ r_m . Med gradniki, ki so daleč narazen, namreč ni kovalentnih oziroma van der Waalsovih vezi, saj so le-te krajšega dosega.

Elastični mrežni model je najpogosteje uporabljen na primeru proteinov, pri katerem za gradnike modela izberemo vse ogljik- α atome. Referenčne strukture, iz katerih pridobimo informacije o lokacijah atomov in njihovih medsebojnih razdaljah, so pogosto javno dostopne v različnih spletnih zbirkah [58]. Mejna razdalja v takšnih modelih ustreza polmeru koordinacijske sfere, t.j. sfere, ki zaobjema vse atome, ki so direktno vezani na ogljik- α atom. Tipične vrednosti mejne razdalje so okoli 0.7-1.5

nm [59, 60]. Primera elastičnega mrežnega modela sta prikazana na Sliki 3.3, kjer je r_m enak 0.7 oziroma 1 nm. Poleg mejne razdalje elastični mrežni modeli definirajo tudi konstanto vzmeti, za katero je tipična vrednost okoli 10^{-2} kcal/mol(nm) 2 .



Slika 3.3: Prikaz ogljik- α atomov proteina (a). Sosednji ogljik- α atomi so med seboj povezani s temno sivo črto, ki prikazuje t. i. hrbtenico proteina. Na slikah (b) in (c) sta prikazana dva elastična mrežna modela z mejno razdaljo 7 in 10 nm. Zaradi preglednosti je prikazana samo polovica vseh povezav (vzmeti), ki so obarvane s svetlo sivo barvo.

Potencialno energijo elastičnega mrežnega modela lahko zapišemo kot vsoto prispevkov potencialnih energij vzmeti

$$V = \sum_{i,j>i} k_{ij}(r_{ij} - r_{ij}^0)^2, \quad (3.5)$$

kjer vsota teče po vseh parih ogljik- α atomov; dolžino raztegnjene (ozioroma skrčene) vzmeti med i -tim in j -tim ogljik- α smo označili z r_{ij} , njeno ravnovesno dolžino pa z r_{ij}^0 . Ravnovesna dolžina ustreza razdalji med atomoma iz referenčne strukture, za katero se predpostavi, da ustreza energijskemu minimumu. Funkcija k_{ij} definira povezave med gradniki modela in je odvisna od parametra r_m . V najpreprostejši obliki gre za Heavisidovo funkcijo

$$k_{ij} = \begin{cases} c & \text{če } r_{ij} \leq r_m \\ 0 & \text{če } r_{ij} > r_m \end{cases}, \quad (3.6)$$

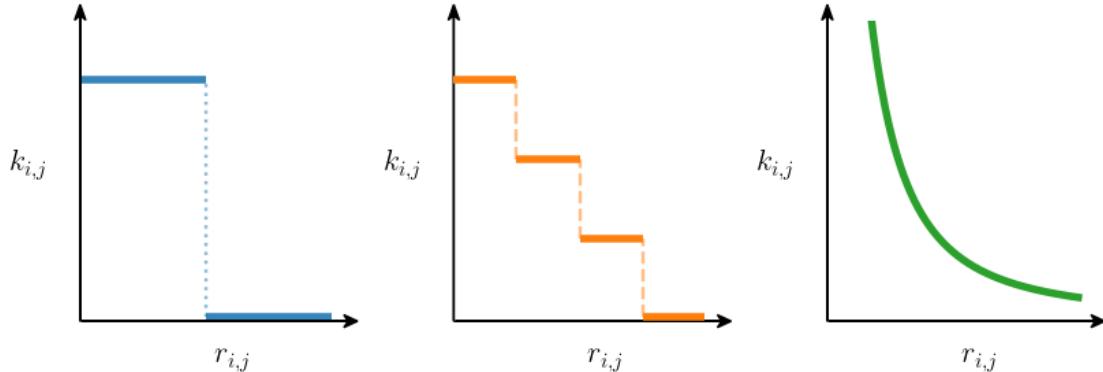
ki je parametrizirana s konstanto vzmeti c . Pogosto je c normiran na 1, saj za analizo osnovnih načinov gibanja makromolekul ni pomembno, kakšen c izberemo. Če je $k_{ij} = 0$, to pomeni, da med i -tim in j -tim ogljik- α atomom ni vzmeti. Razširitve elastičnega mrežnega modela se pogosto poslužujejo drugačnih definicij funkcije k_{ij} , saj enačba (3.6) ne upošteva, da so kemijske vezi med ogljik- α atomi različnih jakosti. Pogosto sta uporabljeni definiciji

$$k_{ij} = \begin{cases} c & \text{če } r_{ij} < r_{m1} \\ c \cdot 10^{-2} & \text{če } r_{m1} < r_{ij} < r_{m2} \\ c \cdot 10^{-4} & \text{če } r_{m2} < r_{ij} < r_{m3} \\ 0 & \text{če } r_{m3} < r_{ij} \end{cases} \quad (3.7)$$

ali

$$k_{ij} = a r_{ij}^p. \quad (3.8)$$

V prvem primeru so definirane tri mejne razdalje r_{m1}, r_{m2}, r_{m3} , konstante vzmeti pa so zaradi enostavnosti modela vse parametrizirane s parametrom c . Skalirni faktorji odražajo različne jakosti kemijskih vezi med ogljik- α atomi. V tem primeru smo predpostavili tri različne jakosti kemijskih vezi med ogljik- α atomi. V enačbi (3.8) pa je k_{ij} definirana kot potenčna funkcija, parametrizirana z dvema parametroma. Model s takšno definicijo poveže med seboj vse gradnike. Tri različne definicije funkcije k_{ij} so prikazana na Sliki 3.4.



Slika 3.4: Tri različne definicije funkcije k_{ij} . Od leve proti desni si sledijo enostopenjska (3.6), tri-stopenjska (3.7) in potenčna (3.8). Enote na vseh treh grafih so arbitrarne.

Elastičen mrežni model lahko analiziramo na dva načina — kot Gaussovski mrežni model ali kot anizotropni mrežni model. Gaussovski mrežni model vsebuje privzetek, da lahko fluktuacije gradnika mrežnega modela okoli ravnovesne lege opišemo z verjetnostno porazdelitvijo, ki je izotropna in “gaussovska” [61], zaradi česar lahko vsak atom opišemo z zgolj enim parametrom, z velikostjo fluktuacij. Anizotropni mrežni model, ki ne vsebuje teh dveh privzetkov, pa je posplošitev Gaussovskega, kjer je vsak gradnik mrežnega modela opisan z vektorjem fluktuacij, namesto zgolj z velikostjo fluktuacij. Konformacijski prostor mrežnega modela ima v tem primeru $3N$ dimenzij, pri čemer je N število gradnikov mrežnega modela. Oba načina sta uporabljeni za izračun fluktuacij gradnikov okoli ravnovesne lege, ki jih izračunamo z analizo lastnih nihajnih načinov.

3.3 Analiza lastnih nihajnih načinov

Analiza lastnih nihajnih načinov je analitično orodje za določanje lastnih nihajnih načinov makromolekul, na primer proteinov [54]. Največja prednost metode je v tem, da se izračunani dolgovalovni lastni nihajni načini dobro ujemajo z osnovnimi načini gibanja makromolekule [62]. Tako lahko preko analitične metode pridobimo informacije o dinamičnih lastnostih makromolekule. Analiza lastnih nihajnih načinov je bila s pridom uporabljena v kontekstu študij prehajanja proteinov med različnimi konformacijami, t. i. odprtimi in zaprtimi formacijami [63]. Prav tako je bila uspešno uporabljena za primerjavo proteinov znotraj istih družin ter za iskanje podobnosti in razlik v dinamičnih lastnostih različnih proteinov [64], kot tudi za iskanje funkcionalnih domen proteinov [65].

Analiza normalnih načinov je pogosto izvedena na elastičnih mrežnih modelih, obstajajo pa tudi druge implementacije, kjer analiza ni izvedena na poenostavljeni strukturi, pač pa na strukturi vseh atomov makromolekule [66]. V primeru, da analiziramo predhodno skonstruiran elastičen mrežni model, lastne načine izračunamo preko Hessejeve matrike

$$H = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1N} \\ h_{21} & h_{22} & \dots & h_{2N} \\ \vdots & & & \vdots \\ h_{N1} & h_{N2} & \dots & h_{NN} \end{bmatrix}, \quad (3.9)$$

kjer so h_{ij} enaki

$$h_{ij} = \begin{bmatrix} \partial^2 V / \partial r_{i_1} \partial r_{j_1} & \partial^2 V / \partial r_{i_1} \partial r_{j_2} & \partial^2 V / \partial r_{i_1} \partial r_{j_3} \\ \partial^2 V / \partial r_{i_2} \partial r_{j_1} & \partial^2 V / \partial r_{i_2} \partial r_{j_2} & \partial^2 V / \partial r_{i_2} \partial r_{j_3} \\ \partial^2 V / \partial r_{i_3} \partial r_{j_1} & \partial^2 V / \partial r_{i_3} \partial r_{j_2} & \partial^2 V / \partial r_{i_3} \partial r_{j_3} \end{bmatrix}. \quad (3.10)$$

Hessejeva matrika je torej simetrična $3N \times 3N$ matrika drugih odvodov potenciala V . Superelementi h_{ij} so 3×3 matrike, ki vsebujejo druge odvode potenciala V vzdolž koordinat r_{i_x} , kjer je \mathbf{r}_i lokacija i -tega ogljik- α atoma, indeks x pa označuje prostorsko koordinato (npr. $x = 1, 2, 3$ predstavljajo prvo, drugo in tretjo prostorsko dimenzijo). Element $\partial^2 V / \partial r_{i_1} \partial r_{j_1}$ se tako nanaša na odvod potenciala vzdolž x koordinate i -tega in j -tega ogljik- α atoma. Za analizo lastnih načinov elastičnega mrežnega modela je potencial sistema enak temu v Enačbi (3.5). Izkaže se [67], da lahko druge odvode, ki nastopajo v matrikah h_{ij} izrazimo kot

$$\frac{\partial^2 V}{\partial r_{i_x} \partial r_{j_y}} = -k_{ij} \frac{(r_{j_x} - r_{i_x})(r_{j_y} - r_{i_y})}{r_{ij}^2} \Bigg|_{r_{ij}=r_{ij}^0}, \quad (3.11)$$

pri čemer pa so elementi matrik h_{ii} enaki

$$\frac{\partial^2 V}{\partial r_{i_x} \partial r_{i_y}} = \sum_{j \neq i} k_{ij} \frac{(r_{j_x} - r_{i_x})(r_{j_y} - r_{i_y})}{r_{ij}^2} \Bigg|_{r_{ij}=r_{ij}^0}, \quad (3.12)$$

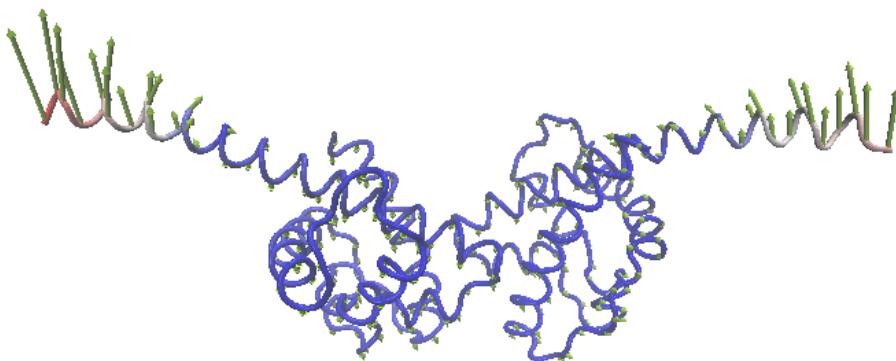
POGLAVJE 3. METODE

kjer je razdalja med i -tim in j -tim atomom r_{ij} izvrednotena pri ravnovesni razdalji r_{ij}^0 . Za ravnovesno razdaljo se enako kot v elastičnem mrežnem modelu vzame kar razdaljo iz referenčne strukture, za katero se predpostavi, da je v energijskem minimumu. Na desni strani Enačb (3.11) in (3.12) lahko prepoznamo funkcijo k_{ij} , ki je v najpreprostejši obliki (Enačba (3.6)) enaka bodisi 0 bodisi konstanti $vzmeti c$. Na isti strani enačbe nastopa tudi ulomek, ki se izvrednoti v brez-dimenzijski številski faktor. Desna stran obeh enačb je zelo sorodna frekvenci nihanja, kar lahko uvidimo, če se spomnimo izraza za frekvenco harmonskega oscilatorja — $\omega^2 = k/m$. Če bi torej celo matriko H delili z $-1/m$, bi vsak element v h_{ij} predstavljal kvadrat frekvence nihanja. Ker imajo vsi ogljik- α atomi enako maso, lahko ta korak izpustimo, saj bi s tem zgolj skalirali celotno matriko za številski faktor, kar ne bi vplivalo na nadaljnje korake v analizi.

Hessejevo matriko H lahko diagonaliziramo in s tem poiščemo lastni sistem matrike. Problem iskanja lastnih vrednosti lahko zapišemo kot

$$H = U\Omega U^{-1}, \quad (3.13)$$

kjer je U matrika lastnih vektorjev in Ω diagonalna matrika lastnih vrednosti. Z dekompozicijo Hessejeve matrike poiščemo lastni prostor sistema, kjer vsaka lastna vrednost predstavlja frekvenco lastnega načina, vsak lastni vektor pa ustrezne premike ogljik- α atomov, povezane z lastnim načinom. Izkaže se, da je prvih šest lastnih vrednosti enakih nič, saj so ti načini povezani s tremi translacijami in tremi rotacijami v prostoru, pri katerih ni nihanja. Lastni vektorji, ki jih dobimo z diagonalizacijo, so velikosti $3N$ in vsebujejo N tridimenzionalnih vektorjev, po enega za vsak gradnik elastičnega mrežnega modela. Vsak izmed N -tih vektorjev definira smer v prostoru, vzdolž katere bo gradnik zanihal. Norma vektorja pa skupaj s frekvenco lastnega načina določi amplitudo nihaja. Lastne načine makromolekul lahko vizualiziramo s programi, kot je VMD [68], pri katerem na spektru od modre do rdeče prikažemo amplitudo oscilacij, kot je prikazano na Sliki 3.5.



Slika 3.5: Vizualizacija prvega lastnega načina proteina (PDB: 8IZU [69]). Na mestih ogljik- α atomov so s puščicami prikazane smeri in amplitudo nihanja posameznega ogljik- α atoma. Lastni vektorji so interpolirani preko celotne hrbtenice proteina, ki je obarvana glede na amplitudo nihanja posameznega segmenta.

Pri interpretaciji rezultatov se je potrebno zavedati, da analiza lastnih načinov sloni na dveh pomembnih predpostavkah — da je referenčna struktura, iz katere analiza izhaja, v energijsko stabilnem stanju, in da je gibanje gradnikov makromolekule v okolini tega minimuma približno linearne. Ker smo težko prepričani,

da je referenčna struktura zares v energijskem minimumu, je potrebna previdnost pri interpretacijah visoko-frekvenčnih lastnih načinov, saj le-ti opisujejo lokalizirana gibanja makromolekule in imajo variacije pozicij atomov v referenčni strukturi na njih večji vpliv. Zaradi predpostavke linearnosti pa se je pomembno zavedati, da so nihajni načini točni le za majhne izmike iz ravnovesne lege.

3.4 Metoda bistvene dinamike

Metoda bistvene dinamike (ang. *Essential dynamics, ED*) omogoča konstrukcijo GZ modela, pri katerem ohranimo bistvene dinamične lastnosti proteina [67]. Kot je pogosto pri GZ modelih, je glavni cilj metode poiskati najpreprostejšo molekularno reprezentacijo, ki še odraža nekatere fizikalne lastnosti molekule. V primeru metode bistvene dinamike so to dinamične lastnosti modeliranega proteina (kako se protein upogiba, katera področja so bolj prožna od drugih ipd.). Dinamične lastnosti, ki jih metoda bistvene dinamike ohrani, so bile v prvi obliki metode pridobljene z analizo poglavitnih komponent (ang. *Principal components analysis*) [70]. V tem primeru najprej opravimo simulacijo molekularne dinamike in na rezultatih s pomočjo analize poglavitnih komponent iščemo osnovne načine gibanja takšne makromolekule. Kasneje je bila razvita še verzija metode, ki definira dinamične lastnosti glede na elastičen mrežni model in analizo lastnih nihajnih načinov [67]. V obeh primerih je osnoven princip konstrukcije GZ modela enak — iščemo domene znotraj proteina, ki se, glede na prepozname dinamične lastnosti, gibljejo na enak način. Obe inačici metode dajeta dobre rezultate, vendar je ta, ki je osnovana na elastičnemu mrežnemu modelu bistveno manj računsko zahtevna, saj ne potrebujejo časovno potratnih simulacij molekularne dinamike. Iz tega razloga bomo v kontekstu te naloge obravnavali le obliko metode bistvene dinamike, ki sloni na elastičnemu mrežnemu modelu.

Metoda bistvene dinamike je bila uporabljena za študij več-domenskih proteinov, proteinskih kompleksov in molekularnih motorjev [71], kjer je bilo potrjeno, da se tako pridobljene dinamične domene dobro ujemajo s funkcionalnimi domenami, t. j. predeli, ki skupaj opravljam neko funkcijo proteina.

V metodi bistvene dinamike so dinamične lastnosti definirane preko lastnih nihajnih načinov proteina. V tem smislu prvih nekaj lastnih načinov predstavlja načine gibanja proteina, za katere želimo, da jih lahko poustvarimo tudi z GZ modelom. Ti najnižji lastni nihajni načini definirajo t. i. podprostor bistvenih načinov gibanja proteina. Podprostor je v tem kontekstu mišljen kot del prostora vseh načinov gibanja. Metodo bistvene dinamike lahko strnemo v nekaj korakih:

1. Izberemo število GZ gradnikov M .
2. Naključno izberemo $M - 1$ izmed vseh ogljik- α atomov proteina in jih postavimo za mejne ogljik- α atome, saj definirajo meje med dinamičnimi domenami proteina (glej Sliko 3.6).
3. Izvrednotimo residual χ^2 , ki je definiran z

$$\chi^2 = \frac{1}{3M} \sum_{S=1}^M \sum_{i \in S} \sum_{j \geq i \in S} \langle (\Delta \mathbf{r}_i^{\text{ED}})^2 - 2\Delta \mathbf{r}_i^{\text{ED}} \mathbf{r}_j^{\text{ED}} + (\mathbf{r}_j^{\text{ED}})^2 \rangle, \quad (3.14)$$

POGLAVJE 3. METODE

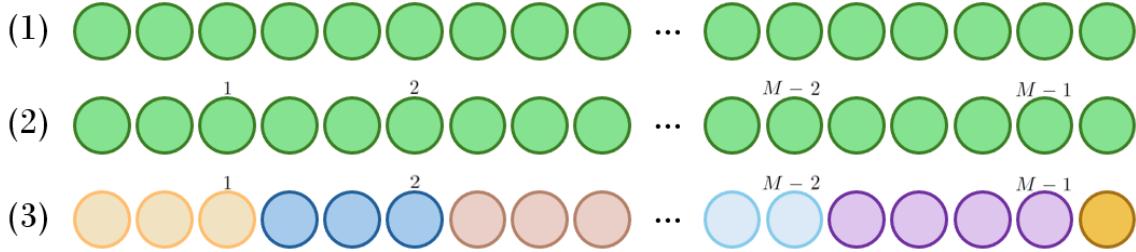
kjer je

$$\langle (\Delta \mathbf{r}_i^{\text{ED}})^2 \rangle = k_B T \sum_{x=1}^3 \sum_{q=7}^{3M} \mathbf{u}_q^{i_x} \omega_q^{-1} \mathbf{u}_q^{i_x}. \quad (3.15)$$

4. Določimo nove mejne ogljik- α atome.

(Koraka 3 in 4 ponavljamo, dokler ne dosežemo minimuma residuala χ^2 .)

5. Gradnike GZ modela postavimo v masna središča dinamičnih domen.



Slika 3.6: Shema razdelitve proteina na domene, ki je uporabljena v metodi bistvene dinamike. (1) Protein, predstavljen kot veriga ogljik- α atomov (vsaka zelena kroglica ustreza enemu atomu). (2) Naključen izbor $M - 1$ mejnih ogljik- α atomov. (3) Definiranje M domen glede na mejne ogljik- α atome.

Residual χ^2 je minimalen, ko uspemo protein razdeliti na domene, znotraj katerih se ogljik- α atomi gibljejo na karseda podoben način. Iščemo torej razdelitev proteina na domene, ki bodo združile skupaj atome, ki se znotraj podprostora bistvenih načinov gibanja premikajo korelirano — nihajo v isto smer. To nam razkriva tudi podrobni pregled Enačbe (3.14): prva vsota teče po vsaki dinamični domeni (označeni s S) posebej, druga in tretja vsota pa po vseh parih atomov i in j , ki pripadajo domeni S. Za vsak tak par izračunamo povprečje kvadrata fluktuacij atoma i ($\Delta \mathbf{r}_i^{\text{ED}}$)² in atoma j ($\Delta \mathbf{r}_j^{\text{ED}}$)² znotraj podprostora bistvenih načinov gibanja nihala povsem enako, bi veljalo $\Delta \mathbf{r}_i^{\text{ED}} = \Delta \mathbf{r}_j^{\text{ED}}$ in bi bil tako $2\Delta \mathbf{r}_i^{\text{ED}} \mathbf{r}_j^{\text{ED}}$ enak ($\Delta \mathbf{r}_i^{\text{ED}}$)², zaradi česar bi bil ta prispevek k residualu χ^2 enak 0. Predfaktor $1/(3M)$ normira celoten residual glede na izbrano število dinamičnih domen oziroma GZ gradnikov in ustreza številu prostostnih stopenj GZ modela z M gradniki.

Povprečje kvadrata fluktuacij je za posamezen atom izračunano glede na Enačbo (3.15). Izraz izhaja iz enakosti s področja teorije grafov [72]

$$\langle (\Delta \mathbf{r}_i^{\text{ED}})^2 \rangle = k_B T \text{tr}[\mathbf{h}_{ii}^{-1}], \quad (3.16)$$

kjer je s T označena temperatura sistema, k_B je Boltzmannova konstanta, $\text{tr}[\mathbf{h}_{ii}^{-1}]$ pa označuje sled inverza matrike \mathbf{h}_{ii} , ki jo lahko izrazimo kot

$$\text{tr}[\mathbf{h}_{ii}^{-1}] = \sum_{x=1}^3 \sum_{q=7}^{3M} \mathbf{u}_q^{i_x} \omega_q^{-1} \mathbf{u}_q^{i_x}, \quad (3.17)$$

kjer smo z \mathbf{u}_q označili q -ti najmanjši lastni vektor, z ω_q pa pripadajočo lastno vrednost. Če združimo obe zvezi, dobimo točno Enačbo (3.15). Ker lahko lastne vektorje zapišemo kot

$$\mathbf{u}_q = \begin{bmatrix} \mathbf{u}_q^1 \\ \mathbf{u}_q^2 \\ \vdots \\ \mathbf{u}_q^N \end{bmatrix} = \left[\left(u_q^{1,1} u_q^{1,2} u_q^{1,3} \right) \left(u_q^{2,1} u_q^{2,2} u_q^{2,3} \right) \dots \left(u_q^{N,1} u_q^{N,2} u_q^{N,3} \right) \right]^T, \quad (3.18)$$

u_q^{ix} ustreza torej x prostorski koordinati i -tega ogljik- α atoma glede na q -ti lastni način. Druga vsota v Enačbi (3.15) se začne pri $q = 7$, saj je prvih 6 lastnih vrednosti ω_q vedno enakih 0, saj ti načini ustrezano prosti translaciji in rotaciji v prostoru. Pogosto je velikost podprostora bistvenih nihajnih načinov, ki nastopa kot zgornja meja vsote po q v Enačbi 3.15, enaka številu prostostnih stopenj GZ modela z M atomi, torej $3M$.

Koraka 3 in 4 sta izvedena z uporabo numeričnih metod za iskanje globalnega minimuma funkcije. V kombinaciji sta uporabljena simulirano ohlajanje (ang. *simulated annealing*) in gradientni spust (ang. *gradient descent*). Implementirana sta na način, da na vsakem koraku minimizacijskega algoritma naključno izberemo eno izmed mej med domenami in jo zamaknemo za premik, izžreban iz neke verjetnostne porazdelitve, ki se nam zdi smiselna. Novo razdelitev na domene sprejmemo glede na spremembo vrednosti residuala $\Delta\chi^2$ in Metropolisov kriterij:

- sprejmemo, če
 - $\Delta\chi^2 < 0$ ali
 - $r \leq \exp(-\Delta\chi^2/T)$, kjer je r naključno število z intervala $[0, 1]$,
- sicer zavrnemo.

Na vsakem koraku postopoma znižujemo temperaturo sistema T , s čimer se manjša verjetnost za sprejem razdelitev, pri katerih se ni zmanjšal residual. Na ta način dopustimo, da se vmes vrednost residuala tudi poveča, za ceno tega, da se izognemo ujetosti v lokalne minime. Simulirano ohlajanje zaključimo, ko T pade na 0. Lahko si obetamo, da bo takrat sistem v bližini minima, za katerega slutimo, da je globalen. Po simuliranem ohlajanju opravimo še gradientni spust, kjer iz okolice minima preidemo v dejanski minimum residuala. To storimo na način, da za vsako mejo med domenami posebej pogledamo, kako se spremeni residual, če jo zamaknemo za eno mesto v levo ali desno in premik sprejmemo le za zamike, ki so zmanjšali residual. Na tak način minimiziramo residual glede na lokacijo vsake meje med domenami posebej. Ker kombinacija uporabe simuliranega ohlajanja in gradientnega spusta ne zagotavlja, da bomo našli globalni minimum residuala, opravimo več minimizacij z različnimi začetnimi razmejitvami med domenami, na koncu pa izberemo tisto z najmanjšo vrednostjo residuala.

Eden izmed večjih izzivov pri uporabi metode bistvene dinamike je odločitev, na koliko dinamičnih domen bomo razdelili protein. Eden izmed načinov, ki se ga lahko poslužimo, je, da izračunamo, koliko lastnih načinov potrebujemo zaobjeti v podprostor bistvenih načinov gibanja, da z njimi opišemo npr. 90% vseh fluktuacij, nato število domen določimo tako, da se število prostostnih stopenj GZ modela

POGLAVJE 3. METODE

ujema z velikostjo podprostora. Na ta način sicer fizikalno opravičimo število izbranih domen, a problem prestavimo na vprašanje, kolikšen procent fluktuacij je smiselno upoštevati, kar v splošnem tudi ni očitno.

Druga večja omejitev metode pa je v načinu iskanja minimuma residuala, ki predpostavlja, da so domene neprekinjene glede na zaporedje ogljik- α atomov, kar ni nujno fizikalno smiselno. Kot odgovor na to problematiko sta bila razvita dva alternativna načina določevanja mej domen, prostorski [73] in bisekcijski [74]. Pri prostorskem algoritmu dopustimo združevanje poljubnih ogljik- α atomov v GZ gradnike, združujemo pa jih na podlagi njihove prostorske oddaljenosti, sam algoritem pa je zelo podoben kot pri metodi k-means. Bisekcijski algoritem pa je pravzaprav soroden bisekcijskemu k-means algoritmu, pri katerem celoten protein razdelimo na dve domeni, nato pa postopoma delimo domene na manjše enote na način, da na dva dela razdelimo tisto domeno, pri kateri smo z minimizacijskim algoritmom našli razdelitev z najnižjo vrednostjo residuala. Izkaže se, da je mogoče s prostorskim modelom v nekaterih primerih najti boljše GZ modele, a je postopek iskanja vedno bistveno daljši v primerjavi z osnovnim modelom. Z bisekcijskim modelom navadno pridelamo podobne rešitve kot z osnovnim, pri čemer je minimizacija hitrejša, zaradi česar je posebno uporaben za modeliranje velikih proteinov.

4. Rezultati

V naslednjem poglavju so zbrani vsi rezultati tega magistrskega dela. Najprej so predstavljeni rezultati, ki se navezujejo na GZ modelle. Prikazani so GZ modeli, ki temeljijo na metodi k-means in modeli, ki jih generira metoda bistvene dinamike. Drugi del je posvečen iskanju korelacij med dinamičnimi lastnostmi, ki so povezane z lastnimi nihajnimi načini in hitrostjo agregacije. V prvem podpoglavlju tega sklopa hitrost agregacije koreliramo s frekvencami lastnih nihajnih načinov, v drugem pa z amplitudo nihanja v variabilni regiji monoklonskih protiteles. Vsi rezultati izhajajo iz podatkovnega seta 21 monoklonskih protiteles [18].

4.1 Grobo-zrnati modeli

GZ modeli monoklonskih protiteles ponavadi vsebujejo najmanj 3 oziroma največ 26 psevdo-atomov, medtem ko je najbolj pogosto uporabljen reprezentacija z 12 gradniki. V večini primerov je združevanje atomov v psevdo-atome pogojeno z informacijami o mejah med funkcionalnimi domenami monoklonskega protitelesa. V nadaljevanju sta predstavljena dva alternativna pristopa konstrukcije GZ modela, ki teh informacij ne potrebujejo in bi bila lahko posplošena tudi za uporabo na drugih proteinih. Metoda strojnega učenja k-means išče predele molekule, kjer so atomi gosto skupaj in jih združi v psevdo-atom. V tem oziru gre za GZ model, ki temelji na gostoti porazdelitve atomov po prostoru. Poleg metode k-means smo preizkusili tudi nekatere druge metode gručenja s področja strojnega učenja, kot so Gaussovske mešanice, hierarhično gručenje in DBSCAN [75], a se je metoda k-means zaradi enostavnosti uporabe in obetavnih rezultatov izkazala za najbolj primerno.

Drugi način konstrukcije GZ modela, ki je predstavljen v nadaljevanju, je metoda bistvene dinamike, pri kateri so atomi združeni v psevdo-atome tako, da se ohranijo dinamične lastnosti molekule.

Ker je namen te naloge preizkusiti ali sta predlagana načina konstrukcije GZ modela smiselna, je prvi cilj reprodukcija rezultatov, ki se pojavljajo drugod v literaturi in so prikazani tudi na Sliki 2.5. Za možnosti kvalitativne primerjave so zato podani modeli s 3, 4, 6 in 12 psevdo-atomi. Ker podrobni podatki GZ modelov, kot sta referenčna struktura in meje med psevdo-atomimi, pogosto niso javnega značaja, kvantitativne primerjave niso mogoče. V vseh primerih smo GZ modelirali prvo monoklonsko protitelo iz podatkovnega seta [18], ki ustreza molekuli tocilizumab. Referenčna struktura je relaksirana.

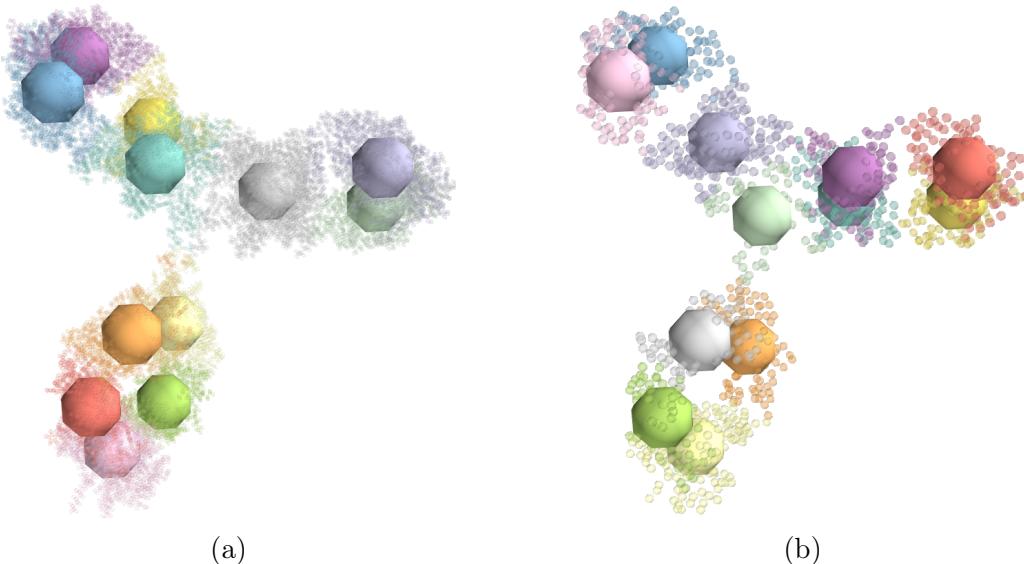
4.1.1 GZ modeliranje z metodo k-means

Uporabljena je python implemetacija gručenja k-means, ki je dostopna znotraj knjižnice `scikit-learn` [76]. Poleg števila gruč (psevdo-atomov) določimo tudi inicializacijski algoritem (`k-means++`), maksimalno število iteracij (100) in število inicializacij (100), s čimer povečamo verjetnost, da se ob koncu algoritma res nahaj-

POGLAVJE 4. REZULTATI

jamo v globalnem minimum.

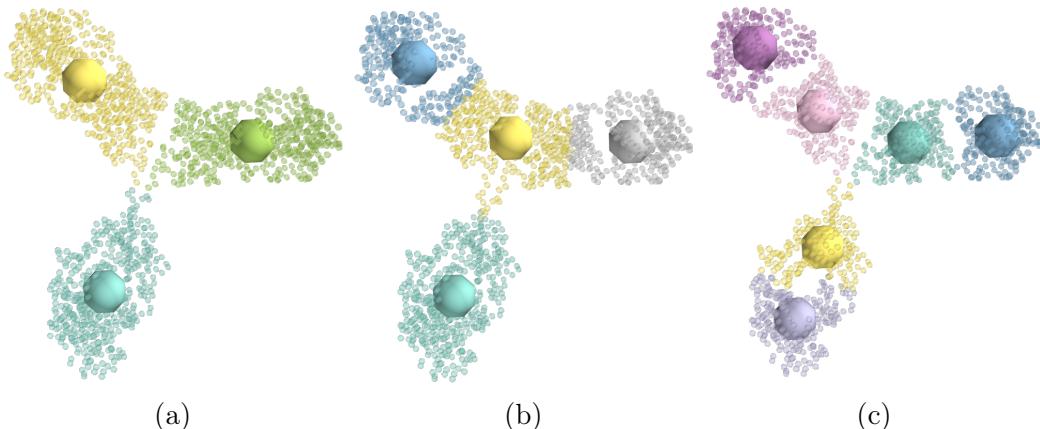
Algoritem gručenja smo sprva izvedli na vseh atomih molekule. Primer z 12 psevdo-atomi je prikazan na Sliki 4.1a, kjer so atomi obarvani glede na to, kateremu psevdo-atomu pripadajo. Psevdo-atomi so prikazani kot večje krogle. V reprezentaciji z vsemi atomi meje med domenami niso več tako jasno definirane, zaradi česar jih algoritem k-means ne prepozna povsem. Levi krak molekule je razdeljen na 4 psevdo-atome, kot v referenčnih modelih iz literature, medtem ko sta desni in spodnji krak razdeljena na 3 oziroma 5 psevdo-atomov. Ker vemo, da je molekula skoraj povsem simetrična glede na levo in desno polovico, lahko sklepamo, da rešitev ni optimalna. Nekoliko boljše rezultate dobimo, če gručimo le ogljik- α atome, kot je za 12 psevdo-atomov prikazano na Sliki 4.1b. Domene so v tem primeru bolj jasno ločene, kar se odraža tudi v GZ modelu, ki je sedaj bolj podoben referenčnemu — vsak psevdo-atom predstavlja eno domeno, z izjemo konstantnih domen v levem kraku, kjer sta obe opisani z enim psevdo-atomom, zaradi česar je ena gruča “izrinjena” v tečajni predel molekule. Prikazan GZ model je kot prejšnji asimetričen glede na levi in desni krak.



Slika 4.1: Gručenje k-means, izvedeno na vseh atomih monoklonskega protitelesa (a) in na ogljik- α atomih monoklonskega protitelesa (b). Atomi so obarvani glede na to, v kater psevdo-atom so združeni. Velikosti psevdo-atomov so določene arbitralno.

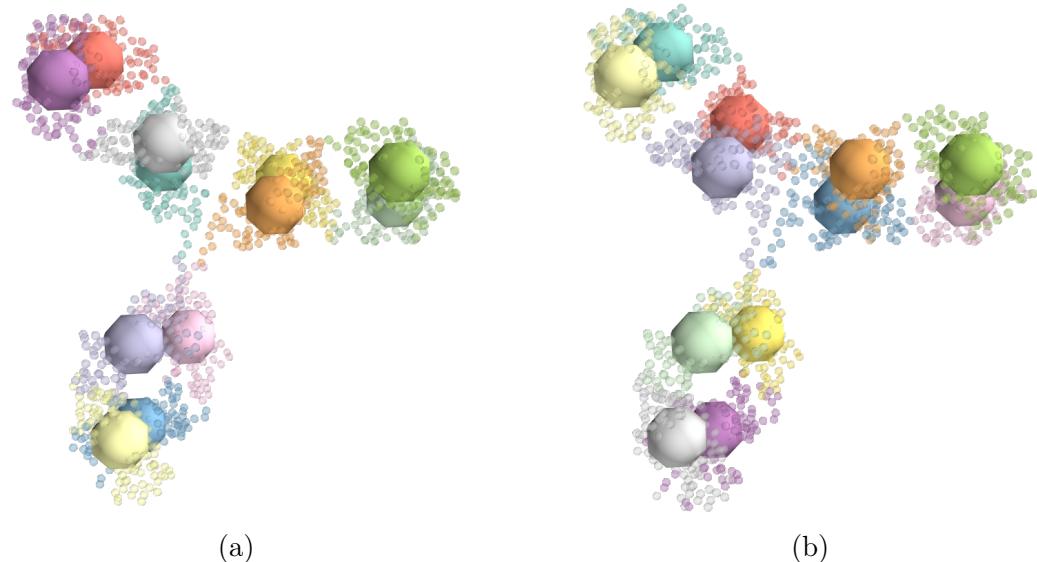
Reproducirati poskusimo tudi GZ modele s 3, 4 ali 6 psevdo-atomimi. Gručenja na ogljik- α atomih so prikazana na Sliki 4.2. Modela s 3 in 6 psevdo-atomimi se kvalitativno povsem ujemata z referenčnimi GZ modeli, model s 4 psevdo-atomimi pa je precej drugačen, saj četrти atom ni postavljen v tečajni predel, pač pa zaobjema večji del konstantnih domen levega in desnega kraka.

Da bi zagotovili simetrijo levega in desnega kraka v GZ modelu z 12 psevdo-atomimi, lahko gručenje ločeno opravimo na vsaki verigi monoklonskega protitelesa posebej. Tako izvedemo 4 ločena gručenja, nato pa jih združimo v celoten GZ model. Na ta način vsak psevdo-atom opiše zgolj atome iz ene verige, kar sicer ni nujno smiselno s stališča funkcionalnih domen. Na takšen način lahko skonstruiramo GZ model monoklonskega protitelesa z 12 psevdo-atomimi, ki se kvalitativno ujema z referenčnim. Prikazan je na Sliki 4.3a. Izkaže se, da lahko kvalitativno podobne



Slika 4.2: Gručenja k-means, izvedena na ogljik- α atomih monoklonskega protitelesa s 3 (a), 4 (b) in 6 (c) gručami.

rezultate dobimo tudi z uporabo bisekcijskega k-means gručenja na ogljik- α atomih, kar je prikazano na Sliki 4.3b. Ker se tako izognemo gručenju vsake verige posebej, je takšen pristop nekoliko bolj splošen, hkrati pa dopušča, da se v isti psevdo-atom združijo atomi iz različnih verig.



Slika 4.3: Gručenje k-means, izvedeno na ogljik- α atomih monoklonskega protitelesa na vsaki verigi posebej (a). Težki verigi sta razdeljeni na 4 gruče, lahki pa na 2. Gručenje bisekcijskega k-means, izvedeno na ogljik- α atomih z 12 gručami (b).

Za nekoliko bolj kvantitativno primerjavo med GZ modeloma s Slike 4.3 lahko definiramo metriko

$$P(\text{domena}) = \frac{1}{N} \sum_{C_\alpha^k \in S'_{\text{domena}}} \psi(C_\alpha^k, S'_{\text{domena}}), \quad (4.1)$$

$$\psi(C_\alpha^k, S'_{\text{domena}}) = \begin{cases} 1, & \text{če } C_\alpha^k \in S'_{\text{domena}}, \\ 0, & \text{če } C_\alpha^k \notin S'_{\text{domena}} \end{cases}$$

POGLAVJE 4. REZULTATI

ki ovrednoti podobnost dveh gruč različnih GZ modelov, ki opisujeta isto domeno monoklonskega protitelesa. Seta S_{domena} in S'_{domena} vsebujeta vse ogljik- α atome prvega oziroma drugega modela, ki se nahajajo v določeni domeni. Metrika prešteje, koliko ogljik- α atomov se nahaja tako v enem kot v drugem setu. Normirana je na število različnih ogljik- α atomov v obeh setih N . V Tabeli 4.1 so za vsako izmed domen monoklonskega protitelesa navedeni pripadajoči ogljik- α atomi za bisekcijski k-means in k-means, izveden na vsaki verigi posebej.

domena	pripadajoči ogljik- α atomi	P
C_{H3}	248 - 258, 312 - 313, 316 - 317, 342 - 354, 366 - 448, 803 - 814, 864, 896 249 - 257, 343 - 448	76.5%
C_{H2}	235 - 247, 259 - 311, 314 - 315, 318 - 341 233 - 248, 258 - 342	91.1%
C_{H1}	124 - 137, 139 - 148, 156 - 171, 186 - 234, 1016 - 1023, 1108 - 1110 121 - 232	69.6%
V_H	1 - 42, 46 - 102, 105, 107 - 119, 939, 990 - 992 1 - 120	91.1%
C_L	120 - 123, 138, 149 - 155, 172 - 185, 1004 - 1015, 1024 - 1107 1005 - 1061, 1063, 1066, 1068 - 1110	68.4%
V_L	43 - 45, 103 - 104, 106, 897 - 938, 940 - 989, 993 - 1003 897 - 1004, 1062, 1064 - 1065, 1067	87.3%
C_{H3}^*	355 - 365, 696 - 706, 760 - 761, 764 - 765, 790 - 802, 815 - 863, 865 - 895 697 - 705, 791 - 896	75.9%
C_{H2}^*	683 - 695, 707 - 759, 762 - 763, 766 - 789 681 - 696, 706 - 790	91.1%
C_{H1}^*	572 - 585, 587 - 596, 604 - 619, 634 - 682, 1230 - 1237, 1322 - 1324 569 - 680	69.6%
V_H^*	449 - 490, 494 - 550, 553, 555 - 567, 1153, 1204 - 1206 449 - 568	91.1%
C_L^*	568 - 571, 586, 597 - 603, 620 - 633, 1218 - 1229, 1238 - 1321 1219 - 1275, 1277, 1280, 1282 - 1324	68.4%
V_L^*	491 - 493, 551 - 552, 554, 1111 - 1152, 1154 - 1203, 1207 - 1217 1111 - 1218, 1276, 1278 - 1279, 1281	87.3%

Tabela 4.1: Primerjava GZ modelov, skonstruiranih z bisekcijskim k-means in s k-means, izvedenim na vsaki verigi posebej. V prvem stolpcu se nahajajo označke domen monoklonskega protitelesa, v drugem pa so v vsaki celici zgoraj napisani pripadajoči ogljik- α atomi glede na bisekcijski k-means GZ model, spodaj pa pripadajoči ogljik- α atomi glede na k-means, izvedenim na vsaki verigi posebej. V zadnjem stolpcu so navedene vrednosti metrike prekrivanja P za posamezno domeno.

Simetrični pari domen so označeni z zvezdico, nomenklatura pa je enaka kot na Sliki 2.3. V zadnjem stolpcu tabele so izračunane metrike prekrivanja obeh modelov za posamezno domeno. Izkaže se, da so prekrivanja domen in njihovih simetričnih parov praktično enaka, kar kaže na simetričnost in hkrati smiselnost obeh modelov. Najboljše se ujemata domeni C_{H2} in V_H , oziroma njuna simetrična para, kjer je prekrivanje 91.1%. Podobno močno prekrivanje opazimo tudi pri domeni V_L , kjer je

$P = 87.3\%$. Najslabši prekrivanji sta med domenami C_{H1} in C_L , ki znašata 69.6% oziroma 68.4%. Edino odstopanje med simetričnima paroma opazimo pri domeni C_{H3} .

Ker za strukturo GZ modela ni pomembno, kateri vsi atomi določajo posamezen psevdo-atom (čeprav tudi to postane pomembno, ko GZ psevdo-atomom pripisemo pospoljene naboje in druge lastnosti), lahko modela primerjamo glede na razdaljo med težiščema posameznih psevdo-atomov. Definiramo lahko relativno razliko med težiščema psevdo-atomov

$$\Delta\mu_{\text{domena}} = \frac{\|\mathbf{R}_{\text{domena}} - \mathbf{R}'_{\text{domena}}\|}{\langle d_{C_\alpha} \rangle}, \quad (4.2)$$

kjer smo z $\mathbf{R}_{\text{domena}}$ in $\mathbf{R}'_{\text{domena}}$ označili težišči psevdo-atomov obeh modelov, ki opisujeta določeno domeno. Razliko težišč normiramo s povprečno razdaljo med zaporednimi ogljik- α atomi $\langle d_{C_\alpha} \rangle$, tako da postane količina brez-dimenzijska.

V Tabeli 4.2 so navedene vrednosti $\Delta\mu_{\text{domena}}$ za posamezno domeno. Zgornja vrednost v posamezni celici se nanaša na pripisano domeno, spodnja pa na njen simetričen par. Podobno kot pri prekrivanju domen se edina znatna razlika pojavi med domenama C_{H3} , in C'_{H3} . Razlike med težišči psevdo-atomov so v vseh primerih relativno majhne, največje odstopanje se pojavi pri C_{H1} , kjer je tudi prekrivanje med psevdo-atomoma najmanjše in znaša 1.181. Domeni C_{H2} in V_H se sicer enako prekrivata ($P = 91.1\%$), a se nekoliko razlikujeta v razdalji med težiščema.

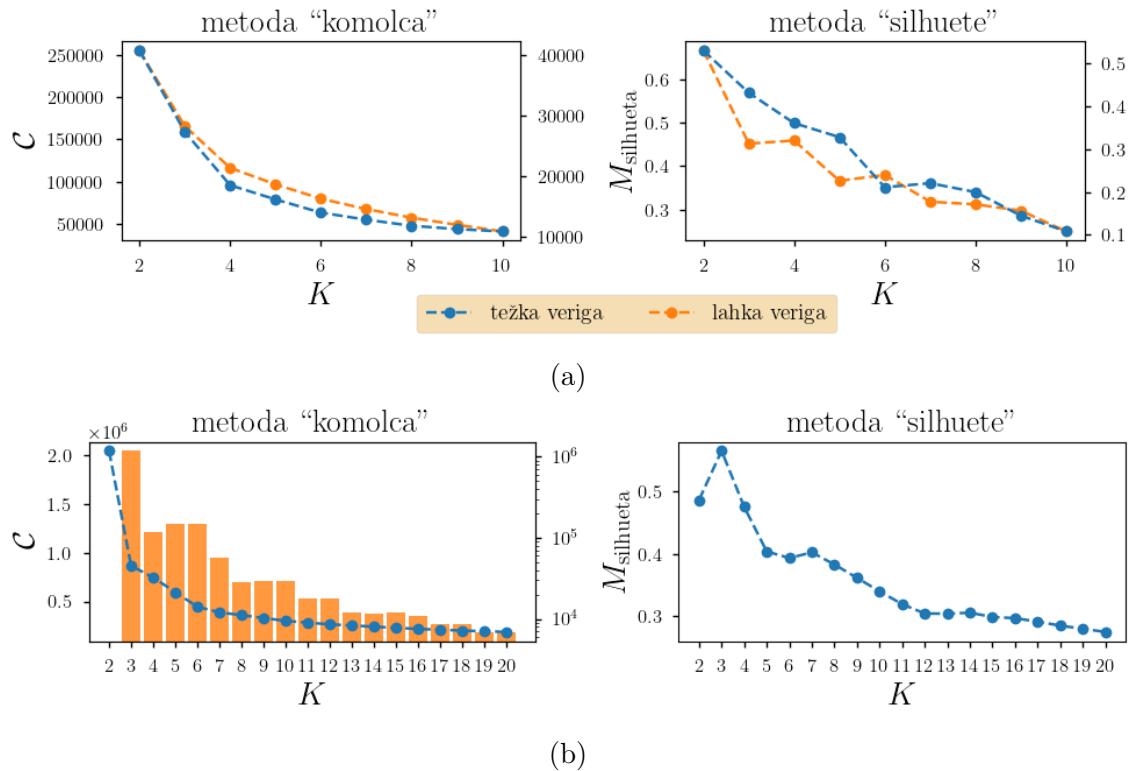
	C_{H3}	C_{H2}	C_{H1}	V_H	C_L	V_L
$\Delta\mu_{\text{domena}}$	0.794	0.23	1.181	0.095	1.072	0.234
	0.914	0.23	1.181	0.095	1.072	0.234

Tabela 4.2: Primerjava strukture GZ modelov preko razlike med težišči psevdo-atomov $\Delta\mu_{\text{domena}}$. V celicah sta združeni vrednosti metrike, ki se navezujeta na domeno (zgoraj) in njen simetričen par (spodaj).

V primeru modeliranja protiteles so število gruč določevali referenčni modeli, ki jih želimo reproducirati, da pa bi preverili, ali se da metodo uporabiti tudi brez referenčnega GZ modela na kakšnem drugem tipu molekul, si lahko ogledamo metodi "komolca" in "silhuete". V kolikor lahko smiselna števila gruč izluščimo tudi s tema metodama, bi lahko k-means metodo uporabili tudi za konstrukcijo GZ modelov za tipe molekul, kjer referenčni model še ni znan. Na Slikah 4.4a in 4.4b so prikazani grafi obeh metod za k-means, izveden na vsaki verigi posebej, in bisekcijski k-means.

Graf metode "silhuete" s Slike 4.4a nakazuje, da je primerno vsako verigo razdeliti na dva dela, s čimer bi dobili 8-delčni model, ki se sicer ne pojavlja v literaturi. Graf metode "komolca" s Slike 4.4a se za težko verigo res prelomi pri $K = 4$. Oba grafa, ki se navezujeta na bisekcijski k-means, najbolj jasno nakazujeta, da je optimalno število gruč enako 3, kar ustrezza najpreprostejšemu referenčnemu GZ modelu. Graf cenovne funkcije pa nakazuje tudi na smiselnost gručenja pri $K = 6, 10$ in 12 , saj se odbitki cenovne funkcije po vsakem izmed navedenih K znatno zmanjšajo. Gručenji s $K = 6$ in 12 ustrezata referenčnim modelom, medtem ko gručenje pri $K = 10$ ustrezza GZ modelu, kjer sta obe konstantni domeni levega in desnega kraka združeni v en delec. Takšni GZ modeli se v literaturi še ne pojavljajo, bi pa morda lahko bili smiselnii.

POGLAVJE 4. REZULTATI



Slika 4.4: Metodi ‘komolca’ in ‘silhuete’ za k-means, izveden na vsaki verigi posebej (a), in bisekcijski k-means (b). Na grafih zgoraj se leva y -os nanaša na podatke za težko verigo (označena z modrimi točkami), desna pa za lahko verigo (označena z oranžnimi točkami). Graf metode ‘komolca’ za bisekcijski k-means vsebuje podatke o tem, za koliko se cenovna funkcija \mathcal{C} na vsakem koraku zmanjša, kar je prikazano z oranžnim stolpičnim diagramom. Diagram je skaliran logaritemsko, kot nakazuje desna y -os.

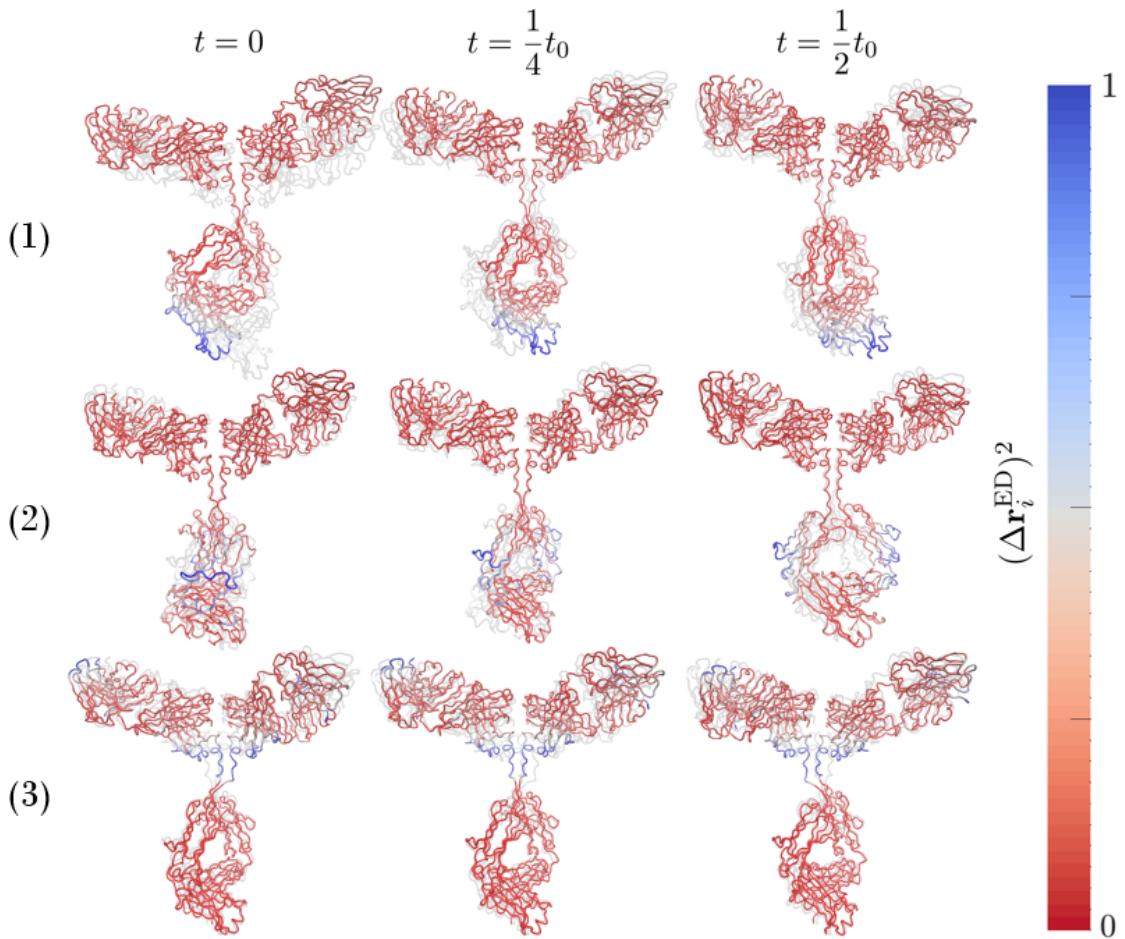
4.1.2 GZ modeliranje z metodo bistvene dinamike

Za uporabo metode bistvene dinamike je potrebno molekularne strukture iz [18] nekoliko prirediti, saj metoda razdeli molekulo v dinamične domene glede na zaporedje ogljik- α atomov, zaradi česar postane pomembno, v kakšnem zaporedju si sledijo verige iz referenčne strukture. V nasprotnem primeru se lahko pojavijo artefakti v GZ modelu, pri katerih v isto domeno združimo ogljik- α atome, ki so prostorsko oddaljeni med seboj. Strukture tako priredimo na način, da se stikajo skupaj tisti konci verig, ki so blizu skupaj, recimo v zaporedje: $C_L - V_L - V_H - C_{H1} - C_{H2} - C_{H3} - C_{H3}^* - C_{H2}^* - C_{H1}^* - V_H^* - V_L^* - C_L^*$.

Za uporabo metode bistvene dinamike je potrebno najprej skonstruirati elastičen mrežni model in na njem izvesti analizo normalnih načinov. To storimo z uporabe python knjižnice `prody` [77]. Na relaksirani referenčni strukturi protitelesa je uporabljen anizotropni elastični model z mejno razdaljo 1.5 nm — preizkušeni so bili tudi modeli z več mejnimi razdaljami, a niso doprinesli bistvene spremembe h končnemu rezultatu¹. Znotraj knjižnice izračunamo lastne nihajne načine preko ukaza `calcModes()`, ki vrne set lastnih načinov in njihovih frekvenc, ki so razvrščeni od

¹Za takšne modele je znotraj knjižnice `prody` potrebna lastna implementacija izračuna Hessejeve matrike.

dolgovalovnih do kratkovalovnih načinov. Vizualizacije prvih treh lastnih načinov so prikazane na Sliki 4.5.

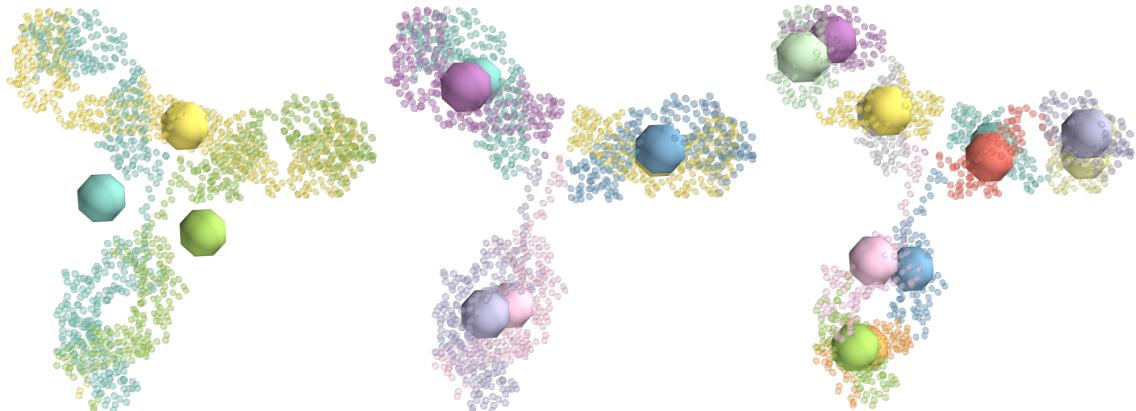


Slika 4.5: Prvi trije lastni nihajni načini monoklonskega protitelesa, prikazani ob treh različnih trenutkih, glede na nihajni čas t_0 . Hrbtenica molekule (povezani zaporedni ogljik- α atomi) je obarvana glede na velikost fluktuacij $(\Delta r_i^{\text{ED}})^2$, ki so normirane na interval $[0, 1]$. S sivo so prikazane strukture molekule ob $t - t_0/4$.

Lastni načini so prikazani glede nihajni čas t_0 . Prvi način opisuje nihanje spodnjega kraka molekule proti levemu in desnemu kraku, kjer se najbolj premikajo atomi na dnu molekule, saj prepotujejo največjo pot ob nihaju. Drugi lastni način je povezan z vrtenjem spodnjega dela molekule okoli navpične osi, tretji način pa s hkratno rotacijo levega in desnega kraka okoli vodoravne osi. Vsak izmed prvih treh načinov opisuje rotacijo okoli druge prostorskih osi. Iz teh nihajnih načinov se jasno vidi, da se nekateri deli molekule premikajo korelirano. Animacije izbranih lastnih nihajnih načinov so dostopne na [78]. Višji nihajni načini, ki opisujejo hitrejše oscilacije, so za nas manj zanimivi, saj ne opisujejo kolektivnega gibanja molekule, prav tako pa so bolj občutljivi na napake v referenčni strukturi in zato manj zanesljivi.

Na izračunanih lastnih načinih izvedemo metodo bistvene dinamike, ki je implementirana znotraj knjižnice OpenMSCG [79], kjer kot edini parameter nastopa število psevdono-atomov GZ modela. Vizualizacije rezultatov za 3,6 in 12 psevdono-atomov se nahajajo na Sliki 4.6.

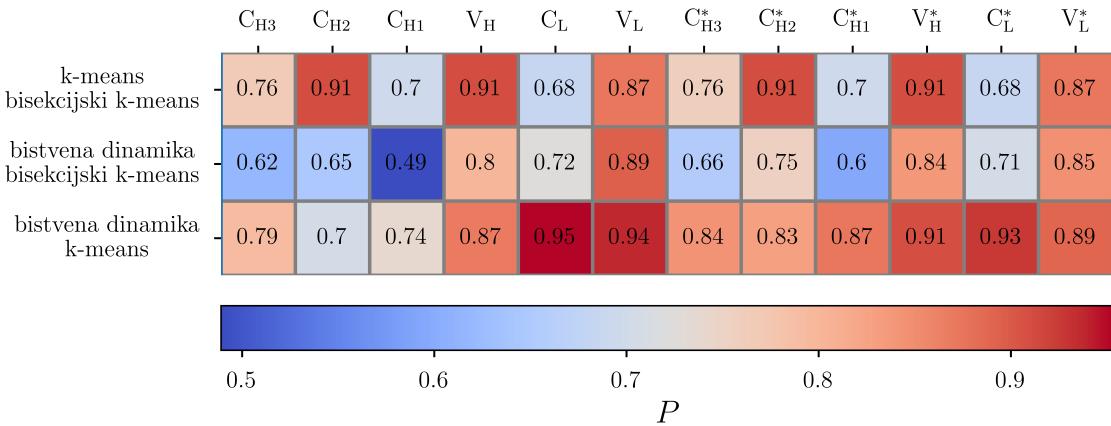
POGLAVJE 4. REZULTATI



Slika 4.6: GZ modeli s tremi, šestimi in dvanajstimi psevdo-atomimi, konstruirani z metodo bistvene dinamike.

GZ model s tremi psevdo-atomimi se bistveno razlikuje od referenčnih, saj se izkaže, da je v oziru nihajnih načinov smiselno z enim psevdo-atomom opisati obe lahki verigi skupaj in z enim vsako težko verigo. Takšna razdelitev je smiselna, saj npr. v drugem lastnem načinu s Slike 4.5 vidimo, da obe težki verigi (ozioroma njuni C_{H2} in C_{H3} domeni) nihata ravno v nasprotni smeri in je zato smiselno, da ju ločimo v različna psevdo-atoma. Kljub temu je primernost GZ modela, ki ga generira takšna razdelitev, vprašljiva, saj njegova struktura ne odraža dejanske strukture monoklon-skega protitelesa. Model s šestimi psevdo-atomimi je podoben temu s tremi gradniki, le da vsako izmed domen razdeli na dva dela. Tako sta obe lahki verigi predstavljeni s svojim psevdo-atomom, težki verigi pa razdeljeni na zgornji in spodnji del. Tudi tak model se razlikuje od referenčnega, pri katerem so v isti gradnik združeni ogljik- α atomi iz različnih verig, za razliko od rezultatov metode bistvene dinamike. 12-delčni GZ model je kvalitativno podoben referenčnemu, saj je vsaka funkcionalna domena predstavljena s svojim psevdo-atomom, podobno kot pri rezultatih gručenja k-means (Slika 4.3). Kljub kvalitativni podobnosti vseh treh modelov pa s Slike 4.7 vidimo, da dinamične domene, prepoznane z metodo bistvene dinamike, niso povsem enake gručam večje gostote, določenimi z metodama k-means. Najboljše se ujemata modela k-means, ki smo ga izvedli za vsako verigo posebej, in model bistvene dinamike, pri katerih so največja prekrivanja med domenami C_L , V_L in V_H . Bistvene razlike med temi modela se pojavijo predvsem v tečajnjem predelu, torej na meji med C_{H1} in C_{H2} . GZ modela bisekcijskega k-means in bistvene dinamike se slabše ujemata, največja odstopanja se ponovno pojavijo v tečajnjem predelu, kjer je prekrivanje domen C_{H1} zgolj 49%. Model bistvene dinamike ni simetričen, zato tudi prekrivanja med simetričnimi pari domen niso enaka.

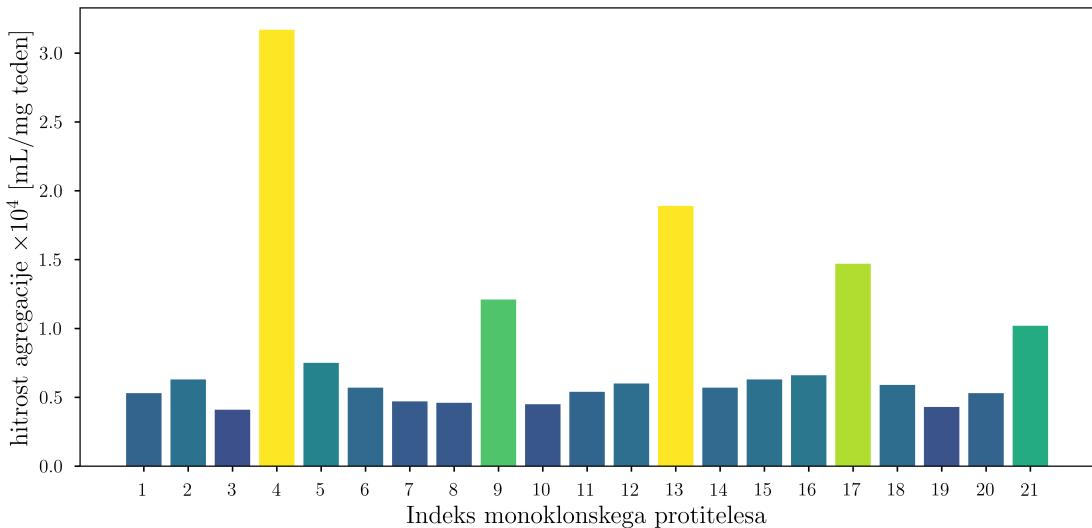
4.2. KORELACIJE LASTNIH NIHAJNIH NAČINOV IN HITROSTI AGREGACIJE



Slika 4.7: Prekrivanje P med GZ modeli k-means, bisekcijskim k-means in bistvene dinamike. Prekrivanja so obarvana, kot ponazarja barvna shema na dnu slike.

4.2 Korelacije lastnih nihajnih načinov in hitrosti agregacije

Del podatkovnega seta 21 monoklonskih protiteles so tudi eksperimentalni podatki o hitrosti agregacije. Gre za hitrosti, izračunane glede na dinamiko (2.1) iz meritev koncentracije monomerov (neaggregiranih protiteles) v časovnem obdobju dveh tednov. Podrobnejši podatki o izvajanju meritev so opisani v [18]. Podatki o hitrosti agregacije so prikazani na Sliki 4.8, s katere vidimo, da ima večino monoklonskih protiteles v setu nizko hitrost agregacije, ki znaša okoli 0.5×10^4 mL/mg teden, pri petih monoklonskih protitelesih pa opazimo hitrejšo agregacijo.



Slika 4.8: Hitrosti agregacije monoklonskih protiteles iz podatkovnega seta [18] podane v enotah mL/mg teden. Stolpci so glede na hitrosti agregacije obarvani od temno modre do rumene barve, kjer je hitrost agregacije najvišja.

Različni dejavniki lahko vodijo do agregacije protiteles, vsem pa je skupno, da vplivajo tudi na dinamične lastnosti molekule. Cilj tega podpoglavlja je raziskati, ali

POGLAVJE 4. REZULTATI

lahko iz dinamičnih lastnosti molekule, določenih preko elastičnega mrežnega modela in analize lastnih nihajnih načinov, povemo kaj o hitrosti agregacije monoklonskih protiteles. Povezave iščemo z računanjem linearnih korelacij različnih dinamičnih lastnosti molekule in hitrostjo agregacije.

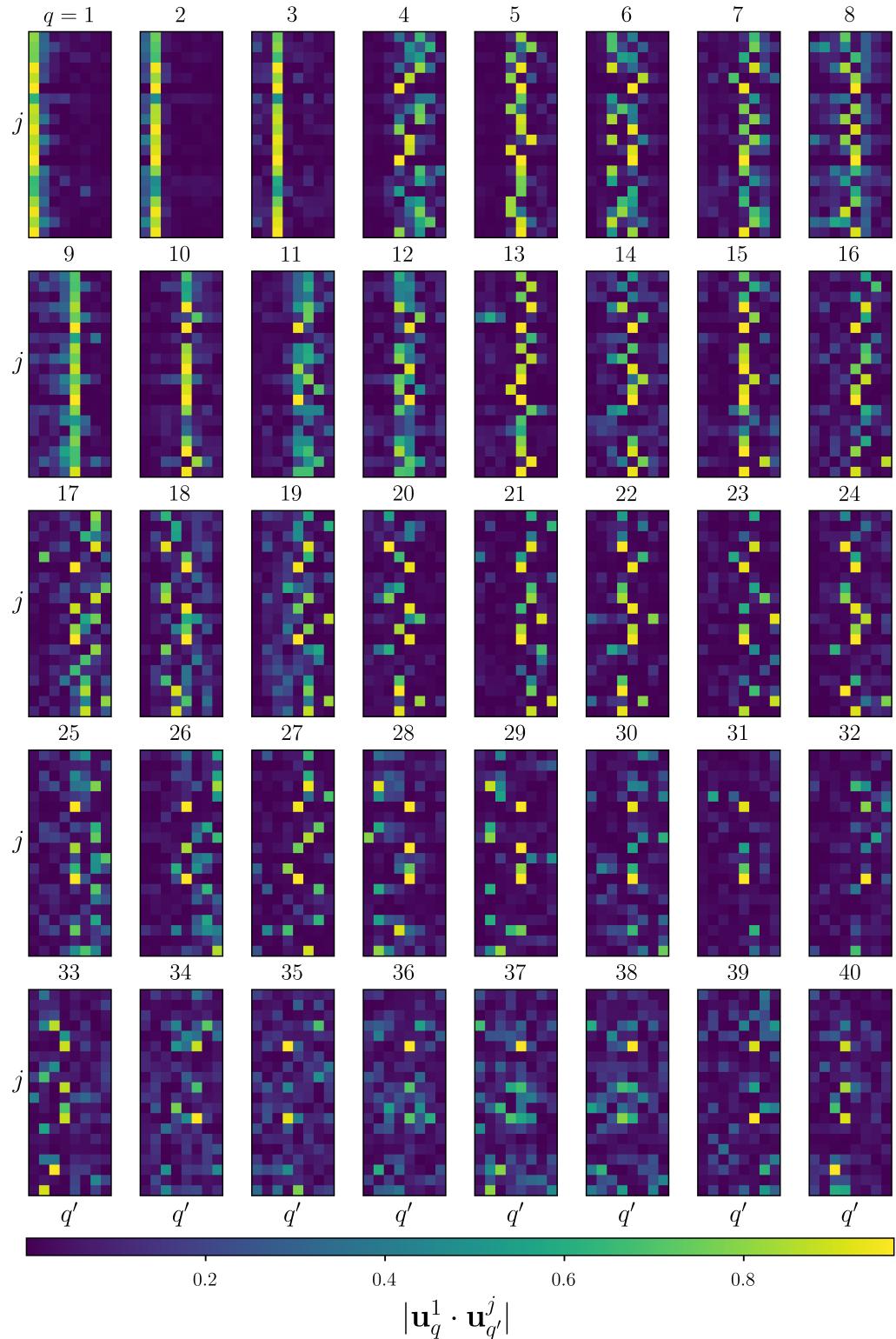
4.2.1 Frekvence lastnih načinov

Z analizo lastnih nihajnih načinov pridobimo lastne vektorje, ki opisujejo lastne načine in lastne vrednosti, ki predstavljajo frekvence nihanja lastnih načinov. Osrednji namen tega razdelka je ugotoviti, ali frekvence lastnih načinov korelirajo s hitrostjo agregacije monoklonskih protiteles, pri čemer se omejimo zgolj na iskanje linearnih korelacij. Fizikalna motivacija za iskanje korelacij je v tem, da je morda kakšen izmed lastnih načinov povezan z dinamiko monoklonskega protitelesa v procesu agregaciji. Če takšna povezava obstaja, bi lahko bila frekvenca takšnega nihanja povezana s hitrostjo agregacije.

Primerjati želimo frekvence specifičnih nihajnih načinov 21 monoklonskih protiteles, pri čemer se omejimo na najnižje lastne načine, za katere vemo, da so povezani s kolektivnim gibanjem molekule — izberemo prvih 40 nihajnih načinov. Ker uporabljena implementacija analize lastnih nihajnih načinov vrne le-te v vrstnem redu, začenši s tistimi z najnižjo frekvenco, se izkaže, da q -ti lastni način posameznega monoklonskega protitelesa ne opisuje nujno enakega načina gibanja za različna protitelesa. q -ti lastni način enega izmed protiteles lahko ustrezza nihajnemu načinu q' nekega drugega protitelesa, ali pa se izkaže, da je q -ti lastni način za to monoklonsko protitelo specifičen, saj ga ne opazimo pri drugih protitelesih. Za smiselnoprimerjavo frekvenc je tako potrebno najprej osnovati sete podobnih lastnih nihajnih načinov. Za primerjavo lastnih načinov najprej prirežemo vse lastne vektorje monoklonskih protiteles na dolžino najkrajšega, kar v praksi pomeni, da bomo iz primerjava izpustili do deset ogljik- α atomov konstantne domene ene lahke verige. Ta predel je izbran namenoma, saj se nahaja v osrčju molekule, a ni del tečajnega predela, zato tam pričakujemo manj gibanja. Podobne lastne načine iščemo glede na lastne načine prvega monoklonskega protitelesa v setu. Za vsak lastni način \mathbf{u}_q^1 (indeks 1 označuje, da gre za prvo monoklonsko protitelo iz seta) poiščemo lastni način $\mathbf{u}_{q'}^j$ za j -to monoklonsko protitelo iz seta, ki se najbolje ujema z njim. Podobnost lastnih nihajnih načinov določimo glede na skalarni produkt vektorja nihajnih načinov, pri čemer vzamemo njegovo absolutno vrednost, saj so nekateri nihajni načini enaki, le zamaknjeni za fazo π . Če so vsi vektorji premika posameznih ogljik- α atomov za \mathbf{u}_q^1 in $\mathbf{u}_{q'}^j$ točno enaki, je $\mathbf{u}_q^1 \cdot \mathbf{u}_{q'}^j = 1$. Na Sliki 4.9 so prikazani skalarni produkti prvih 40 nihajnih načinov prvega monoklonskega protitelesa z nekaterimi lastnimi načini ostalih protiteles. q -ti lastni način prvega monoklonskega protitelesa je primerjan z devetimi lastnimi načini vsakega j -tega monoklonskega protitelesa, tako da je $q' \in [q - 4, q + 4]$.

Z izjemo prvih treh lastnih načinov, kjer je $q = q'$ v prvem, drugem oziroma tretjem stolpcu, enakost $q = q'$ velja za sredinski stolpec. Pri $q = 1, 2$ in 3 dobimo največji skalarni produkt, ko je $q = q'$, iz česar sklepamo, da so prvi trije nihajni načini podobni za vsa monoklonska protitelesa. To potrdi tudi ogled animacij lastnih nihajnih načinov znotraj programa VMD. Pri nekaterih lastnih načinih najdemo podobne načine pri čisto vseh drugih protitelesih — peti nihajni način (glede na prvo monoklonsko protitelo) se pojavi pri drugih protitelesih kot četrti, peti ali šesti lastni način, najnižji skalarni produkt pa je enak 0.75. Predvsem v višjih nihajnih

4.2. KORELACIJE LASTNIH NIHAJNIH NAČINOV IN HITROSTI AGREGACIJE

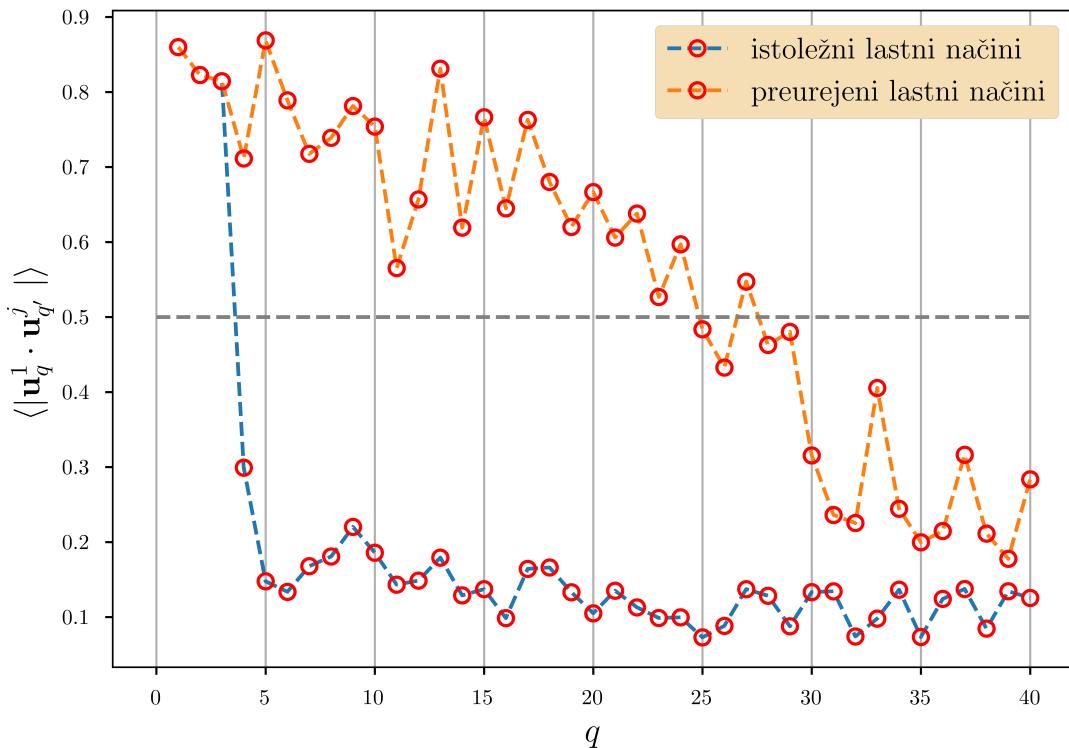


Slika 4.9: Skalarni produkti q -tih lastnih načinov prvega monoklonskega protitelesa s q' -timi lastnimi načini vseh ostalih protiteles. Vsaka sličica ustrez enemu izmed lastnih načinov q . Na y -osi vsake sličice je prikazan indeks monoklonskega protitelesa $j \in [2, 21]$, na x -osi pa q' -ti lastni načini, kjer je $q' \in [q-4, q+4]$. Velikost skalarnega produkta je obarvana glede na barvno shemo na dnu slike.

POGLAVJE 4. REZULTATI

načinih pa se pogosto zgodi, da se podobni načini ne pojavijo pri drugih protitelesih. Pojavijo se tudi primeri, ko je sicer skalarni produkt relativno nizek (npr. $\mathbf{u}_1^1 \cdot \mathbf{u}_1^8 = 0.64$), kvalitativno pa sta lastna načina podobna (glede na ogled animacij), kar kaže na pomanjkljivost uporabe skalarnega produkta prirezanih lastnih načinov kot metrike podobnosti.

Glede na izračunane skalarne produkte sestavimo sete lastnih načinov, ki opisujejo podoben tip gibanja. To storimo na način, da v q -ti set dodelimo tiste lastne načine protiteles (po enega na protitelo), ki dajo največjo vrednost skalarnega produkta. Ker so v nekaterih primerih vsi skalarni produkti nizki, ni posebnega zagotovila, da bi bil tisti z najvišjo vrednostjo nujno kvalitativno najbolj podoben, zato uvedemo minimalni prag, ki ga mora skalarni produkt preseči, da ga dodamo v set. Če za kakšno protitelo ne najdemo nobenega lastnega načina, ki bi dal dovolj velik skalarni produkt, v set dodelimo kar istoležni nihajni način. Po analizi skalarnih produktov in kvalitativni primerjavi lastnih načinov postavimo minimalen prag na 0.5. Za sete prvih 40 lastnih načinov izračunamo povprečen skalarni produkt in ga primerjamo s povprečnim skalarnim produkтом, ki ga dobimo, če bi v sete združevali istoležne nihajne načine, kar je prikazano na Sliki 4.10.



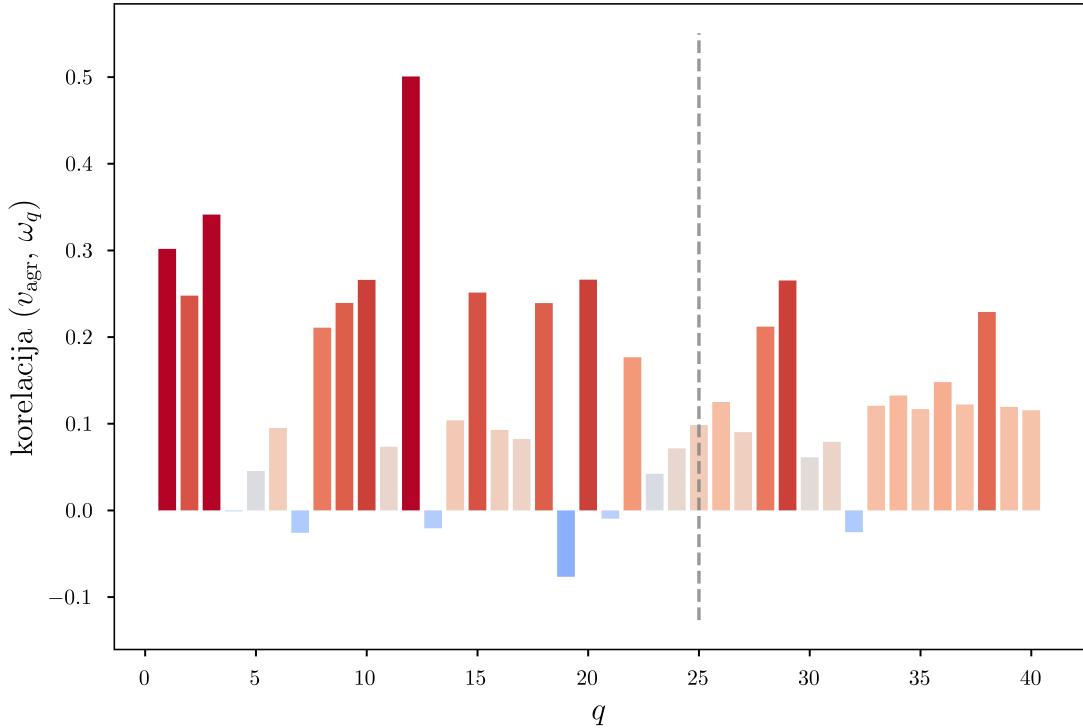
Slika 4.10: Primerjava povprečnega skalarnega produkta setov lastnih načinov, ko združujemo skupaj istoležne lastne načine in ko vrstni red preuredimo, tako da združimo skupaj lastne načine z največjim skalarnim produkтом. S črtko je vrisan prag minimalnega skalarnega produkta.

Za prve tri lastne načine je povprečni skalarni produkt enak ne glede na preurejanje, saj so prvi trije načini za vsa monoklonska protitelesa enaki. Pri vseh ostalih načinih imajo preurejeni seti mnogo višji povprečni skalarni produkt, kar potrjuje smiselnost tvorjenja preurejenih setov. Pri $q = 25$ vrednost povprečnega skalarnega produkta prvič pada pod prag minimalnega skalarnega produkta, zato ni verjetno,

4.2. KORELACIJE LASTNIH NIHAJNIH NAČINOV IN HITROSTI AGREGACIJE

da bi seti višjih lastnih načinov zares opisovalni podoben način gibanja protitelesa. Vse morebitne korelacijske z višjimi načini tako interpretiramo kot statističnega izvora zaradi majhnega seta podatkov.

Za preurejen set lastnih nihajnih načinov izračunamo linearne korelacijske frekvence s hitrostjo agregacije. Rezultati so prikazani na Sliki 4.11.



Slika 4.11: Stolpični diagram linearnih korelacijskih frekvenc nihajnih načinov s hitrostjo agregacije. Stolpci so obarvani z rdečo v primeru pozitivne korelacijske frekvence in z modro v primeru negativne korelacijske frekvence. Z vertikalno črtkano črto je označen 25. lastni način.

Najvišjo korelacijsko frekvenco, ki znaša 0.49, opazimo pri 12. lastnem nihajnjem načinu, nato pa po velikosti sledita prvi in tretji lastni način. Vse ostale korelacijske frekvence so nižje od 0.3 in jih ne smatramo za znatne. Glede na razmeroma močno korelacijsko frekvenco 12. lastnega načina in hitrosti agregacije lahko predvidevamo, da je način gibanja, ki ga opisuje ta lastni način, res povezan s procesom agregacije. V tem oziru bi lahko višja frekvenca pomenila hitrejše nihanje in s tem potencialno ugodnejše pogoje za tvorjenje polimerov. Za potrditev te interpretacije bi bilo potrebno opraviti dodatne analize in eksperimente molekularne dinamike.

Potrebeno se je zavedati, da so rezultati osnovani glede na preurejen set nihajnih načinov, ki združi skupaj tiste načine, ki imajo velik skalarni produkt, kar morda ni najboljša metrika za določanje podobnosti. Vprašljiva je tudi zato, ker računamo skalarni produkt glede na istoležne ogljik- α atome, in ne med tistimi, ki se nahajajo v isti točki prostora. Pojavlja se tudi vprašanje, kakšne korelacijske frekvence bi dobili, če bi konstruirali tudi manjše sete lastnih načinov, v katerih bi bila le tista monoklonska protitelesa, pri katerih bi našli skalarni produkt, ki bi bil večji od izbranega praga.

Poleg linearnih korelacijskih frekvenc smo iskali tudi korelacijske z izpeljanimi količinami, kot so vsota prvih M frekvenc in vsota kombinacij frekvenc, a nismo našli

znatnih korelacij.

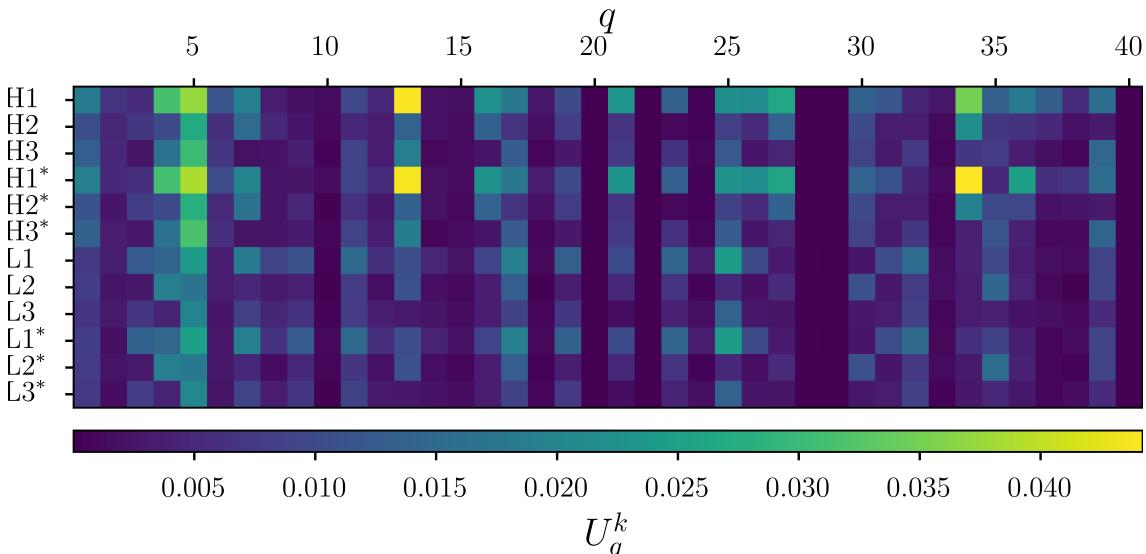
4.2.2 Nihanja v variabilnih regijah

Monoklonska protitelesa se med seboj razlikujejo predvsem v predelih variabilnih regij, najbolj izrazito v 12 CDR zankah, zato pričakujemo, da se bodo lastni načini posameznih protiteles razlikovali v načinu nihanja znotraj CDR zank. Namen tega podpoglavlja je preveriti, ali obstajajo korelacije med vsoto amplitud nihanja v variabilnih domenah in hitrostjo agregacije. Definiramo lahko vsoto amplitud nihanja q -tega lastnega načina v k -ti CDR zanki kot

$$U_q^k = \sum_{i \in S_{\text{CDR}}^k} \|\mathbf{u}_q^i\|, \quad (4.3)$$

kjer vsota teče po vseh ogljik- α atomih i , ki so v S_{CDR}^k , torej znotraj seta atomov k -te CDR zanke. Z \mathbf{u}_q^i je označen vektor premika i -tega ogljik- α atoma v q -tem lastnem načinu. U_q^k je sestavljen iz norm vektorjev premika in predstavlja mero za znatnost nihanja v CDR zanki. Ker so vsi lastni vektorji normirani, U_q^k hkrati predstavlja tudi relativen delež vsote amplitud nihanja CDR zanke v q -tem lastnem načinu.

Lokacije CDR zank znotraj aminokislinskega zaporedja monoklonskega protitelesa določimo preko spletnega klasifikatorja PyIgClassify2 [80], ki glede na referenčno strukturo protitelesa poišče, katere aminokisline (in s tem ogljik- α atomi) sestavljajo posamezno CDR zanko². Za vsako protitelo izračunamo U_q^k za $q \in [1, 40]$ in $k \in [1, 12]$. Rezultati za prvo monoklonsko protitelo so prikazani na Sliki 4.12.



Slika 4.12: Vsota amplitud nihanja U_q^k prvega monoklonskega protitelesa za prvih 40 nihajnih načinov in vseh 12 CDR zank. Vsaka vrstica označuje eno izmed zank. Oznaka H1 pomeni prvo CDR zanko težke (ang. *heavy*) verige. Zanke simetričnih verig so označene z zvezdico.

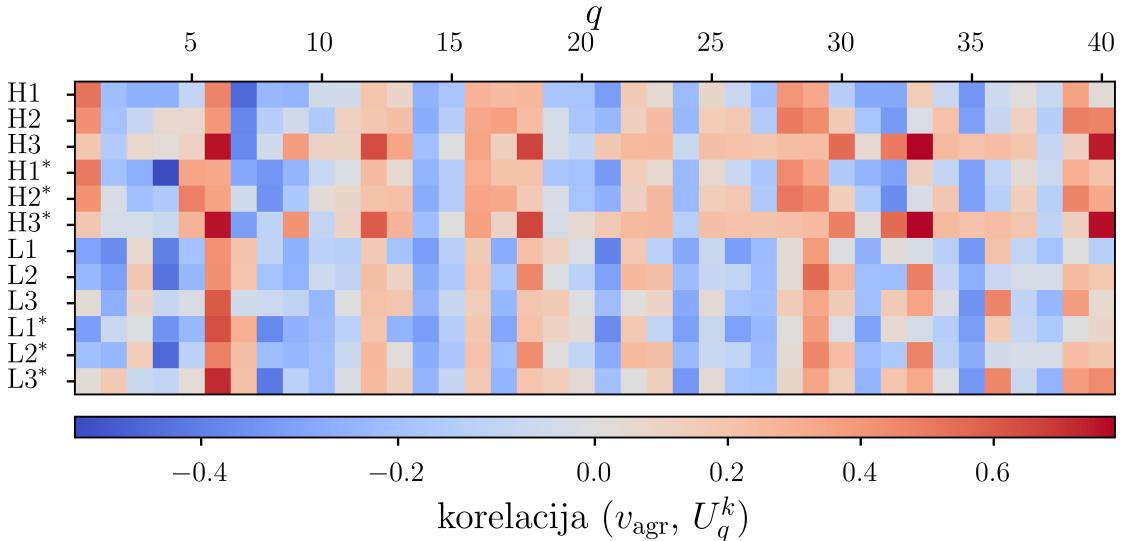
Najmočnejše nihanje prvega monoklonskega telesa opazimo v 13. in 5. lastnem načinu v prvi CDR zanki težke verige (H1). Znatno nihanje se pojavi sicer tudi v

²Trenutna verzija klasifikatorja ne omogoča več klasifikacije na poljubni referenčni strukturi, pač pa služi kot podatkovna baza že klasificiranih protiteles.

4.2. KORELACIJE LASTNIH NIHAJNIH NAČINOV IN HITROSTI AGREGACIJE

34. lastnem načinu, a mu ne pripisemo posebnega pomena, saj za tako visoke lastne načine ne moremo osnovati smiselnih setov za računanje korelacije. V večini primerov je nihanje enako močno v obeh simetričnih parih posamezne CDR zanke. Izjeme se pojavljajo predvsem v višjih nihajnih načinih, kjer je nihanje bolj lokalizirano.

Za sete preurejenih nihajnih načinov izračunamo korelacije med posameznimi U_q^k in hitrostjo agregacije. Rezultati so prikazani na Sliki 4.13.



Slika 4.13: Korelacija hitrosti agregacije in vsote amplitud nihanja U_q^k za $q \in [1, 40]$ in $k \in [1, 12]$. Vsaka slikovna točka predstavlja linearno korelacijo in je obravljana glede na barvno shemo na dnu slike.

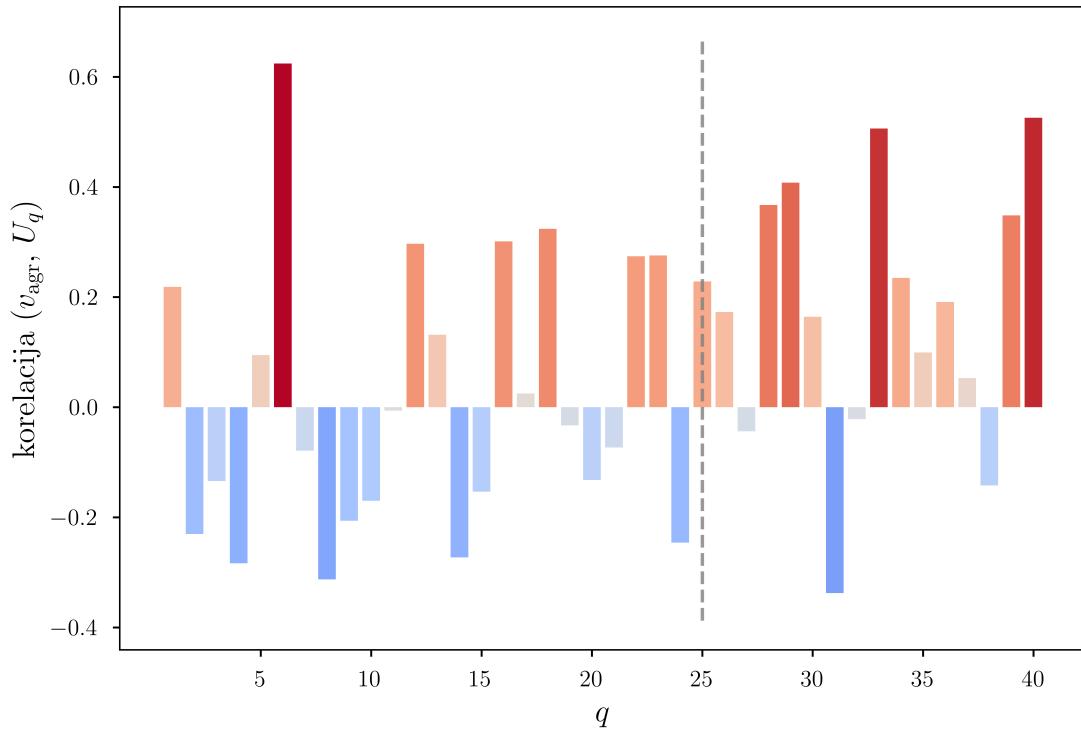
V rezultatih so močnejše pozitivne korelacije kot negativne, kar je smiselno, saj je aggregacija v splošnem povezana s povečanim gibanjem molekule. Najvišje pozitivne korelacije znašajo okoli 0.7, negativne pa -0.5. Zdi se, da je H3 zanka pogosteje povezana z višjo korelacijo, saj se močen signal tam pojavi v več lastnih načinih — v 6., 12., 18. in v višjih načinih. Močne korelacije opazimo tudi pri vseh CDR zankah šestega nihajnega načina, kar bi lahko bil znak, da je nihanje CDR zank v tovrstnem načinu povezano z aggregacijo: če monoklonsko protitelelo v 6. nihajnemu načinu (glede na nihajne načine prvega monoklonskega protitelesa) izrazito niha v predelu variabilne domene, verjetno tudi hitreje aggregira. Korelacijski signal pri nihajnih načinih višjih od $q = 25$ interpretiramo kot signal statistične narave, saj tisti seti lastnih načinov niso nujno smiselnii.

Takšen način iskanja korelacij kot je prikazan na Sliki 4.13, je gotovo podvržen statističnim artefaktom, ki se pojavijo zaradi majhnosti podatkovnega seta. Da smemo označiti določeno korelacijo kot resničen indikator hitrejše aggregacije, jo je potrebno fizikalno upravičiti. V tem oziru je za določeno CDR zanko težko upravičiti izjavo o korelaciji, saj gre za zelo specifične izjave. Da bi se znebili statističnih artefaktov, koreliramo hitrost aggregacije z vsoto amplitud nihanja čez celotno variabilno domeno, kar je prikazano na Sliki 4.14.

Podobno kot na Sliki 4.13 se najmočnejša korelacija pojavi v 6. nihajnjem načinu in znaša približno 0.6. Znatne (> 0.3) korelacije se pojavijo zgolj še v višjih nihajnih načinih, a jih kot prej interpretiramo kot signal statistične narave.

Glede na velikost korelacije pri 6. nihajnjem načinu lahko sklenemo podoben zaključek kot pri interpretaciji Slike 4.13 — če protitelelo v 6. nihajnjemu načinu

POGLAVJE 4. REZULTATI



Slika 4.14: Korelacje vsote amplitud nihanja celotne variabilne domene s hitrostjo agregacije. Kot pri drugih vizualizacijah, so pozitivne korelacie nakazane z rdečo, negativne pa z modro barvo. Vertikalna črtkana črta pri $q = 25$ označuje do katerega lastnega načina sete nihajnih načinov obravnavamo kot smiselne.

izrazito niha v predelu variabilne domene, verjetno tudi hitreje agregira. Za potrditev takšne interpretacije je potrebna dodatna statistična analiza rezultatov (npr. z metodo ponovnega vzorčenja (ang. *bootstrapping*)), povečanje podatkovnega seta in analiza ujemanja lastnih načinov z dinamiko molekule v procesu agregacije.

5. Zaključek

V magistrskem delu smo pokazali, da je z metodo strojnega učenja k-means mogoče skonstruirati GZ modele, ki so kvalitativno podobni referenčnim, že preizkušenim GZ modelom za monoklonska protitelesa. Uspešno smo reproducirali modele s 3, 6 in 12 psevdo-atomi, medtem ko modelov, ki postavijo psevdo-atom v tečajni predel, z metodo k-means ni mogoče poustvariti. Posebej smo se osredotočili na model z 12 psevdo-atomi, saj je najpogosteje uporabljen v literaturi, in ga konstruirali z bisekcijskim k-means in k-means algoritmom, izvedenim na vsaki verigi protitelesa posebej. Modela sta si kvalitativno podobna glede na strukturo psevdo-atomov, saj se njihove pozicije le malo razlikujejo, bolj pa se razlikujeta v načinu, katere ogljik- α atome združita v posamezen psevdo-atom. Domene se ujemajo v najmanj 68.4% in največ 91.1% pripadajočih ogljik- α atomih, kar bi postalo pomembno, ko bi določevali naboj psevdo-atomov. Ker je cilj naloge poiskati GZ model, ki bi se ga dalo prenesti tudi na druge vrste proteinov, smo pogledali, ali bi lahko brez informacije, da je 12-delčni model smiseln, vseeno uganili, koliko psevdo-atomov je potrebnih za dober opis strukture monoklonskega protitelesa. V ta namen smo uporabili metodi "komolca" in "silhuete", ki pa nista izkazali močne indikacije, da bi bila takšna napoved mogoča.

Monoklonska protitelesa smo modelirali tudi z uporabo metode bistvene dinamike, s katero smo reproducirali smiseln 12-delčni GZ model. Primerjali smo ga z 12-delčnima k-means modeloma in ugotovili, da je precej podoben GZ modelu, ki ga dobimo z metodo k-means, izvedeno na vsaki verigi monoklonskega protitelesa posebej. Med konstrukcijo GZ modela smo pridobili tudi lastne nihajne načine monoklonskih protiteles, ki bi lahko bili uporabni še v drugih kontekstih.

Vzpodbudni rezultati obeh preizkušenih metod konstrukcije GZ modela nakazujejo, da bi jih bilo mogoče uporabiti tudi na drugih proteinih, predvsem ko so le-ti razmeroma veliki in želimo določiti poenostavljenou strukturo za potrebe simulacij molekularne dinamike. Proces konstrukcije GZ modela z metodo k-means in metodo bistvene dinamike predstavlja nov pristop modeliranja monoklonskih protiteles in se od uveljavljenih postopkov, ki se že pojavljajo v literaturi, razlikuje predvsem v tem, da za konstrukcijo ne potrebujemo informacij o funkcionalnih domenah protiteles.

Da bi bili skonstruirani GZ modeli lahko uporabljeni za *in silico* eksperimente, je potrebno po že uveljavljenih postopkih z njimi določiti tudi nekatere lastnosti psevdo-atomov, kot so naboj, dipolni moment, masa ipd. Vse osnovane GZ modele bi bilo potrebno uporabiti za simulacije molekularne dinamike in preizkusiti, ali reproducirajo katere izmed dobro poznanih lastnosti monoklonskih protiteles.

V drugem delu magistrske naloge smo frekvence lastnih načinov korelirali s hitrostjo agregacije. Pri tem smo oblikovali sete lastnih načinov, ki so opisovali podobno gibanje glede na skalarni produkt lastnih načinov. V 12. lastnem načinu smo opazili največjo korelacijo, ki je znašala 0.49. Ta vrednost bi lahko pomenila, da je 12.

POGLAVJE 5. ZAKLJUČEK

nihajni način povezan s procesom agregacije, a magistrsko delo samo ne daje dovolj trdnih dokazov, da je temu tako.

Hitrost agregacije smo korelirali tudi z vsoto amplitud nihanja U_q^k v CDR zankah. Rezultati nakazujejo na to, da je tudi 6. nihajni način povezan s procesom agregacije, saj korelacija s hitrostjo agregacije in U_q znaša 0.62. Za podkrepitev obeh pomembnejših korelacij bi bila potrebna dodatna statistična analiza kot tudi ponovitev izračuna na večjem podatkovnem setu. Prav tako bi bila potrebna analiza procesa agregacije, da bi ugotovili, ali se mehanizmi nastajanja polimerov ujemajo s kakšnim izmed lastnih nihajnih načinov. Možnosti dodatnih raziskav se pojavljajo tudi pri načinu konstrukcije setov podobnih nihajnih načinov.

6. Literatura

- [1] G. Walsh in E. Walsh, *Biopharmaceutical Benchmarks 2022*, Nature Biotechnology **40**, 1722 (2022).
- [2] H. M. Shepard, G. L. Phillips, C. D. Thanos in M. Feldmann, *Developments in Therapy with Monoclonal Antibodies and Related Proteins*, Clinical Medicine **17**, 220 (2017).
- [3] V. Bayer, *An Overview of Monoclonal Antibodies*, Seminars in Oncology Nursing Immunotherapy in Oncology, **35**, 150927 (2019).
- [4] S. Singh, N. K. Tank, P. Dwiwedi, J. Charan, R. Kaur, P. Sidhu in V. K. Chugh, *Monoclonal Antibodies: A Review*, Current Clinical Pharmacology **13**, 85 (2018).
- [5] M. S. Castelli, P. McGonigle in P. J. Hornby, *The Pharmacology and Therapeutic Applications of Monoclonal Antibodies*, Pharmacology Research & Perspectives **7**, e00535 (2019).
- [6] V. P. Chavda, R. Prajapati, D. Lathigara, B. Nagar, J. Kukadiya, E. M. Redwan, V. N. Uversky, M. N. Kher in R. Patel, *Therapeutic Monoclonal Antibodies for COVID-19 Management: An Update*, Expert Opinion on Biological Therapy **22**, 763 (2022).
- [7] R. Ghosh, C. Calero-Rubio, A. Saluja in C. J. Roberts, *Relating Protein–Protein Interactions and Aggregation Rates from Low to High Concentrations*, Journal of Pharmaceutical Sciences **105**, 1086 (2016).
- [8] W. F. Weiss, T. M. Young in C. J. Roberts, *Principles, Approaches, and Challenges for Predicting Protein Aggregation Rates and Shelf Life*, Journal of Pharmaceutical Sciences **98**, 1246 (2009).
- [9] A. Saluja, R. M. Fesinmeyer, S. Hogan, D. N. Brems in Y. R. Gokarn, *Diffusion and Sedimentation Interaction Parameters for Measuring the Second Virial Coefficient and Their Utility as Predictors of Protein Aggregation*, Biophysical Journal **99**, 2657 (2010).
- [10] D. Roberts, R. Keeling, M. Tracka, C. F. van der Walle, S. Uddin, J. Warwick in R. Curtis, *The Role of Electrostatics in Protein–Protein Interactions of a Monoclonal Antibody*, Molecular Pharmaceutics **11**, 2475 (2014).
- [11] M. M. C. van Beers in M. Bardor, *Minimizing Immunogenicity of Biopharmaceuticals by Controlling Critical Quality Attributes of Proteins*, Biotechnology Journal **7**, 1473 (2012).

POGLAVJE 6. LITERATURA

- [12] S. J. Shire, Z. Shahrokh in J. Liu, *Challenges in the Development of High Protein Concentration Formulations*, Journal of Pharmaceutical Sciences **93**, 1390 (2004).
- [13] M. S. Neergaard, D. S. Kalonia, H. Parshad, A. D. Nielsen, E. H. Møller in M. van de Weert, *Viscosity of High Concentration Protein Formulations of Monoclonal Antibodies of the IgG1 and IgG4 Subclass – Prediction of Viscosity through Protein–Protein Interaction Measurements*, European Journal of Pharmaceutical Sciences **49**, 400 (2013).
- [14] L. Li, S. Kumar, P. M. Buck, C. Burns, J. Lavoie, S. K. Singh, N. W. Warne, P. Nichols, N. Luksha in D. Boardman, *Concentration Dependent Viscosity of Monoclonal Antibody Solutions: Explaining Experimental Behavior in Terms of Molecular Properties*, Pharmaceutical Research **31**, 3161 (2014).
- [15] J. S. Kingsbury, A. Saini, S. M. Auclair, L. Fu, M. M. Lantz, K. T. Halloran, C. Calero-Rubio, W. Schwenger, C. Y. Airiau, J. Zhang in Y. R. Gokarn, *A Single Molecular Descriptor to Predict Solution Behavior of Therapeutic Antibodies*, Science Advances **6**, eabb0372 (2020).
- [16] C. O. Calero-Rubio, *Protein Interactions, Unfolding and Aggregation from Low to High Protein Concentrations via Coarse-Grained Molecular Modeling and Experimental Characterization*, Doktorska disertacija, University of Delaware (2017).
- [17] P.-K. Lai, J. W. Swan in B. L. Trout, *Calculation of Therapeutic Antibody Viscosity with Coarse-Grained Models, Hydrodynamic Calculations and Machine Learning-Based Parameters*, mAbs **13**, 1907882 (2021).
- [18] P.-K. Lai, A. Fernando, T. K. Cloutier, Y. Gokarn, J. Zhang, W. Schwenger, R. Chari, C. Calero-Rubio in B. L. Trout, *Machine Learning Applied to Determine the Molecular Descriptors Responsible for the Viscosity Behavior of Concentrated Therapeutic Antibodies*, Molecular Pharmaceutics **18**, 1167 (2021).
- [19] A. Chaudhri, I. E. Zarraga, T. J. Kamerzell, J. P. Brandt, T. W. Patapoff, S. J. Shire in G. A. Voth, *Coarse-Grained Modeling of the Self-Association of Therapeutic Monoclonal Antibodies*, The Journal of Physical Chemistry B **116**, 8045 (2012).
- [20] A. Chaudhri, I. E. Zarraga, S. Yadav, T. W. Patapoff, S. J. Shire in G. A. Voth, *The Role of Amino Acid Sequence in the Self-Association of Therapeutic Monoclonal Antibodies: Insights from Coarse-Grained Modeling*, The Journal of Physical Chemistry B **117**, 1269 (2013).
- [21] T. Bereau in M. Deserno, *Generic Coarse-Grained Model for Protein Folding and Aggregation*, The Journal of Chemical Physics **130**, 235106 (2009).
- [22] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman in S.-J. Marrink, *The MARTINI Coarse-Grained Force Field: Extension to Proteins*, Journal of Chemical Theory and Computation **4**, 819 (2008).

-
- [23] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura in D. Baker, *Protein Structure Prediction Using Rosetta*, v *Methods in Enzymology*, Numerical Computer Methods, Part D, Zv. 383 (Academic Press, 2004) str. 66–93.
 - [24] J. Maupetit, P. Tuffery in P. Derreumaux, *A Coarse-Grained Protein Force Field for Folding and Structure Prediction*, Proteins: Structure, Function, and Bioinformatics **69**, 394 (2007).
 - [25] S. Izvekov in G. A. Voth, *A Multiscale Coarse-Graining Method for Biomolecular Systems*, The Journal of Physical Chemistry B **109**, 2469 (2005).
 - [26] H. Shahfar, J. K. Forder in C. J. Roberts, *Toward a Suite of Coarse-Grained Models for Molecular Simulation of Monoclonal Antibodies and Therapeutic Proteins*, The Journal of Physical Chemistry B **125**, 3574 (2021).
 - [27] D. L. Nelson, A. L. Lehninger in M. M. Cox, *Lehninger Principles of Biochemistry*, 5. izd. (W.H. Freeman, New York, 2008).
 - [28] C. I. Branden in J. Tooze, *Introduction to Protein Structure*, 2. izd. (Garland Science, New York, 1998).
 - [29] *CH450 and CH451: Biochemistry - Defining Life at the Molecular Level* (2019).
 - [30] R. C. Robinson, K. Turbedsky, D. A. Kaiser, J. B. Marchand, H. N. Higgs, S. Choe in T. D. Pollard, *Crystal Structure of Arp2/3 Complex*, Science (New York, N.Y.) **294**, 1679 (2001).
 - [31] E. C. Meng, T. D. Goddard, E. F. Pettersen, G. S. Couch, Z. J. Pearson, J. H. Morris in T. E. Ferrin, *UCSF ChimeraX: Tools for Structure Building and Analysis*, Protein Science : A Publication of the Protein Society **32**, e4792 (2023).
 - [32] J. W. Goding, *Monoclonal Antibodies: Principles and Practice*, 3. izd. (Academic Press, San Diego, 1996).
 - [33] P. N. Nelson, G. M. Reynolds, E. E. Waldron, E. Ward, K. Giannopoulos in P. G. Murray, *Monoclonal antibodies*, Molecular Pathology **53**, 111 (2000).
 - [34] M. Zidar, *Analysis and Prediction of Aggregation and Degradation in Protein-Based Biopharmaceuticals*, Doktorska disertacija, Univerza v Ljubljani, Fakulteta za matematiko in fiziko (2020).
 - [35] D. Arzenšek, *Physics of Colloidal Interactions in Protein Aggregation Processes*, Doktorska disertacija, Univerza v Ljubljani, Fakulteta za matematiko in fiziko (2015).
 - [36] A. Vaziri in A. Gopinath, *Cell and Biomolecular Mechanics in Silico*, Nature Materials **7**, 15 (2008).
 - [37] C. A. López, A. J. Rzepiela, A. H. de Vries, L. Dijkhuizen, P. H. Hünenberger in S. J. Marrink, *Martini Coarse-Grained Force Field: Extension to Carbohydrates*, Journal of Chemical Theory and Computation **5**, 3195 (2009).

POGLAVJE 6. LITERATURA

- [38] D. A. Potoyan, A. Savelyev in G. A. Papoian, *Recent Successes in Coarse-Grained Modeling of DNA*, WIREs Computational Molecular Science **3**, 69 (2013).
- [39] M. Orsi, D. Y. Haubertin, W. E. Sanderson in J. W. Essex, *A Quantitative Coarse-Grain Model for Lipid Bilayers*, The Journal of Physical Chemistry B **112**, 802 (2008).
- [40] S. Kmiecik, D. Gront, M. Kolinski, L. Witeska, A. E. Dawid in A. Kolinski, *Coarse-Grained Protein Models and Their Applications*, Chemical Reviews **116**, 7898 (2016).
- [41] L. E. Kay, *NMR Studies of Protein Structure and Dynamics*, Journal of Magnetic Resonance Magnetic Moments, **213**, 477 (2011).
- [42] A. Ilari in C. Savino, *Protein Structure Determination by X-Ray Crystallography*, v *Bioinformatics: Data, Sequence Analysis and Evolution*, ur. J. M. Keith (Humana Press, Totowa, NJ, 2008) str. 63–87.
- [43] K. M. Yip, N. Fischer, E. Paknia, A. Chari in H. Stark, *Atomic-Resolution Protein Structure Determination by Cryo-EM*, Nature **587**, 157 (2020).
- [44] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli in D. Hassabis, *Highly Accurate Protein Structure Prediction with AlphaFold*, Nature **596**, 583 (2021).
- [45] V. Tozzini, *Coarse-Grained Models for Proteins*, Current Opinion in Structural Biology Theory and Simulation/Macromolecular Assemblages, **15**, 144 (2005).
- [46] M. A. Blanco, H. W. Hatch, J. E. Curtis in V. K. Shen, *Evaluating the Effects of Hinge Flexibility on the Solution Structure of Antibodies at Concentrated Conditions*, Journal of Pharmaceutical Sciences **108**, 1663 (2019).
- [47] R. Dandekar in A. M. Ardekani, *Monoclonal Antibody Aggregation near Silicone Oil-Water Interfaces*, Langmuir **37**, 1386 (2021).
- [48] G. Wang, Z. Varga, J. Hofmann, I. E. Zarraga in J. W. Swan, *Structure and Relaxation in Solutions of Monoclonal Antibodies*, The Journal of Physical Chemistry B **122**, 2867 (2018).
- [49] T. M. Mitchell, *Machine Learning* (McGraw-Hill, New York, 1997).
- [50] E. A. Engel, A. Anelli, M. Ceriotti, C. J. Pickard in R. J. Needs, *Mapping Uncharted Territory in Ice from Zeolite Networks to Ice Structures*, Nature Communications **9**, 2173 (2018).
- [51] C. Barth in C. Becker, *Machine Learning Classification for Field Distributions of Photonic Modes*, Communications Physics **1**, 1 (2018).

-
- [52] S. Nakajima, H. Hoshina, M. Yamashita, C. Otani in N. Miyoshi, *Terahertz Imaging Diagnostics of Cancer Tissues with a Chemometrics Technique*, Applied Physics Letters **90**, 041102 (2007).
 - [53] K. Teknomo, *K-means clustering tutorial*, Medicine **100**, 3 (2006).
 - [54] Y. Togashi in H. Flechsig, *Coarse-Grained Protein Dynamics Studies Using Elastic Network Models*, International Journal of Molecular Sciences **19**, 3899 (2018).
 - [55] M. M. Tirion, *Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis*, Physical Review Letters **77**, 1905 (1996).
 - [56] I. Bahar, A. R. Atilgan in B. Erman, *Direct Evaluation of Thermal Fluctuations in Proteins Using a Single-Parameter Harmonic Potential*, Folding and Design **2**, 173 (1997).
 - [57] L. Yang, G. Song in R. L. Jernigan, *Protein Elastic Network Models and the Ranges of Cooperativity*, Proceedings of the National Academy of Sciences **106**, 12347 (2009).
 - [58] wwPDB consortium, *Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data*, Nucleic Acids Research **47**, D520 (2019).
 - [59] S. Hayward in B. de Groot, *Normal Modes and Essential Dynamics*, Methods in molecular biology (Clifton, N.J.) **443**, 89 (2008).
 - [60] E. Fuglebakk, S. P. Tiwari in N. Reuter, *Comparing the Intrinsic Dynamics of Multiple Protein Structures Using Elastic Network Models*, Biochimica et Biophysica Acta (BBA) - General Subjects Recent Developments of Molecular Dynamics, **1850**, 911 (2015).
 - [61] A. J. Rader, C. Chennubhotla, L.-W. Yang in a. I. Bahar, *The Gaussian Network Model: Theory and Applications*, v Normal Mode Analysis (Chapman and Hall/CRC, New York, 2005).
 - [62] Q. C. Bahar, Ivet, ur., *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems* (Chapman and Hall/CRC, New York, 2005).
 - [63] F. Tama in Y.-H. Sanejouand, *Conformational Change of Proteins Arising from Normal Mode Calculations*, Protein Engineering, Design and Selection **14**, 1 (2001).
 - [64] A. Leo-Macias, P. Lopez-Romero, D. Lupyan, D. Zerbino in A. R. Ortiz, *An Analysis of Core Deformations in Protein Superfamilies*, Biophysical Journal **88**, 1291 (2005).
 - [65] S. Kundu, D. C. Sorensen in G. N. Phillips, *Automatic Domain Decomposition of Proteins by a Gaussian Network Model*, Proteins: Structure, Function, and Bioinformatics **57**, 725 (2004).
 - [66] L. Skjaerven, S. M. Hollup in N. Reuter, *Normal Mode Analysis for Proteins*, Journal of Molecular Structure: THEOCHEM **898**, 42 (2009).

POGLAVJE 6. LITERATURA

- [67] Z. Zhang, J. Pfaendtner in G. A. Voth, *Defining Coarse-Grained Representations of Large Biomolecules and Biomolecular Complexes from Elastic Network Models*, Biophysical Journal **97**, 2327 (2009).
- [68] W. Humphrey, A. Dalke in K. Schulten, *VMD – Visual Molecular Dynamics*, Journal of Molecular Graphics **14**, 33 (1996).
- [69] X. Ni in J. Lei, *RCSB PDB - 8IZU: Crystal Structure of the n-Terminal Domain (Residues 1-137) of MPXV A7* (2024).
- [70] Z. Zhang, L. Lu, W. G. Noid, V. Krishna, J. Pfaendtner in G. A. Voth, *A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules*, Biophysical Journal **95**, 5073 (2008).
- [71] Z. Zhang, *Systematic Methods for Defining Coarse-Grained Maps in Large Biomolecules*, Advances in Experimental Medicine and Biology **827**, 33 (2015).
- [72] P. J. Flory, M. Gordon, P. J. Flory in N. . G. McCrum, *Statistical Thermodynamics of Random Networks*, Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences **351**, 351 (1997).
- [73] Z. Zhang in G. A. Voth, *Coarse-Grained Representations of Large Biomolecular Complexes from Low-Resolution Structural Data*, Journal of Chemical Theory and Computation **6**, 2990 (2010).
- [74] M. Li, J. Z. H. Zhang in F. Xia, *A New Algorithm for Construction of Coarse-Grained Sites of Large Biomolecules*, Journal of Computational Chemistry **37**, 795 (2016).
- [75] R. Xu in D. Wunsch, *Survey of clustering algorithms*, IEEE Transactions on Neural Networks **16**, 645 (2005).
- [76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot in É. Duchesnay, *Scikit-Learn: Machine Learning in Python*, Journal of Machine Learning Research **12**, 2825 (2011).
- [77] A. Bakan, L. M. Meireles in I. Bahar, *Prody: Protein Dynamics Inferred from Theory and Experiments*, Bioinformatics **27**, 1575 (2011).
- [78] S. Perovnik, *Lastni nihajni načini monoklonskih protiteles* (2024).
- [79] Y. Peng, A. J. Pak, A. E. P. Durumeric, P. G. Sahrmann, S. Mani, J. Jin, T. D. Loose, J. Beiter in G. A. Voth, *OpenMSCG: A Software Tool for Bottom-up Coarse-Graining*, The Journal of Physical Chemistry B **127**, 8537 (2023).
- [80] S. Kelow, B. Faezov, Q. Xu, M. Parker, J. Adolf-Bryfogle in R. L. Dunbrack, *A Penultimate Classification of Canonical Antibody CDR Conformations*, bioRxiv 10.1101/2022.10.12.511988 (2022).