

Sveučilište u Zagrebu  
Prirodoslovno-matematički fakultet

Domagoj Fijan  
Teorijska konformacijska analiza alifatskih spojeva i peptida

Zagreb, 2014.

Ovaj rad izrađen je u Fizikalnom zavodu Kemijskog odsjeka Prirodoslovno-matematičkog fakulteta pod vodstvom prof. dr. sc. Tomice Hrenara i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2013./2014.

## Popis korištenih kratica

Kratica	Značenje
PES	engl. <i>Potential energy surface</i> , ploha potencijalne energije
MD	Molekulska dinamika
HF	Hartree-Fock
SCF	engl. <i>Self consistent field method</i> , metoda samousklađenog polja
GGA	engl. <i>Generalized Gradient Approximation</i>
$k_b=1.3806488 \times 10^{-23} \text{ J K}^{-1}$	Boltzmannova konstanta
PCA	engl. <i>Principal Component Analysis</i> , Analiza glavnih komponenti
NIPALS	engl. <i>Non-iterative partial least squares</i> , neiterativna parcijalna metoda najmanjih kvadrata

## Sadržaj

<b>1. Uvod.....</b>	<b>1</b>
<b>1.1. Konformacijska analiza.....</b>	<b>1</b>
<b>1.2. Alifatski spojevi.....</b>	<b>1</b>
1.2.1. Etan.....	1
1.2.2. Cikloheksan.....	2
<b>1.3. Peptidi.....</b>	<b>3</b>
<b>2. Opći i specifični ciljevi.....</b>	<b>4</b>
<b>3. Teorijski uvod.....</b>	<b>5</b>
<b>3.1. Molekulska Dinamika.....</b>	<b>5</b>
3.1.1. Integriranje Velocity Verletovim algoritmom.....	5
3.1.2. Računanje sila programskim paketom Gaussian 09.....	7
3.1.3. Veza između molekulske dinamike i statističke termodinamike.....	9
<b>3.2. Analiza glavnih komponenti.....</b>	<b>10</b>
3.2.1. Algoritam za neiterativnu parcijalnu metodu najmanjih kvadrata.....	10
3.2.2. Izbor koordinata za analizu glavnih komponenti.....	11
3.2.3. Transformacija koordinata.....	11
<b>3.3. Račun raspodjele vjerojatnosti.....</b>	<b>12</b>
3.3.1. Optimizacija raspodjele vjerojatnosti.....	13
<b>4. Materijali i metode.....</b>	<b>14</b>
<b>4.1. Proučavanje konformacijskog prostora MD-om.....</b>	<b>14</b>
4.1.1. Molekulska dinamika alifatskih spojeva.....	14
4.1.2. Molekulska dinamika peptida.....	16
<b>4.2. Analiza glavnih komponenti.....</b>	<b>17</b>
4.2.1. Transformacija koordinata.....	17
4.2.2. Analiza glavnih komponenti alifatskih spojeva.....	18
4.2.3. Analiza glavnih komponenti peptida.....	18
<b>4.3. Raspodjela vjerojatnosti i optimizacija.....</b>	<b>20</b>
4.3.1. Raspodjela vjerojatnosti i optimizacija glavnih komponenti alifatskih spojeva.....	20
4.3.2. Raspodjela vjerojatnosti i optimizacija glavnih komponenti peptida.....	20
<b>4.4. Analiza konformera.....</b>	<b>21</b>
4.4.1. Analiza konformera alifatskih spojeva.....	21
4.4.2. Analiza konformera komponenti peptida.....	21
<b>5. Rezultati i rasprava.....</b>	<b>23</b>
<b>5.1. Etan.....</b>	<b>23</b>
5.1.1. Analiza u kartezijevim koordinatama.....	23
5.1.2. Analiza u internim koordinatama dobivenim funkcijom acos.....	24
<b>5.2. Cikloheksan.....</b>	<b>26</b>
5.2.1. Analiza u kartezijevim koordinatama.....	26
5.2.2. Analiza u internim koordinatama dobivenim funkcijom acos.....	27
<b>5.3. Glicin dipeptid.....</b>	<b>30</b>
5.3.1. Analiza u kartezijevim koordinatama.....	30
5.3.2. Analiza u internim koordinatama dobivenim funkcijom acos.....	33

5.3.3. Analiza u internim koordinatama dobivenim funkcijom $\text{atan}2$ .....	34
5.3.4. Analiza za sve moguće udaljenosti.....	36
<b>5.4. Alanilglicin.....</b>	<b>40</b>
5.4.1. Analiza u kartezijevim koordinatama.....	40
5.4.2. Analiza u internim koordinatama dobivenim funkcijom $\text{acos}$ .....	43
5.4.3. Analiza u internim koordinatama dobivenim funkcijom $\text{atan}2$ .....	44
5.4.4. Analiza za sve moguće udaljenosti.....	46
<b>5.5. Alanin dipeptid.....</b>	<b>48</b>
5.5.1. Analiza u kartezijevim koordinatama.....	48
5.5.2. Analiza u internim koordinatama dobivenim funkcijom $\text{acos}$ .....	50
5.5.3. Analiza u internim koordinatama dobivenim funkcijom $\text{atan}2$ .....	51
5.5.4. Analiza za sve moguće udaljenosti.....	52
<b>6. Zaključak.....</b>	<b>55</b>
<b>7. Sažetak.....</b>	<b>56</b>
<b>8. Summary.....</b>	<b>57</b>
<b>Literatura.....</b>	<b>58</b>

# 1. Uvod

## 1.1. Konformacijska analiza

Konformacijska analiza je pretraga plohe potencijalne energije (engl. *Potential energy surface*, PES) u svrhu pronalaska minimuma koje nazivamo konformeri. PES je hiperploha koja opisuje energiju molekule u ovisnosti o koordinatama jezgara. U nekom realnom sustavu molekule će poprimati razne konformacije. Najzastupljeniji će biti konformeri, i to oni nižih energija. Što je energija strukture niža to će njegova zastupljenost u nekom uzorku biti veća. Ukoliko je poznata zastupljenost većine glavnih konformera te njihova geometrija moguće je predvidjeti ponašanje sustava u određenim uvjetima. Iz geometrija dominantnih konformera neke molekule moguće je zaključiti mnogo o kemiji takve vrste - od predviđanja reaktivnosti, do računanja raznih svojstava poput dipolnog momenta ili odgovarajućih spektara.

## 1.2. Alifatski spojevi

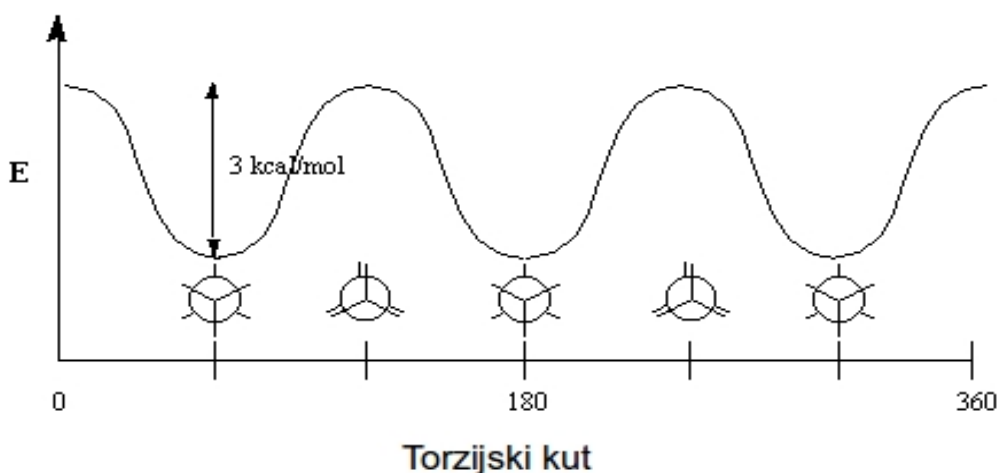
Alifatski spojevi ili alifati je skupni izraz za organske spojeve koji se sastoje od ugljika i vodika ali nisu aromatični. U ovu skupinu tako spadaju alkani i cikloalkani, dio alkena i alkina. Konformacijski prostori malih alkana poput etana, propana i butana, te cikloalkana poput cikloheksana su vrlo dobro istraženi, te su kao takvi vrlo dobre mete za testiranje metoda konformacijske analize.

### 1.2.1. Etan

Etan je alkan formule  $C_2H_6$ . Pri standardnim uvjetima i sobnoj temperaturi nalazi se u plinovitom stanju. Bezbojan je i nema mirisa, u prirodi se nalazi u prirodnom plinu iz kojeg se najčešće i dobiva industrijskim postupcima.

Etan može poslužiti kao jedan od najjednostavnijih sustava za konformacijsku analizu. Etan ima 3 kemijski ekvivalentna konformera odijeljena rotacijskom barijerom od 3 kcal mol<sup>-1</sup>. Treba naglasiti da je ova 3 konformera moguće razlikovati u računalnim simulacijama jer su svi atomi označeni. Najčešći doprinosi energetske barijerama kod rotacija su repulzije uzrokovane steričkim smetnjama između nasuprotnih skupina (u ovom slučaju vodikovih atoma). Etan je vrlo specifičan jer hiperkonjugacija ima značajan doprinos barijeri.<sup>1</sup> Konformacija kada se svi vodici preklapaju u Newmanovoj projekciji

(maksimumi na slici 1.) zove se zasjenjena, a kada su vodici maksimalno udaljeni (minimumi na slici 1.) nezasjenjena.

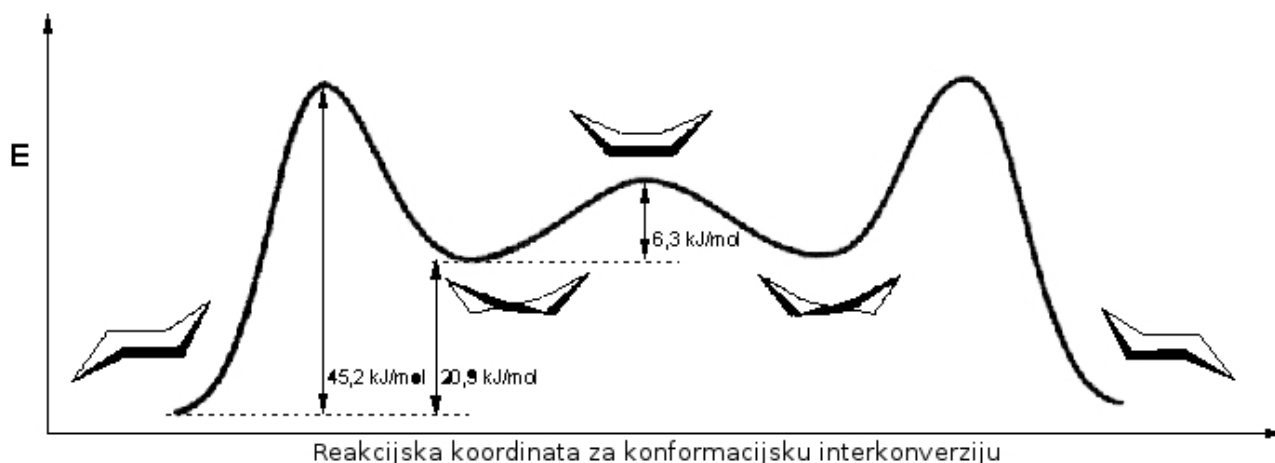


Slika 1. Konformacijski prostor etana.

### 1.2.2. Cikloheksan

Cikloheksan je najpoznatiji cikloalkan kemijske formule  $C_6H_{12}$ . Najčešće se koristi kao nepolarano otapalo u kemijskoj industriji (pri sobnoj temperaturi i standardnim uvjetima je tekućina). Konformacijski prostor cikloheksana je temeljito istražen tokom zadnjih stotinjak godina. Cikloheksan se smatra klasičnim primjerom konformacijskog izomerizma.

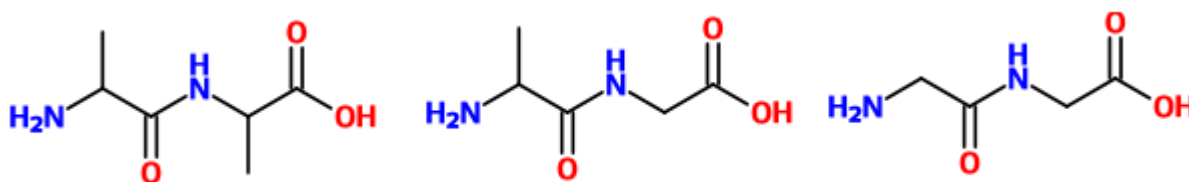
Kutevi između C-C veza odgovaraju idealnom kutu za  $sp^3$  hibridiziran ugljikov lanac  $109.5^\circ$ . Takva geometrija ima najmanje naprezanje te će molekula cikloheksana uvijek težiti takvoj geometriji. Najstabilniji konformer cikloheksana je konformacija stolca. Kod konformacije stolca svaka C-C veza je nezasjenjena. Preko polustolca cikloheksan može poprimiti geometriju lokalnog minimuma koji se naziva savijena kolijevka. Prelazak između dviju savijenih kolijevki odvija se preko konformacije kolijevke. Nova stolica se može formirati preko polustolca. Promjena iz jedne konformacije stolca u novu konformaciju stolice naziva se izvrtanje prstena ili stolca.



Slika 2. Konformacijski prostor cikloheksana.

### 1.3. Peptidi

Peptidi su kratki lanci aminokiselinskih monomera. Aminokiseline se vežu peptidnim vezama. Peptidna veza nastaje kad karboksilna skupina jedne aminokiseline reagira s amino skupinom druge aminokiseline. Razlika između peptida i proteina je isključivo u veličini odnosno broju aminokiselinskih monomera. Peptidi mogu poslužiti kao model za proučavanje proteina, odnosno strukturnih motiva proteina ali na manjoj skali.



Slika 3. Proučavani dipeptidi : alanin dipeptid, alanilglicin i glicin dipeptid.

Proučavani peptidi sastoje se od glicina i alanina. Glicin je najjednostavnija i najmanja aminokiselina, često prisutna u zavojevima proteina. Alanin je malo veći i najčešće se nalazi u  $\alpha$ -zavojnica.

Proučavanje konformacija dipeptida sastavljenih od ovih aminokiselina može dati uvid u proces savijanja proteina u nativnu konformaciju ili geometriju proteina. Savijanje proteina u nativnu konformaciju je vrlo važan proces jer proteini u ne nativnim konformacijama nisu biološki aktivni.



## 2. Opći i specifični ciljevi

Opći cilj ovoga rada je razviti i testirati metodu konformacijske analize iz molekulske dinamike uz pomoć analize glavnih komponenti i računa distribucije vjerojatnosti. U tu svrhu napravljena je analiza konformacijskog prostora nekoliko jednostavnih alifatskih spojeva kako bi se testirala metoda.

Dipeptidi predstavljaju nešto kompliciraniji, konkretniji i zahtjevniji sustav za testiranje konformacijske analize. Glavni cilj je istražiti konformacijski prostor dipeptida, odnosno odrediti karakteristike takvih konformacija koje su najdominantnije i njihovu povezanost sa sekundarnim strukturama proteina.

## 3. Teorijski uvod

### 3.1. Molekulska Dinamika

Molekulska Dinamika (MD) je metoda simulacije sustava u računalnoj fizici i kemiji. U MD simulaciji numerički se rješavaju jednadžbe gibanja klasičnog sustava od  $N$  čestica u diskretnim vremenskim intervalima. Klasični sustav je ovdje definiran Newtonovim jednadžbama gibanja kojima se opisuje gibanje čestica. MD simulacija, kao rezultat, generira trajektoriju. Trajektorija je funkcija koja opisuje položaj čestica sustava u ovisnosti o vremenu. Analizom trajektorije moguće je dobiti uvid u mnoge dinamičke karakteristike simuliranog sustava. Međutim, ako je trajektorija dovoljno duga moguće je dobiti i mnoge termodinamičke podatke.

Ključni korak u generiranju trajektorije je integracija jednadžbi gibanja i računanje sila. Svaka simulacija počinje sa početno zadanom konfiguracijom sustava - početnim koordinatama čestica i pripadnim brzinama. Svaki sljedeći korak, čestice u sustavu su pomaknute ovisno o njihovoj brzini i sili koja djeluje na njih. Jednadžbe su integrirane Velocity Verletovim algoritmom dok je za računanje sila korišten programski paket Gaussian 09.<sup>2</sup> Računanje sila je uvijek najsporiji korak svake MD simulacije, a odabir modela kojim se sile računaju često diktira fizikalnost simulacije. Stoga je odabir metode i modela za računanje sila vrlo bitan dio svake MD simulacije.

#### 3.1.1. Integriranje Velocity Verletovim algoritmom

Integriranjem Newtonovih jednadžbi gibanja dobivamo nove geometrije na trajektoriji. Postoje različiti algoritmi kojima se provodi integracija u MD-u. Velocity Verlet je najčešće korišten algoritam u ovu svrhu. Njegova glavna prednost je jednostavnost implementacije ali i činjenica da za duge trajektorije daje vrlo dobro očuvanje ukupne energije.

Velocity Verletov algoritam temelji se na Verletovom algoritmu.<sup>3</sup> Izvod jednadžbi za Verletov algoritam započinje Taylorovim razvojem u red koordinate čestice oko vremena  $t$

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 + \frac{\ddot{r}}{3!}\Delta t^3 + O(\Delta t^4) \quad . \quad (3.1)$$

Analogno može se zapisati izraz ako gledamo unazad

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 - \frac{\ddot{r}}{3!}\Delta t^3 + O(\Delta t^4) . \quad (3.2)$$

Sumiranjem ove dvije jednačbe dobiva se

$$r(t+\Delta t) + r(t-\Delta t) = 2r(t) + \frac{f(t)}{m}\Delta t^2 + O(\Delta t^4) . \quad (3.3)$$

Sređivanjem sada moguće je dobiti izraz koji nam govori kako koordinata čestice u vremenu ovisi o prethodnom položaju

$$r(t+\Delta t) \approx 2r(t) - r(t-\Delta t) + \frac{f(t)}{m}\Delta t^2 . \quad (3.4)$$

Brzina se može izračunati kao

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\Delta t} . \quad (3.5)$$

Velocity Verlet<sup>4</sup> algoritam uključuje brzine u izraz za položaj čestica

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{f(t)\Delta t^2}{2m} , \quad (3.6)$$

dok sam izraz za brzinu postaje

$$v(t+\Delta t) = v(t) + \frac{f(t) + f(t+\Delta t)}{2m}\Delta t . \quad (3.7)$$

Standardna shema implementacije Velocity Verletovog algoritma je sljedeća :

1. Računanje položaja prema formuli 3.6
2. Računanje brzina prema formuli  $v(t+\frac{\Delta t}{2}) = v(t) + \frac{f(t)}{2m}\Delta t$
3. Računanje sila  $f(t+\Delta t)$  iz novo izračunatih položaja čestica
4. Računanje brzine prema formuli  $v(t+\Delta t) = v(t+\frac{\Delta t}{2}) + \frac{f(t+\Delta t)}{2m}\Delta t$  .

Zakruživanje Taylorovog razvoja u izvodu jednačbi za integraciju Velocity Verletovim algoritmom će prouzročiti dugoročnu akumulaciju greške prilikom integracije. Nakon dovoljno dugo koraka očuvanje ukupne energije će biti izgubljeno, iako će ona lokalno na

nekom manjem segmentu biti očuvana. Akumulacija greške prilikom integracije će prouzročiti eksponencijalno odstupanje računate trajektorije od one realne. Ovaj efekt poznat pod imenom Lyapunova nestabilnost djeluje kao veliki problem međutim njegov utjecaj na termodinamičke podatke koje možemo izvući iz trajektorije je zanemariv. Iako ova činjenica nije još dokazana, vjeruje se da je istinita, a u prilog joj idu i mnogi numerički primjeri. Tako se MD trajektorije najčešće povezuju s takozvanim sjena-orbitama. Sjena-orbite su realne trajektorije koje tijesno slijede numeričke trajektorije za vrijeme koje je dovoljno dugo da u usporedbi s vremenom potrebnim da se razvije Lyapunova nestabilnost.

### 3.1.2. Računanje sila programskim paketom Gaussian 09

Gaussian 09<sup>2</sup> je kvantno kemijski program za računanje elektronske strukture. Podržava korištenje različitih *ab-initio* te semiempirijskih metoda. Glavna prednost *ab-initio* metoda je njihova točnost u rješavanju elektronske Schrödingerove jednadžbe. Semiempirijske metode koriste isti pristup ali zanemaruju ili aproksimiraju dio dvoelektronskih integrala u Hartree-Fockovoj teoriji empirijskim vrijednostima. Ove metode su brže od *ab-initio* metoda jer ne troše vrijeme na računanje dvoelektronskih integrala što rezultira smanjenoj točnosti proračuna. Kako bi se povećala točnost ovih metoda one se često parametriziraju na temelju eksperimentalnih ili vrlo točnih teorijskih podataka kako bi davale što bolje slaganje. Pošto je za pretraživanje većine dostupnog konformacijskog prostora potrebna relativno duga trajektorija korištenje semiempirijske metode predstavlja optimalan izbor za metodu kojom se računa sila za MD.

Za računanje sile potrebno je izračunati gradijent energije

$$F = -\nabla V \quad (3.8)$$

Gradijenti za semiempirijske metode u Gaussianu 09 su analitički. Za računanje energija semiempirijskim metodama koristi se Hartree-Fock(HF) pristup. HF teorija kreće od Born-Oppenheimerove aproksimacije. Ona podrazumijeva zanemarivanje gibanja jezgri sprem elektrona pod izlikom da se jezgre gibaju mnogo sporije od elektrona. To nam omogućuje odvajanje odvajanje elektronske od nuklearne valne funkcije. Tada ukupnu je ukupna valna funkcija

$$\Psi_{\text{tot}} = \psi_{\text{jezgara}} \psi_{\text{elektronska}} \quad (3.9)$$

Ova aproksimacija je vrlo dobra za osnovno stanje i često je u potpunosti opravdana. Pošto je od interesa isključivo osnovno stanje pretpostavljeno je da ova aproksimacija uvijek vrijedi. Relativistički efekti se također zanemaruju što je opravdano činjenicom da se ne istražuju spojevi koji imaju atome teže od kisika. Kod takvih atoma brzine elektrona su daleko ispod brzine svjetlosti. U HF teoriji elektronska energija je opisana Fockovim operatorom

$$\hat{F}(i) = \hat{h}(i) + \sum_{j=1}^{n/2} [2\hat{J}_j(i) - \hat{K}_j(i)] \quad (3.10)$$

Fockov operator sastoji se od jednoelektronskog dijela (takozvanog *core* hamiltonijana  $\hat{h}(i)$ ) i dvoelektronskog dijela. Dvoelektronski dio sastoji se od Coulombovog operatora  $\hat{J}_j(i)$  koji opisuje odbijanje  $i$ -tog i  $j$ -tog elektrona, te operatora izmjene  $\hat{K}_j(i)$  koji opisuje kvantni efekt izmjene 2 elektrona. Fockov operator sam po sebi mora biti definiran za neki zadani skup orbitala te je stoga nelinearni operator. Rješavanjem problema svojstvenih vrijednosti za Fockov operator

$$\hat{F}(i)|j\rangle = \sum_{i=1}^n \epsilon_{ij}|j\rangle \quad (3.11)$$

dobiva se novi set orbitala  $|j\rangle$  i vlastitih vrijednosti odnosno energija  $\epsilon$ . Problem vlastitih vrijednosti se rješava dijagonalizacijom Fockovog operatora. Fockov operator je potrebno transformirati nekim drugim unitarnim operatorom

$$\hat{f} = U^T \hat{F} U \quad (3.12)$$

$$E = U^T \epsilon U \quad (3.13)$$

$$\hat{f}|j\rangle = E_j|j\rangle \quad (3.14)$$

Ovaj postupak mora se ponoviti dok se ne postigne odgovarajuća konvergencija. Ova procedura poznata je pod imenom self-consistent Field method (SCF). Semiempirijske metode ne računaju dvoelektronske integrale nego ih ili zanemaruju ili ih zamjenjuju nekim empirijskim vrijednostima. Ovo uvelike smanjuje vrijeme računanja ali i smanjuje točnost proračuna. Za potrebe ovog rada velika točnost u energiji nije ključna. Bitno je samo da su relativne razlike u energijama proporcionalne pravim vrijednostima.

### 3.1.3. Veza između molekulske dinamike i statističke termodinamike

Glavna pretpostavka kod računanja termodinamičkih podataka iz MD-a je ergodička hipoteza,<sup>5</sup> temelj statističke termodinamike. Ona govori da je prosjek trajektorije u vremenu ekvivalentan prosjeku po statističkom ansamblu. Ansambl je skup virtualnih kopija nekog sustava od kojih svaki predstavlja neku realnu konfiguraciju. Velocity Verletov algoritam simulira mikrokanonski ansambl gdje su konstantni broj čestica, volumen i ukupna energija. Ravnotežno stanje u mikrokanonskom ansamblu je definirano kada je entropija ekstremalna (drugi zakon termodinamike). Dokaz da je entropija termodinamički potencijal se da lako izvesti iz diferencijala ukupne energije

$$dE = TdS - pdV + \mu dN \quad (3.15)$$

$$dS = \frac{1}{T} dE + \frac{p}{T} dV - \frac{\mu}{T} dN \quad (3.16)$$

Tada diferencijal entropije možemo zapisati kao. Iz ovoga se vidi da je u uvjetima mikrokanonskog ansambla ( $N, V, E$  konstantni) entropija ekstremalna, odnosno ona je odgovarajući termodinamički potencijal. Entropiju možemo definirati prema Boltzmannovoj formuli koja povezuje volumen faznog prostora s entropijom

$$S = k_b \ln W \quad (3.17)$$

Ako sustav teži maksimumu entropije onda će prema Boltzmannu težiti maksimizaciji volumena dostupnog faznog prostora. Volumen faznog prostora je definiran brojem načina na koji možemo rasporediti koordinate i količine gibanja (brzine) za zadane uvjete

$$W = \sum_i w_i, w_i = \begin{cases} \text{const.}, & E + \Delta E < E_i < E \\ 0 & \end{cases} \quad (3.18)$$

Zadanoj potencijalnoj energiji će najčešće odgovarati jedna geometrija odnosno jedan set koordinata. Kinetičku energiju odnosno količine gibanja je moguće rasporediti na mnogo više načina. Tako primjerice ukupnu kinetičku energiju možemo pohraniti u jedan atom molekule ili ju pak ravnomjerno rasporediti svim atomima. Sada je jasno da će maksimizacija kinetičke energije uvjetovati maksimizaciju volumena faznog prostora te time i entropije. Pošto je ukupna energija konstantna maksimizacija kinetičke energije će za posljedicu imati minimizaciju potencijalne energije. Iako je distribucija konfiguracija u mikrokanonskom ansamblu uniformna, mnogo češće će se pojavljivati one konfiguracije

gdje je potencijalna energija minimizirana, jer u tom slučaju je moguće napraviti mnogo više konfiguracija raspodjelama količina gibanja. Što je veća kinetička energija dostupna ukupna količina gibanja će biti veća te će biti moguće napraviti veći broj konfiguracija što će rezultirati povećanjem volumena faznog prostora odnosno entropije. Ako znamo da minimumi na plohi potencijalne energije koja je funkcija energije u prostoru koordinata čestica molekule odgovaraju konformerima, onda će sustav u ravnoteži preferirati konformere. Konformeri niže energije će se češće pojavljivati u trajektoriji jer imaju manju potencijalnu energiju.

### 3.2. Analiza glavnih komponenti

Analiza glavnih komponenti<sup>6</sup> (engl. *Principal Component Analysis*, PCA) je metoda multivarijatne analize podataka. Glavna ideja je promijeniti set moguće koreliranih mjerenja u set varijabli koji je linearno nezavisan. Te nove varijable zovu se glavne komponente i njihov smjer je određen maksimalnom varijancijom podatak. U matematičkim terminima efektivno cilj je napisati matricu  $X$  ranga  $r$  kao sumu  $r$  matrica ranga 1

$$X = M_1 + M_2 + M_3 + \dots + M_r \quad (3.19)$$

Svaka od matrica  $M$  može se zapisati kao produkt

$$M_i = s_i l_i^T \quad (3.20)$$

gdje je  $s_i$  vektor faktorskih bodova (engl. *score*), a  $l_i^T$  je transponirani vektor faktorskih opterećenja koji sadrži projekcije svih početnih jediničnih vektora kojima su opisani početni podatci na  $i$ -tu glavnu komponentu (engl. *loading*). Faktorski bodovi predstavljaju projekciju svakog podatka iz početne matrice na  $i$ -tu glavnu komponentu. Glavni cilj PCA je odrediti glavne komponente i faktorske bodove. PCA može poslužiti kao metoda za redukciju dimenzionalnosti jer je često moguće opisati cijeli skup podataka sa nekoliko najvažnijih glavnih komponenti. Ovaj pristup je već korišten za analizu proteina.<sup>7</sup>

#### 3.2.1. Algoritam za neiterativnu parcijalnu metodu najmanjih kvadrata

Algoritam za neiterativnu parcijalnu metodu najmanjih kvadrata<sup>8</sup> (engl. *Non-iterative partial least squares*, NIPALS) se često koristi za skupove visoko dimenzionalnih podataka u računanju PCA. Glavna prednost ove metode je brzina u računanju prvih par najvažnijih

glavnih komponenti. NIPALS ne računa sve glavne komponente odjednom, već ih računa pojedinačno jednu po jednu.

1. Uzme se vektor  $x_j$  iz  $X$  i proglašuje se  $s_h = x_j$
2. Izračuna se  $l_h^T$  prema  $l_h^T = s_h^T X / s_h^T s_h$
3.  $l_h$  se normalizira  $l_{h, \text{norm}} = l_h / |l_h|$
4. ponovno se izračuna novi  $s_h$   $s_h = X l_{h, \text{norm}} / l_{h, \text{norm}}^T l_{h, \text{norm}}$
5. usporedbom  $s_h$  iz koraka 2 i koraka 4 odredi se je li postignut zadovoljavajući konvergencijski kriterij. U slučaju da je konvergencijski kriterij postignut,  $X$  se zamjenjuje s  $E_1 = X - s_1 l_1^T$ , odnosno s  $E_n = E_{n-1} - s_n l_n^T$  za  $n$ -tu glavnu komponentu. U suprotnom ponovno se treba vratiti na korak 2.

Alternativno PCA je moguće provesti određivanjem svojstvenih vrijednosti nekom drugom metodom poput dekompozicije singularnih vrijednosti. Takve metode uključuju računanje matrice kovarijancije što za posljedicu ima dulje vrijeme računanja i nije moguće odrediti glavne komponente jednu po jednu, već ih se sve računa odjednom.

### 3.2.2. Izbor koordinata za analizu glavnih komponenti

Uobičajeno je za MD simulacije koristiti kartezijeve koordinate. Osim u kartezijevim koordinatama, geometriju molekule moguće je prikazati i u mnogim drugim koordinatama. U kemiji su jako popularne interne koordinate. Glavna prednost internih koordinata je eliminacija često neželjenih stupnjeve slobode poput rotacije i translacije centra mase. Također je opis molekule mnogo intuitivniji jer je geometrija opisana međusobnim udaljenostima atoma te kutevima između veza i torzijama. Izbor koordinata imat će velik utjecaj na analizu glavnih komponenti, jer će se varijancija vezana za koordinatu znatno promijeniti.

### 3.2.3. Transformacija koordinata

Interne koordinate definirane su s  $N-1$  duljina veza,  $N-2$  kuteva između veza te  $N-3$  torzija, gdje je  $N$  broj atoma u molekuli. Udaljenost između 2 atoma definirana je kao

$$r_{ab} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2} \quad (3.21)$$



Kutevi među vezama definirani su koordinatama 3 atoma. Neka je  $\vec{v}_1$  vektor koji gleda od atoma 2 prema atomu 1 duž njihove veze

$$\vec{v}_1 = \vec{r}_b - \vec{r}_a \quad (3.22)$$

Neka je  $\vec{v}_2$  vektor koji gleda od atoma 2 prema atomu 3 duž njihove veze

$$\vec{v}_2 = \vec{r}_c - \vec{r}_b \quad (3.23)$$

Tada možemo definirati kut između vektora  $\vec{v}_1$  i  $\vec{v}_2$  kao

$$\varphi = \arccos\left(\frac{\langle \vec{v}_1 | \vec{v}_2 \rangle}{|\vec{v}_1| |\vec{v}_2|}\right) \quad (3.24)$$

Kako bi definirali torziju potrebna su 4 uzastopna atoma povezana vezama. Neka je  $\vec{v}_1$  vektor koji gleda od atoma 1 prema atomu 2 duž njihove veze,  $\vec{v}_2$  vektor koji gleda od atoma 2 prema atomu 3 duž njihove veze, a  $\vec{v}_3$  vektor koji gleda od atoma 4 prema atomu 3 duž njihove veze

$$\vec{v}_1 = \vec{r}_a - \vec{r}_b, \vec{v}_2 = \vec{r}_b - \vec{r}_c, \vec{v}_3 = \vec{r}_d - \vec{r}_c \quad (3.25)$$

Normala na ravninu definiranu vektorima  $\vec{v}_1$  i  $\vec{v}_2$  je definirana kao vektorski produkt  $\vec{n}_1 = \vec{v}_1 \times \vec{v}_2$ . Analogno je normala ravnine definirane vektorima  $\vec{v}_3$  i  $\vec{v}_2$  definirana s  $\vec{n}_2 = \vec{v}_3 \times \vec{v}_2$ . Torziju se sad može izračunati kao kut između dvije normale

$$\theta = \arccos\left(\frac{\langle \vec{n}_1 | \vec{n}_2 \rangle}{|\vec{n}_1| |\vec{n}_2|}\right) \quad (3.26)$$

### 3.3. Račun raspodjele vjerojatnosti

Vjerojatnost pojavljivanja neke geometrije je direktno povezana s njenom potencijalnom energijom. Maksimumi u funkciji gustoće vjerojatnosti pojavljivanja neke geometrije predstavljati će minimume na PES. Funkciju gustoće vjerojatnosti pojavljivanja neke geometrije može se dobiti tako da se zbroje različite konfiguracije u faznom prostoru koje odgovaraju nekoj geometriji.

Nakon PCA dimenzionalnost podataka se može znatno smanjiti. Tako umjesto pretraživanja cijelog konformacijskog prostora (dimenzija  $3N$ , gdje je  $N$  broj atoma) pretraživat će se samo prostor nekoliko glavnih  $k$  komponenti koji opisuju većinu

varijancije. Svaka koordinata, odnosno glavnu komponentu, se raspodijeli na  $ndiv$  intervala, te se potom potraže sve geometrije čiji faktorski bodovi ulaze unutar promatranog  $k$  dimenzionalnog paralelepipeda. Rasponi faktorskih bodova variraju po glavnim komponentama. Korištenjem hipersfere glavne komponente nižeg udjela varijancije ne bi imale doprinosa nakon što radijus prijeđe njihovu maksimalnu vrijednost. Zato je korištenje paralelepipeda bolji odabir.

### 3.3.1. Optimizacija raspodjele vjerojatnosti

Optimizacija podrazumijeva pronalaženje ekstrema neke funkcije. Za optimizaciju gustoće raspodjele vjerojatnosti korišten je jednostavan *brute-force* algoritam. Za svaku je točku potrebno provjeriti je li ona maksimum u svakoj dimenziji. Neka točka je maksimum ako je točka prije i točka poslije manja od promatrane točke. Ukoliko je ovaj uvjet zadovoljen u svim dimenzijama onda je promatrana točka maksimum.

## 4. Materijali i metode

Svi proračuni provedeni su na računalima opremljenim AMD Opteron 852 procesorima snage 2066MHz i openSUSE 12.3 (Dartmouth) (x86\_64) operacijskim sustavom.

### 4.1. Proučavanje konformacijskog prostora MD-om

U mikrokanonskom ansamblu dio konformacijskog prostora bit će ograničen samo danom početnom ukupnom energijom. Svaka MD simulacija započinje iz neke početno zadane geometrije i zadane kinetičke energije u obliku temperature. Time će biti određeno kolika je energija dostupna molekuli za savladavanje barijera pri promjeni konformacije. Pri promjeni konformacije ključno je da su kinetičke energije atoma, čije promjene koordinata najviše doprinose reakcijskoj koordinati za promatranu konformacijsku tranziciju, su dovoljno velike za savladavanje barijere potencijalne energije prijelaznog stanja.

Veza između temperature i kinetičke energije dana je ekviparticijskim teoremom

$$E = \frac{k_b T}{2} \quad (4.1)$$

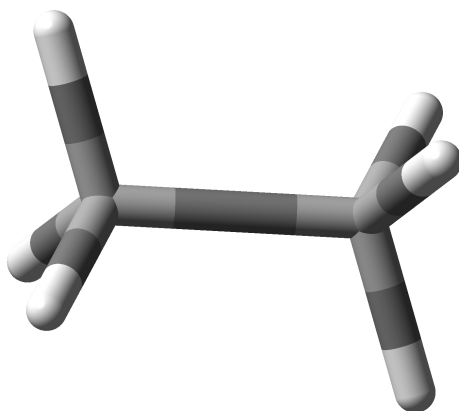
Svaka varijabla o kojoj energija ovisi kvadratično daje doprinos od  $k_b T/2$  ukupnoj energiji. Svaki atom u sustavu ima 3 stupnja slobode za 3 smjera brzine. Onda je ukupna kinetička energija zadana kao  $3NE$ . Kinetičke energije se nasumično raspodjele među stupnjevima slobode tako da ukupna energija ostane očuvana. Iz dane energije se lako odrede brzine koje su vezane uz danu energiju.

Računi sila rađeni su Gaussian 09<sup>2</sup> programskim paketom. Korištena je semiempirijska metoda PM6.<sup>9</sup> Za atome su korišteni najzastupljeniji izotopi : <sup>1</sup>H, <sup>12</sup>C, <sup>14</sup>N i <sup>16</sup>O. U svakom koraku integracije rotacija i translacija centra mase su eliminirane. Kvantni efekti vezani za vibracijske stupnjeve slobode su zanemarivi jer su temperature odabrane tako da se sustav nalazi daleko od kvantne granice. Sve MD simulacije su računane u kartezijevim koordinatama. Vremenski korak iznosio je 2 fs za sve MD simulacije.

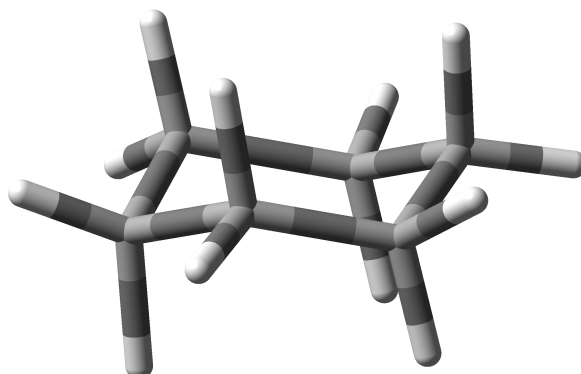
#### 4.1.1. Molekulska dinamika alifatskih spojeva

Molekulska dinamika za etan trajala je 330000 koraka. Početna geometrija etana pripada D<sub>3d</sub> simetrijskoj grupi i globalni je minimum. Početna temperatura za etan je iznosila 2273.15K.

Molekulska dinamika za cikloheksan trajala je 176000 koraka. Početna geometrija cikloheksana je globalni minimum dan na slici. Početna temperatura za cikloheksan iznosila je 1273.15K.

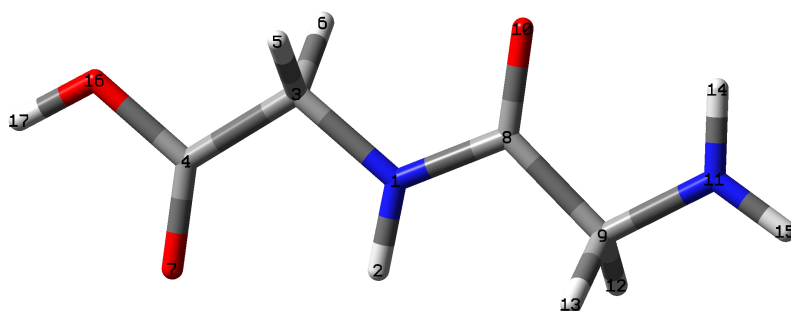


Slika 4. Početna geometrija etana.

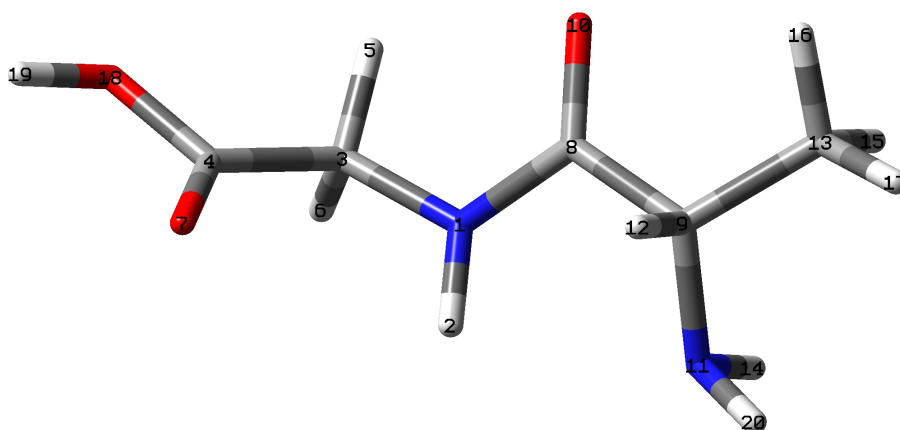


Slika 5. Početna geometrija cikloheksana.

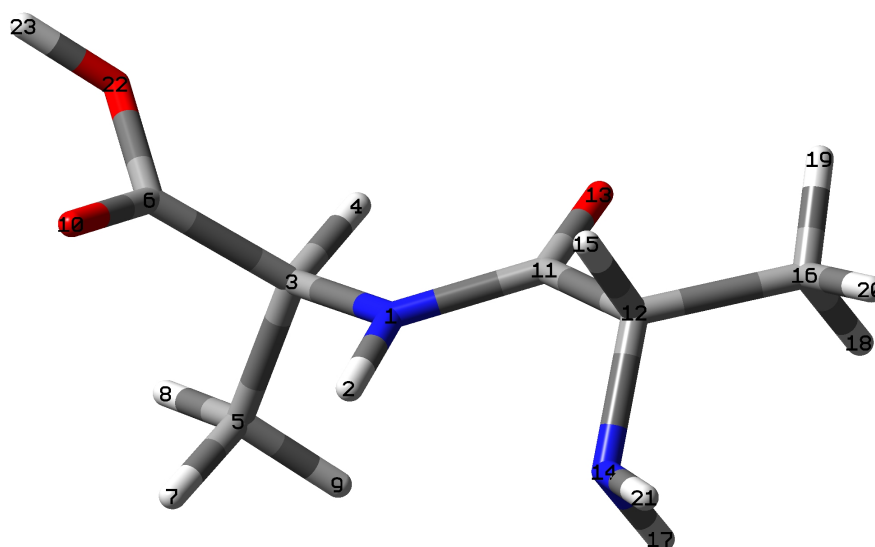
## 4.1.2. Molekulska dinamika peptida



Slika 6. Početna geometrija glicin dipeptida.



Slika 7. Početna geometrija alanilglicina.



Slika 8. Početna geometrija alanin dipeptida.

Početna temperatura za sva 3 dipeptida iznosi 1273.15 K. Trajektorija glicin dipeptida je iznosila 520000 koraka, dok su trajektorije glicilalanina i alanin dipeptida iznosile 1010000 koraka. Početne geometrije dipeptida dane su na slikama.

## 4.2. Analiza glavnih komponenti

Za svaku trajektoriju napravljena je analiza glavnih komponenti u kartezijevim koordinatama i izračunato je prvih 10 glavnih komponenti NIPALS algoritmom. Maksimalan broj iteracija za NIPALS je iznosio 10000 ukoliko prethodno nije postignuta konvergencija od 0.000000000001. Trajektorije su također transformirane iz kartezijevih u interne koordinate te je i na njima napravljena analiza glavnih komponenti.

### 4.2.1. Transformacija koordinata

Duljine veza računane su po formuli 3.21. Za račun torzija i kuteva među vezama osim funkcije  $\text{acos}$  korištena je i  $\text{atan2}$  funkcija u Fortranu 90. Razlog tomu je to što u slučaju torzija  $\text{atan2}$  daje rezultate od  $-\pi$  do  $\pi$ , dok je  $\text{arccos}$  definiran samo na intervalu od 0 do  $\pi$ . Također  $\text{acos}$  je loše numerički definiran pri kutevima jako blizu 0. Stoga je funkcija  $\text{atan2}$  preferirana, iako će i normalni  $\text{acos}$  biti korišten. Kako bi se izračunali kutevi pomoću funkcije  $\text{atan2}$  potrebno je modificirati formule 3.24 i 3.26. Funkcija  $\text{atan2}$  ima 2 argumenta što joj omogućuje određivanje kvadranta u kojem se izračunati kut nalazi. Za neki par realnih brojeva  $x, y$  različitih od 0  $\text{atan2}(x, y)$  je kut u radijanima između pozitivnog dijela  $x$  osi ravnine i točke  $(x, y)$  na toj ravnini. Kut među vezama se  $\text{atan2}$  funkcijom računa

$$\varphi = \text{atan2}(|\vec{v}_1 \times \vec{v}_2|, \langle \vec{v}_1 | \vec{v}_2 \rangle) \quad (4.2)$$

Za izračun torzije potrebno je definirati dodatan vektor  $\vec{m}_1$  koji je dan izrazom

$$\vec{m}_1 = \vec{n}_1 \times \vec{v}_2 \quad (4.3)$$

Izraz za torzijski kut je onda

$$\theta = \text{atan2}\left(\frac{\langle \vec{m}_1 | \vec{n}_2 \rangle}{|\vec{m}_1 \cdot \vec{n}_2|}, \frac{\langle \vec{n}_1 | \vec{n}_2 \rangle}{|\vec{n}_1 \cdot \vec{n}_2|}\right) \quad (4.4)$$

Za uspješnost metode potrebno je imati takav skup ulaznih podataka za PCA koji će jednoznačno opisivati molekulu. Pritom nije bitno kako i koje su torzije kutevi i razmaci između atoma definirani niti dali između tih atoma postoji veza, odnosno dali torzije su za računanje torzija ravnine definirane vektorima kroz koje prolazi veza ili ne (primjerice kod

alanin dipeptida ili alanil glicina). Moguće je dobiti i solidne rezultate ako se u analizu uključe samo glavni atomi kostura. Međutim minimalni potpuni opis molekule zahtijeva  $3N-6$  koordinata i za njega je moguće opisati potpunu konformaciju molekule.

#### 4.2.2. Analiza glavnih komponenti alifatskih spojeva

Za etan i cikloheksan provedena je analiza glavnih komponenti u kartezijevim koordinatama i internim koordinatama dobivenim funkcijom  $\text{acos}$ . Interne koordinate dobivene su računanjem svih mogućih duljina veza, kuteva između veza, te svim mogućim torzijama definiranim vezama.

#### 4.2.3. Analiza glavnih komponenti peptida

Analiza glavnih komponenti provedena je za kartezijeve koordinate, interne koordinate dobivene pomoću funkcija  $\text{acos}$  i  $\text{atan2}$ , te za sve moguće udaljenosti između atoma. Korišteni skupovi internih koordinata dani su u tablici za svaki dipeptid. Broj mogućih

udaljenosti za molekulu s  $N$  atoma iznosi  $\binom{N}{2}$ .

Tablica 1. Veze, kutevi i torzije korišteni za opis peptida u internim koordinatama, svaki broj predstavlja redni broj atoma

Glicin dipeptid		Alanilglicin		Alanin dipeptid	
Veze					
2	1	2	1	2	1
3	1	3	2	3	2
4	3	4	2	4	3
5	3	5	4	5	2
6	3	6	4	6	5
7	4	7	6	7	6
8	1	8	7	8	3
9	8	9	8	9	6
10	8	10	8	10	9
11	9	11	6	11	9
12	9	12	7	12	5
13	9	13	7	13	2
14	11	14	2	14	1

15	11
16	4
17	16

15	3
16	1
17	1
18	1
19	9
20	3

15	1
16	1
17	9
18	11
19	10
20	10
21	10
22	11
23	4

## Kutevi

3	1	2
4	3	1
5	3	1
6	3	1
7	4	3
8	1	3
9	8	1
10	8	1
11	9	8
12	9	8
13	9	8
14	11	9
15	11	9
16	4	3
17	16	4
8	1	2

3	2	1
4	2	1
5	4	2
6	4	2
7	6	4
8	7	6
9	8	7
10	8	7
11	6	7
12	7	6
13	7	6
14	2	4
15	3	2
16	1	2
17	1	2
18	1	2
19	9	8
20	3	2

3	2	1
4	3	2
5	2	3
6	5	2
7	6	5
8	3	2
9	6	5
10	9	6
11	9	6
12	5	2
13	2	5
14	1	2
15	1	2
16	1	2
17	9	6
18	11	9
19	10	9
20	10	9
21	10	9
22	11	9
23	4	3

## Torzije

4	3	1	2
5	3	1	4

4	2	1	3
5	4	2	1

4	3	2	1
5	2	3	4



6	3	1	4	6	4	2	3	6	5	2	3
7	4	3	1	7	6	4	2	7	6	5	2
8	1	3	4	8	7	6	4	8	3	2	5
9	8	1	3	9	8	7	6	9	6	5	2
10	8	1	3	10	8	7	6	10	9	6	5
11	9	8	1	11	6	7	8	11	9	6	5
12	9	8	1	12	7	6	4	12	5	2	1
13	9	8	1	13	7	6	4	13	2	5	6
14	11	9	8	14	2	4	6	14	1	2	5
15	11	9	8	15	3	2	4	15	1	2	5
16	4	3	1	16	1	2	4	16	1	2	5
17	16	4	3	17	1	2	4	17	9	6	5
11	9	8	10	18	1	2	4	18	11	9	6
10	8	1	2	19	9	8	7	19	10	9	6
				20	3	2	4	20	10	9	6
								21	10	9	6
								22	11	9	6
								23	4	3	2

### 4.3. Raspodjela vjerojatnosti i optimizacija

Prostor faktorskih bodova svake glavne komponente podijeljen je na 20 jednakih dijelova ( $n_{div}=20$ ) osim za sve račune kod cikloheksana gdje je prostor faktorskih bodova glavnih komponenti bio podijeljen na 10 dijelova ( $n_{div}=10$ ).

#### 4.3.1. Raspodjela vjerojatnosti i optimizacija glavnih komponenti alifatskih spojeva

Za etan u kartezijevim koordinatama napravljena je distribucija vjerojatnosti za prvih 5 glavnih komponenti, dok je za interne koordinate bilo dovoljno prve 2 glavne komponente.

Kod cikloheksana korišteno je 2 glavne komponente za kartezijeve i 3 za interne koordinate.

#### 4.3.2. Raspodjela vjerojatnosti i optimizacija glavnih komponenti peptida

Za račun raspodjele vjerojatnosti glicin dipeptida u kartezijevim koordinatama analizirani su faktorski bodovi prvih 5 glavnih komponenti, dok je kod internih koordinata korišteno

prvih 6 glavnih komponenti. Za račun svih udaljenosti korištene su prve 3 glavne komponente za glicin dipeptid.

Kod alanil glicina i alanin dipeptida korišteno je prvih 5 glavnih komponenti u analizi internim i kartezijskim koordinatama, te prve 3 glavne komponente za sve udaljenosti.

## 4.4. Analiza konformera

### 4.4.1. Analiza konformera alifatskih spojeva

Za etan analizirane su geometrije prvih 5 najvećih maksimuma za kartezijske i interne koordinate. Za kartezijske koordinate izračunata je prosječna geometrija dok je za interne koordinate uzet reprezentativni uzorak.

Kod cikloheksana za kartezijske koordinate uzet je prosjek prvih 5 najvećih maksimuma, dok je su za prva 4 maksimuma za interne koordinate uzeti reprezentativni uzorci.

### 4.4.2. Analiza konformera komponenti peptida

Za skup geometrija koje su sadržane u promatranim maksimumima dobivenih analizom iz kartezijskih koordinata izračunate su prosječne geometrije. Za maksimume dobivene iz internih koordinata i svih udaljenosti izabrani su reprezentativni uzorci. Geometrije iz maksimuma za sve udaljenosti su izračunate koordinate u dijagramu sličnom Ramachandrovom dijagramu.<sup>10</sup> Za pravi Ramachandrov dijagram bilo bi potrebna još jedna aminokiselina. Pošto se dipeptid sastoji od 2 glicina s karboksilnom terminalnom i amino terminalnom grupom nije moguće napraviti pravi Ramachandrov dijagram. Ipak ako se iz amino terminalnog dijela uzme torziju  $\psi$  (N-C<sub>α</sub>-C-N) opisanu sa atomima 1 8 9 11, te iz karboksilno terminalnog dijela uzme torzija  $\phi$  (C-N-C<sub>α</sub>-C) opisana atomima 8 1 3 4 za glicin dipeptid, rezultati bi trebali biti približno slični onima iz literature.

Za alanilglicin u kartezijskim koordinatama prikazani su prosjeci prva 4 maksimuma, dok su za interne koordinate i sve moguće udaljenosti dani reprezentativni uzorci. Pošto alanil glicin ima 2 različita rezidualne grupe usporedba Ramachandrovim dijagramima iz literature nema smisla.

Za alanin dipeptid u kartezijskim koordinatama dani su prosjeci 4 najveća maksimuma, dok su za interne koordinate dani reprezentativni uzorci. Za analizu svih udaljenosti napravljen je Ramachandrov dijagram. Kut  $\psi$  je definiran 14 12 11 1, dok je  $\phi$  definiran

atomima 11 1 3 6. Također su razmatrana samo 3 najveća maksimuma a rezultati su uspoređeni s literaturom.

## 5. Rezultati i rasprava

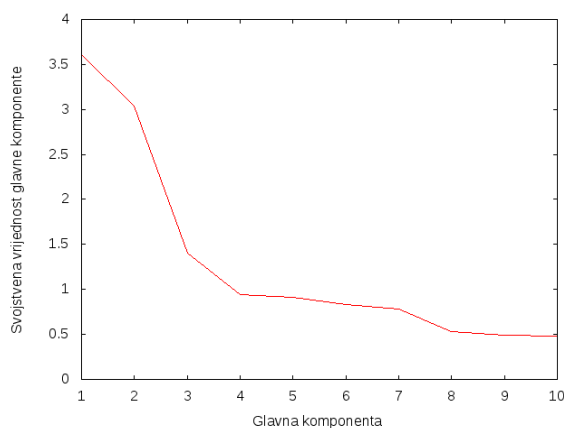
### 5.1. Etan

#### 5.1.1. Analiza u kartezijevim koordinatama

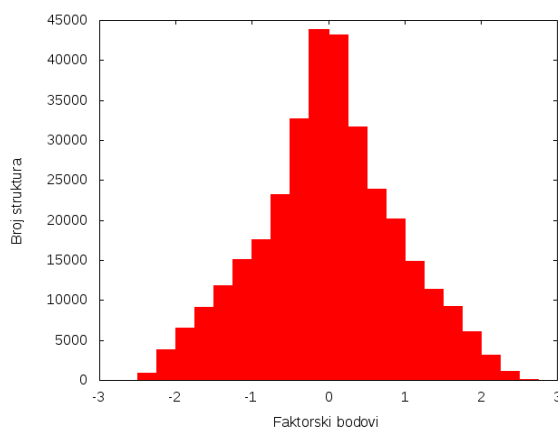
Tablica 2. Rezultati PCA za etan u kartezijevim koordinatama

Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojstvena vrijednost	3.61	3.04	1.40	0.94	0.91	0.83	0.78	0.53	0.49	0.48
Udio opisane varijancije/%	24.93	20.96	9.67	6.49	6.30	5.73	5.37	3.66	3.37	3.28
Kumulativna varijancija/%	24.93	45.89	55.57	62.06	68.36	74.09	79.46	83.12	86.49	89.77

Rezultat PCA dane su u tablici i grafički. Uzimanje prvih 5 glavnih komponenti za daljnju analizu može se opravdati normalnom raspodjelom faktorskih bodova za 6. glavnu komponentu.

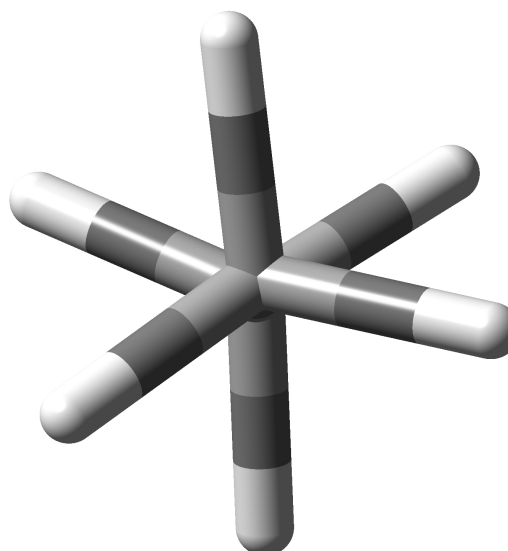


Slika 9. Rezultati PCA za etan u kartezijevim koordinatama.



Slika 10. Distribucija faktorskih bodova za 6. glavnu komponentu u kartezijevim koordinatama.

Dobivene prosječne geometrije za prvih 5 maksimuma su nezasjenjene i pripadaju minimumima na plohi potencijalne energije.



Slika 11. Geometrija maksimuma za etan u kartezijevim koordinatama.

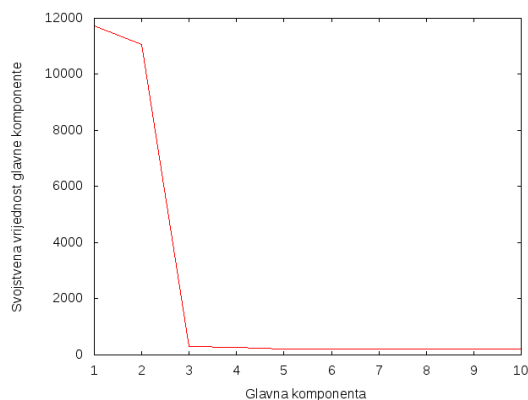
### 5.1.2. Analiza u internim koordinatama dobivenim funkcijom acos

Tablica 3. Rezultati PCA za etan u internim koordinatama

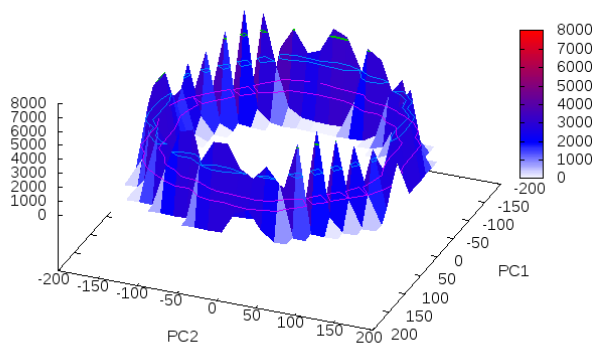
Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojstvena vrijednost	11733	11072	290.48	283.86	214.52	207.67	206.9	202.8	201.3	197.6
Udio opisane varijancije/%	45.31	42.76	1.12	1.10	0.83	0.80	0.80	0.78	0.78	0.76
Kumulativna varijancija/%	45.31	88.06	89.18	90.28	91.11	91.91	92.71	93.49	94.27	95.03

Rezultat PCA dane su u tablici i grafički.

Iz grafa i tablice se dobro vidi da je konvergencija postignuta već nakon prve 2 glavne komponente. Za daljnju analizu korištene su samo prve dvije glavne komponente.

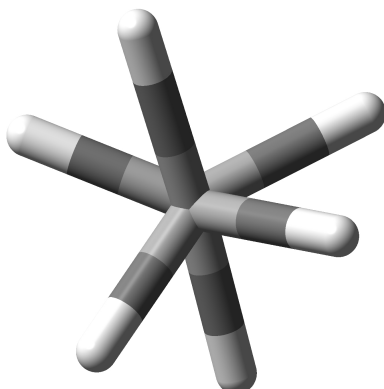


Slika 12. Rezultati PCA za etan u internim koordinatama.



Slika 13. Distribucija vjerojatnosti u prve dvije glavne komponente za etan u internim koordinatama.

Geometrije prvih 5 maksimuma pripadaju nezasićenim konformacijama. Jedina razlika između maksimuma je u tome u kojem od 3 energetska minimuma se nalaze. Prvi, drugi i peti maksimum pokrivaju geometrije sva 3 energetska minimum, stin da svaki predstavlja točno jednu geometriju. Treći i četvrti maksimum imaju geometriju istu kao i prvi maksimum.



Slika 14. Geometrija maksimuma za etan u internim koordinatama.

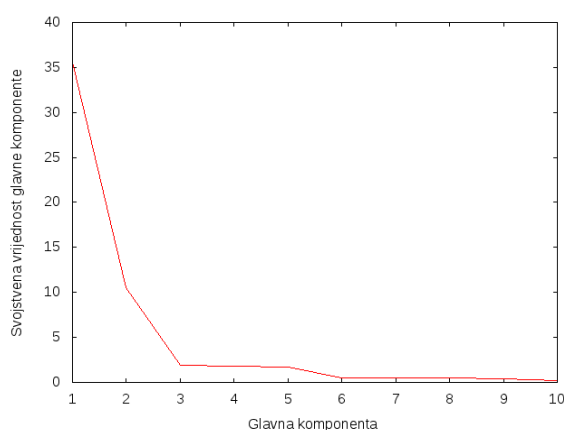
## 5.2. Cikloheksan

### 5.2.1. Analiza u kartezijevim koordinatama

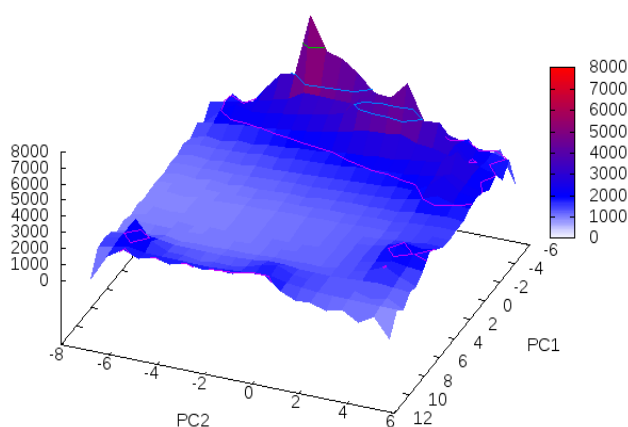
Tablica 4. Rezultati PCA za cikloheksan u kartezijevim koordinatama

Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojstvena vrijednost	35.57	10.57	1.91	1.85	1.69	0.52	0.51	0.48	0.40	0.20
Udio opisane varijancije/%	64.92	19.29	3.48	3.38	3.08	0.96	0.92	0.87	0.73	0.36
Kumulativna varijancija/%	64.92	84.21	87.69	91.07	94.15	95.11	96.03	96.90	97.63	97.99

Rezultat PCA dane su u tablici i grafički. S prve dvije glavne komponente je opisana većina varijancije i dovoljno je provesti ostatak analize za samo prve dvije komponente.

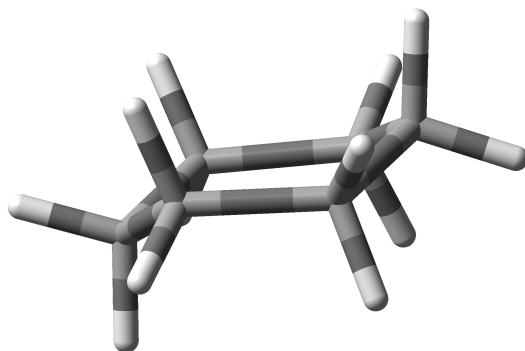


Slika 15. Rezultati PCA za cikloheksan u kartezijevim koordinatama.

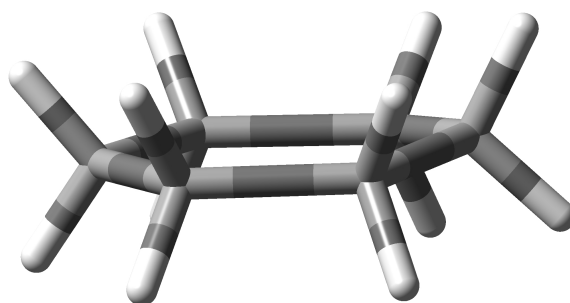


Slika 16. Distribucija vjerojatnosti u prve dvije glavne komponente za cikloheksan u kartezijevim koordinatama.

Geometrija iz prvog i petog maksimuma odgovara globalnom minimumu, odnosno konformaciji stolca. Očekivano je da 3 ostala maksimuma opisuju savijenu kolijevku. Međutim kod geometrija ostala 3 maksimuma svi ugljikovi atomi leže u istoj ravnini. Ovo bi mogla biti posljedica korištenja premalog *ndiv* ili prevelikog koraka u MD simulaciji.



Slika 17. Geometrija prvog i petog maksimuma za kartezijeve koordinate.



Slika 18. Geometrija drugog, trećeg i četvrtog maksimuma za kartezijeve koordinate.

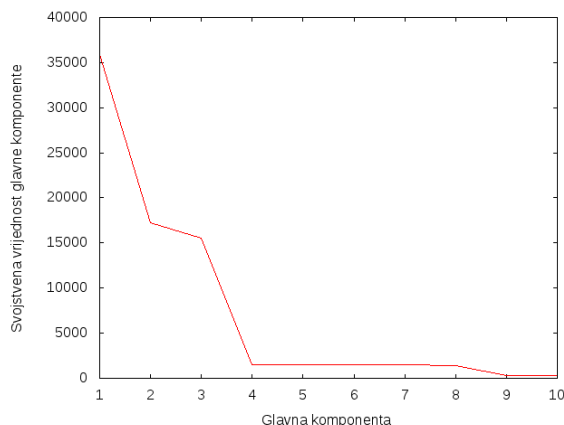
### 5.2.2. Analiza u internim koordinatama dobivenim funkcijom *acos*

Tablica 5. Rezultati PCA za cikloheksan u internim koordinatama

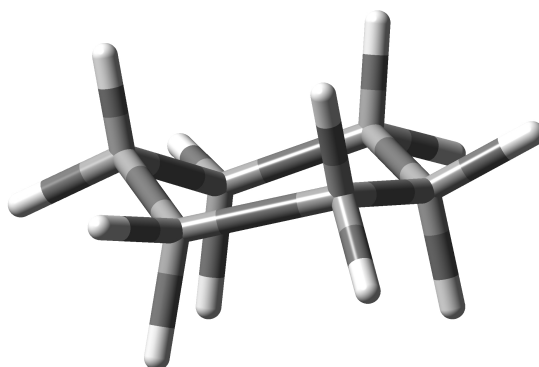
Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojstvena vrijednost	35921	17265	15542	1547	1542	1480	1457	1415	276	265
Udio opisane varijancije/%	44.27	21.28	19.16	1.91	1.90	1.82	1.80	1.74	0.34	0.33
Kumulativna varijancija/%	44.27	65.55	84.71	86.61	88.51	90.34	92.13	93.88	94.22	94.55

Rezultat PCA dane su u tablici i grafički. Nagli pad u opisu varijancije nakon 3. glavne komponente opravdava uzimanje samo prve 3 glavne komponente za daljnju analizu.

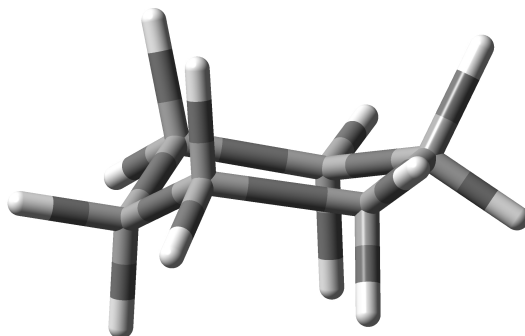




Slika 19. Rezultati PCA za cikloheksan u internim koordinatama.

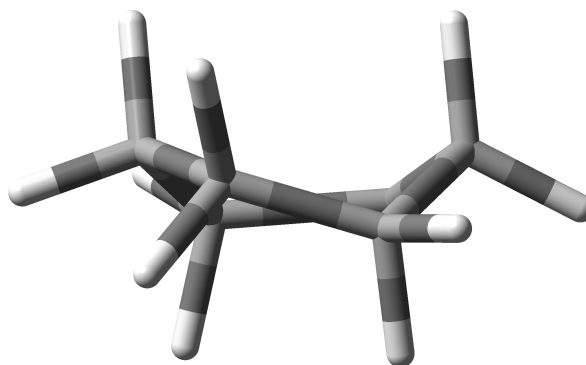


Slika 20. Geometrija najvećeg maksimuma za cikloheksan u internim koordinatama.

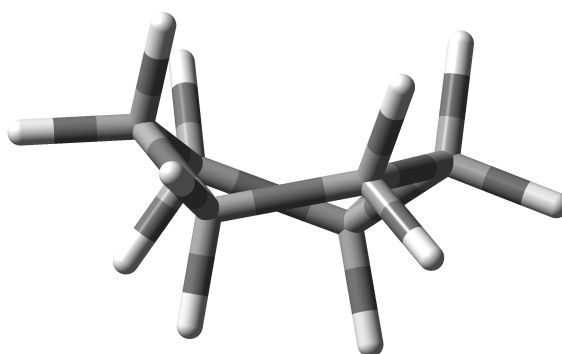


Slika 21. Geometrija drugog najvećeg maksimuma za cikloheksan u internim koordinatama.

Dva najveća maksimuma odgovaraju različitim konformacijama stolice. Sljedeća 2 maksimuma odgovaraju dvjema različitim savijenim kolijevkama. Nađena su sva 4 glavna minimuma na plohi potencijalne energije.



Slika 22. Geometrija trećeg najvećeg maksimuma za cikloheksan u internim koordinatama.



Slika 23. Geometrija četvrtog najvećeg maksimuma za cikloheksan u internim koordinatama.

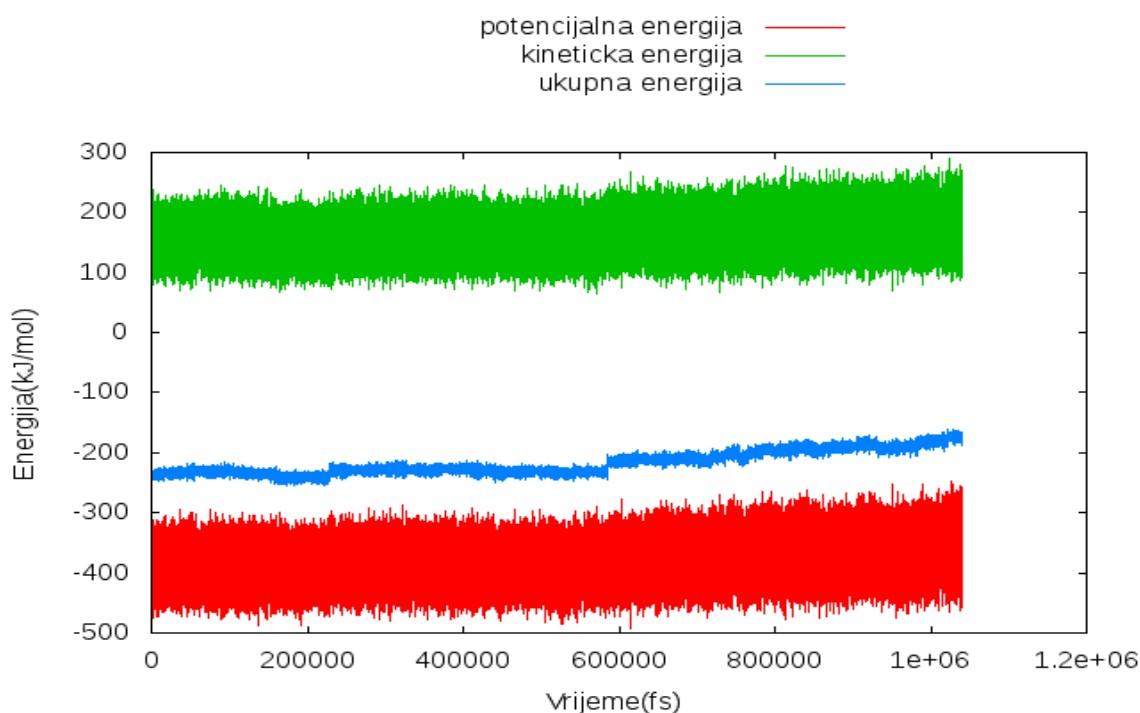
### 5.3. Glicin dipeptid

Statistički podaci vezani uz trajektoriji dani su u tablici.

Tablica 6. Prosječne energije i temperaturu za trajektoriju glicin dipeptida

	Srednja vrijednost	Standardna devijacija
Temperatura/K	769.7	123.5
Potencijalna energija/kJ mol <sup>-1</sup>	-379.2	28.5
Kinetička energija/kJ mol <sup>-1</sup>	163.2	26.2
Ukupna energija/kJ mol <sup>-1</sup>	-216.0	20.9

Iz grafa energijskog profila možemo zaključiti da je očuvanje energije bilo zadovoljavajuće.



Slika 24. Energijski profil trajektorije glicin dipeptida.

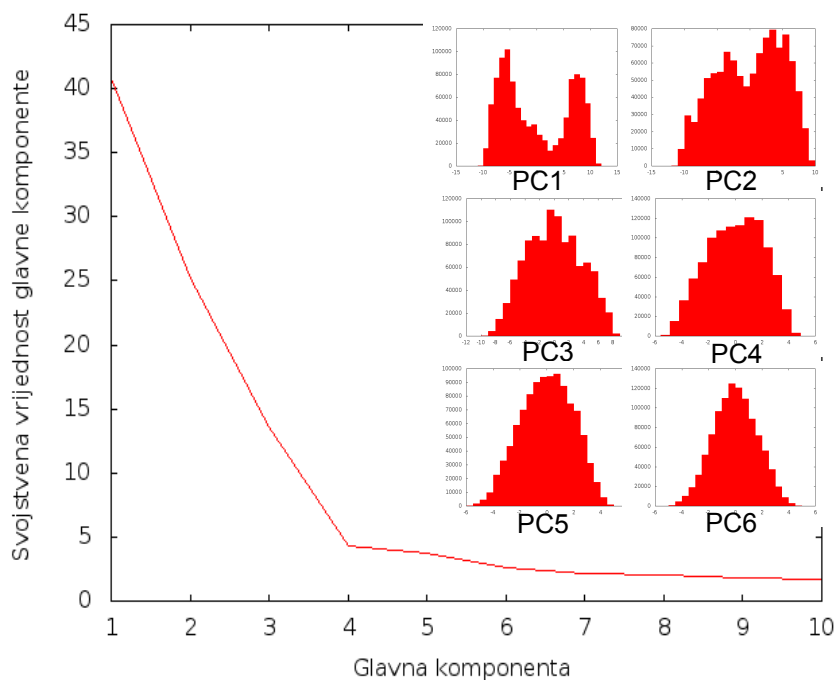
#### 5.3.1. Analiza u kartezijevim koordinatama

Tablica 7. Rezultati PCA za glicin dipeptid u kartezijevim koordinatama

Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojevredna vrijednost	40.77	25.19	13.61	4.28	3.73	2.66	2.21	2.05	1.83	1.73
Udio opisane varijancije/%	36.25	22.4	12.1	3.81	3.32	2.37	1.96	1.82	1.62	1.54
Kumulativna varijancija/%	36.25	58.65	70.75	74.56	77.88	80.25	82.22	84.04	85.66	87.21

Rezultat PCA dane su u tablici i grafički.

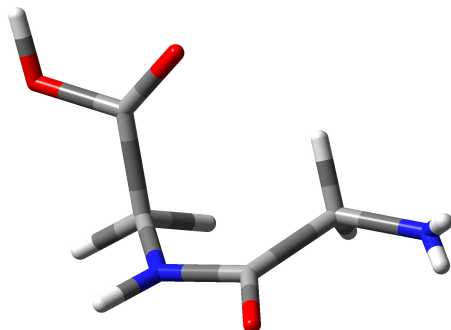
Iz distribucije faktorskih bodova vidi se da 6. glavna komponenta ima normalnu distribuciju. Prvih 5 glavnih komponenti opisuju nešto manje od 78% varijancije te je za njih proveden račun gustoće vjerojatnosti.



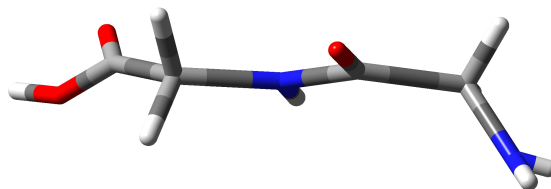
Slika 25. Rezultati PCA za glicin dipeptid u kartezijevim koordinatama. Mali grafovi predstavljaju distribuciju faktorskih bodova za prvih 6 glavnih komponenti.

Optimizacijom je pronađen velik broj maksimuma. Uzeta su 4 najveća maksimuma. Geometrije su dobivene uprosječivanjem geometrija koja pripadaju maksimumu. Geometrija najvećeg maksimuma ima glavni kostur unutar jedne ravnine s iznimkom da je terminalna karboksilna skupina okomita na tu ravninu. Mogući razlog stabilnosti je vodikova veza između karbonilnog kisika iz terminalne karboksilne skupine i vodika koji mu je prostorno najbliže. Iako je odsjenjenost vodika vezanog na ugljik mala, ipak je povoljnije da vodik gleda prema karbonilnom kisiku od cijele aminoskupine. Vodik iz aminoskupine bi tvorio bolju vodikovu vezu međutim cijela aminoskupina je prevelika i ta konfiguracija je sterički jako nepovoljna. Geometrija drugog maksimuma je gotovo ravnolančana uz iznimku terminalne amino skupine koja gleda izvan ravnine. Geometrija trećeg maksimuma je kombinacija prethodne dvije geometrije. Ovdje su repulzije minimizirane na način da su velike terminalne grupe orijentirane u suprotnim smjerovima s obzirom na ravninu u kojoj se nalazi većina glavnog kostura molekule. Zadnja geometrija

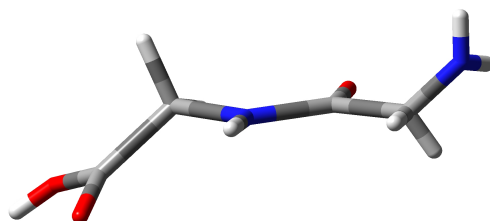
je u potpunosti ispružena ravnolačana konfiguracija. Vodici s terminalne amino skupine orijentirani su prema karbonilnom kisikovom atomu iz peptidne veze, dok je karbonilni kisik iz karboksilne skupine orijentiran prema vodiku vezanom na dušik u peptidnoj vezi.



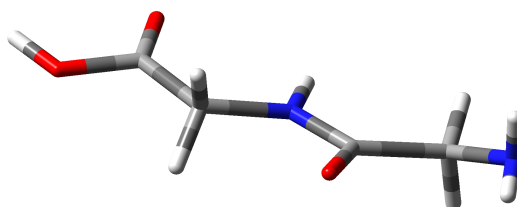
Slika 26. Prosječna geometrija najvećeg maksimuma.



Slika 27. Prosječna geometrija drugog najvećeg maksimuma.



Slika 28. Prosječna geometrija trećeg najvećeg maksimuma.



Slika 29. Prosječna geometrija četvrtog najvećeg maksimuma.

Analizom kartezijevih koordinata dobit će se geometrija u kojoj se trajektorija najdulje zadržala. Pri tome ako se molekula uspije izrotirati na način da se dobije ekvivalentna

konfiguracija ali pri nekim drugačijim apsolutnim kartezijevim položajima ovom metodom se takve geometrije razlikuju, iako su u stvarnosti ekvivalentne.

### 5.3.2. Analiza u internim koordinatama dobivenim funkcijom acos

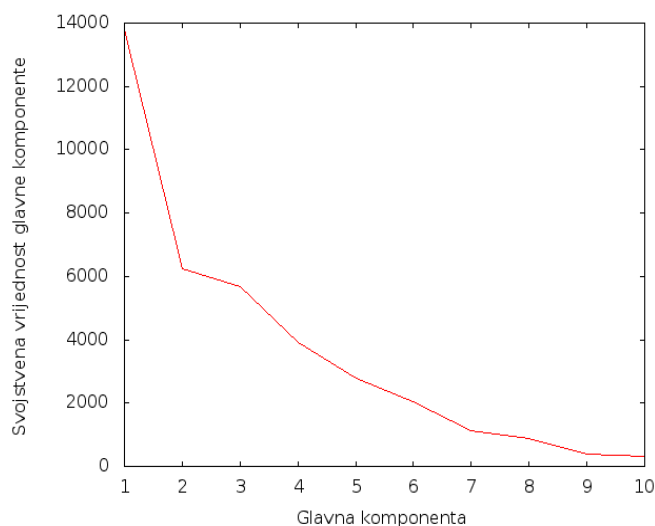
Rezultat PCA dane su u tablici i grafički.

Tablica 8. Rezultati PCA za glicin dipeptid u internim koordinatama dobivenim koristeći acos funkciju za računanje kuteva

Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojtvena vrijednost	13806	6249	5664	3925	2798	2061	1134	892	400	302
Udio opisane varijancije / %	35.82	16.21	14.69	10.18	7.26	5.35	2.94	2.31	1.04	0.78
Kumulativna varijancija / %	35.82	52.03	66.72	76.90	84.16	89.51	92.45	94.77	95.80	96.59
	5.52	-2.39	-0.46	-0.47	-0.64	-31.6	0.41	1.46	14.01	1.82
	-0.09	0.03	0.01	-0.06	-0.01	0.18	0.03	0.02	-0.72	-0.18
	-0.13	0.01	0.02	0.06	-0.01	0.08	-0.02	-0.13	-0.72	-0.05
	3.79	5.86	52.92	-0.01	0.65	-0.53	0.15	0.69	0.21	-0.06
	-5.63	2.78	0.71	0.50	0.95	31.89	-1.81	-3.20	13.68	1.74
	-64.35	12.06	0.50	-0.44	0.04	-2.01	0.43	0.78	0.95	-7.11
	65.48	-12.4	-0.48	0.41	-0.05	1.83	-0.44	-0.62	-1.53	6.87
	18.79	45.88	-5.13	0.35	0.07	-0.87	-4.62	0.65	-0.07	0.25
Faktorska opterećenja torzija	-8.84	-24.5	2.70	-44.3	1.88	0.89	2.98	-0.11	0.09	0.20
	-10.83	-25.1	2.94	44.14	-1.84	0.12	2.84	-0.58	0.08	0.15
	-4.70	-11.51	0.44	2.07	37.15	-1.24	-22.8	4.36	-0.29	-0.21
	-4.54	-10.8	2.26	-2.31	-37.5	0.18	-22.5	4.46	-0.16	-0.23
	-3.74	-5.92	-52.7	0.00	-0.67	0.08	-0.25	-0.58	0.19	-0.01
	3.94	0.55	-2.37	0.84	0.26	5.47	6.90	28.91	0.87	0.23
	-19.05	-46.3	5.21	-0.36	-0.14	0.91	3.36	-0.35	0.11	-0.06
	-65.58	11.56	0.83	-0.30	-0.09	-1.91	-0.21	0.60	-2.46	13.94

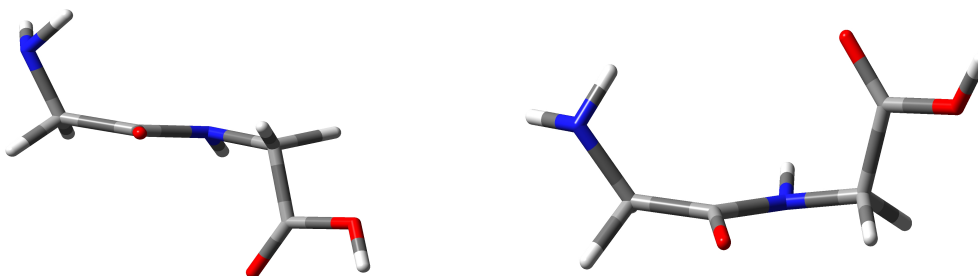
Varijancija u duljinama veza je zanemariva sprem torzija i kuteva. Ovo je očekivano jer u simulaciji ne dolazi do kemijskih promjena. Rotacije oko veza su dobro opisane torzijama. Također magnituda promjena kuteva u stupnjevima (od 0 do 180) je mnogo veća od promjene duljine veza tijekom vibracija (do 0.5). Zato je varijancija u promjeni torzijskih kuteva dominantna.

Opis većine faznog prostora molekule bit će postignut kada u glavnim komponentama sudjeluju sve torzije koje opisuju rotacije oko veza koje čine glavni kostur molekule. Možemo zaključiti da je za potpun opis potrebno provesti račun raspodjele vjerojatnosti za prvih 6 glavnih komponenti.



Slika 30. Rezultati PCA za glicin dipeptid u internim koordinatama dobivenim acos funkcijom.

Rezultati ukazuju da geometrije unutar jednog maksimuma odgovaraju različitim konformacijama. Ovo je očekivan rezultat za nesimetričan sustav poput dipeptida. Pošto funkcija acos nema informaciju o apsolutnoj orijentaciji torzijskog kuta (kodomena funkcije je  $[0, \pi]$ ) konformacije koje imaju neke torzije  $-90^\circ$  ili  $90^\circ$  stupnjeva će imati isti score jer će im funkcija acos dodijeliti vrijednost  $90^\circ$ . Zbog toga se unutar prvog maksimuma nalaze strukture u kojima su terminalne grupe orijentirane nekad u istom smjeru nekad u različitom smjeru, okomito na ravninu molekule.



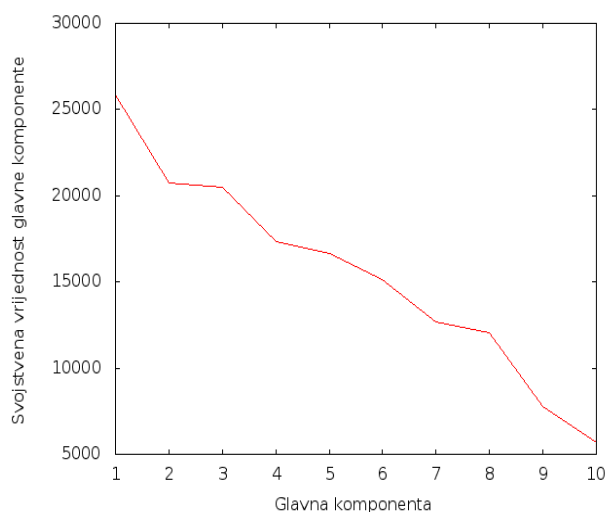
Slika 31. Geometrije sadržane u najvećem maksimumu.

### 5.3.3. Analiza u internim koordinatama dobivenim funkcijom atan2

Rezultati analize glavnih komponenti dani su u tablici.

Tablica 9. Rezultati PCA za glicin dipeptid u internim koordinatama dobivenim koristeći atan2 funkciju za računanje kuteva

Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojtvena vrijednost	25865	20760	20507	17358	16647	15163	12693	12054	7764	5671
Udio opisane varijancije / %	15.05	12.08	11.93	10.10	9.69	8.82	7.39	7.01	4.52	3.30
Kumulativna varijancija / %	15.05	27.13	39.06	49.16	58.85	67.68	75.06	82.08	86.59	89.89
Faktorska opterećenja torzija	0.69	-10.33	19.11	30.80	-6.80	-45.86	5.15	1.38	-2.65	1.48
	-0.07	-0.16	0.00	0.41	0.34	0.06	-0.01	0.03	0.10	-0.08
	-0.06	-0.16	0.03	0.42	0.37	0.06	0.01	0.00	0.08	-0.08
	2.06	3.31	0.74	-28.77	-62.99	-15.10	5.34	-3.88	-0.37	62.50
	-1.32	20.90	-28.92	-69.74	20.89	78.10	-9.72	-1.63	1.85	5.56
	1.11	-21.08	119.2	17.14	-19.89	53.28	1.26	-1.23	21.14	0.54
	0.06	2.89	-28.47	-1.28	5.82	-16.01	0.65	0.83	85.40	0.87
	0.18	95.03	7.80	26.52	-8.41	4.88	0.40	-0.87	-0.76	-1.56
	-0.37	-47.42	-6.07	-12.72	9.95	1.20	41.67	-64.24	-2.09	0.45
	0.23	-47.73	-2.68	-12.46	-4.05	-3.23	-41.99	67.15	-2.02	2.15
-0.64	9.90	-2.60	-0.66	2.19	7.18	67.45	40.55	0.13	2.24	
0.76	7.67	3.67	5.27	-3.93	-6.84	-66.84	-41.56	0.93	2.59	
-4.37	-8.60	1.62	50.78	94.03	13.07	-4.70	3.43	-1.60	41.17	
-160.7	-0.32	0.30	0.19	-4.44	-0.34	-0.45	-0.27	0.20	-0.35	
0.03	-71.87	-6.07	-16.65	4.19	-3.78	0.30	-2.25	0.19	-4.13	
-2.09	28.21	64.23	-80.73	51.78	-57.60	0.82	2.09	0.07	-0.43	



Slika 32. Rezultati PCA za glicin dipeptid u internim koordinatama dobivenim atan2 funkcijom.



Potpun opis glavnog kostura bit će postignut tek s 7 glavnih komponenata. Međutim dobar opis bit će već postignut sa samo 6 glavnih komponenti. Za opis trenda i glavnog nedostatka funkcije atan2 dovoljno je promotriti analizu za 6 glavnih komponenti.

Prva tri najveća maksimuma sadrže desnu geometriju sa Slike 31. Četvrti maksimum sadrži lijevu geometriju ali u nešto izduženom obliku. Iako je na raspolaganju samo dipeptid alanina

U slučaju akiralnih molekula poput glicina interne koordinate postaju loš odabir. Tako u slučaju gdje je raspon torzijskih kuteva  $-180^\circ$  do  $180^\circ$  geometrije čiji faktorski bodovi mogu biti jako različiti mogu biti vrlo bliske. Ovo se dobro vidi u rasponu faktorskih bodova koji obuhvaćaju prva 3 najveća maksimuma. Njihove geometrije su jako slične a faktorski bodovi u potpunosti drugačiji. Apsolutna vrijednost faktorskih bodova je jednaka no ne i njihov predznak. Promjenom apsolutne vrijednosti u faktorskim bodovima ( 4. maksimum) došlo je do značajne promijene geometrije. Kao glavni problem funkcije atan2 nameće se nekontinuiranost kod jako negativnih i jako pozitivnih kuteva. Geometrije čije torzije iznose  $179^\circ$  i  $-179^\circ$  su vrlo slične, no imat će faktorske bodove koji će se jako razlikovati (slučaj prva 3 maksimuma).

Tablica 10. Koordinate centra paralelepipeda korištenog za računanje raspodjele vjerojatnosti za 4 najveća maksimuma

Glavna komponenta	PC1	PC2	PC3	PC4	PC5	PC6
Maksimum 1	170.41	226.26	-77.89	-199.36	-18.08	-88.41
Maksimum 2	170.41	-231.92	71.79	200.88	42.57	83.77
Maksimum 3	-170.48	-262.47	71.79	200.88	12.24	83.77
Maksimum 4	170.41	134.63	251.41	-17.43	-109.05	-122.85

Acos funkcija nema takvih problema no ona ne prepoznaje različite orijentacije u prostoru, pa geometrije sa torzijama  $-90^\circ$  i  $90^\circ$  imaju iste faktorske bodove. Sada je očito da za konformacijsku analizu akiralnih molekula poput glicin dipeptida opis kutevima i torzijama nije zadovoljavajući. Potrebno je naći takav skup internih koordinata koji će moći odrediti orijentaciju torzija u prostoru i istovremeno biti kontinuirane.

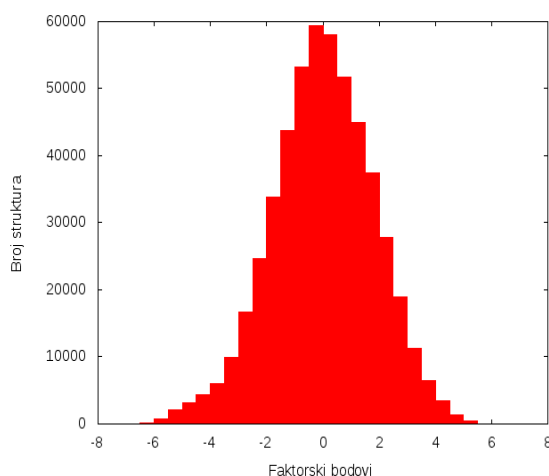
#### 5.3.4. Analiza za sve moguće udaljenosti

Rezultati analize glavnih komponenti dani su u tablici.

Tablica 11. Rezultati PCA za glicin dipeptid u kartezijevim koordinatama

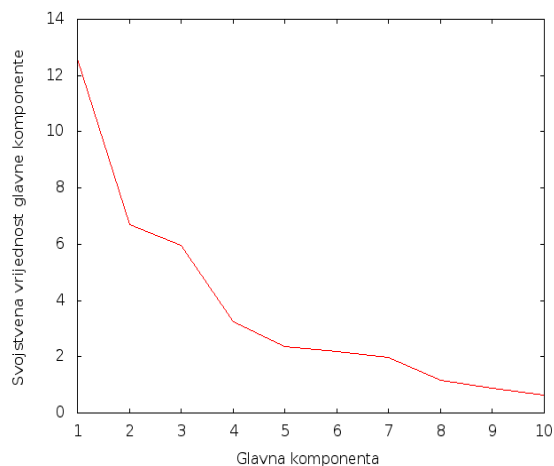
Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojevredna vrijednost	12.59	6.72	5.95	3.25	2.37	2.18	1.96	1.16	0.87	0.64
Udio opisane varijancije / %	28.86	15.39	13.63	7.46	5.43	4.99	4.49	2.66	1.98	1.47
Kumulativna varijancija / %	28.86	44.26	57.89	65.35	70.78	75.76	80.26	82.91	84.89	86.36

Udaljenost između atoma nameće se kao alternativa za probleme koje se susreću kod kuteva. Gledanje udaljenosti neka 4 atoma za koja možemo definirati torziju dat će nam rezultate analogne onima dobivenim funkcijom  $\cos$ . Međutim ako postoji atom u molekuli koji se nalazi izvana ravnine kojom je definiran torzijski kut, udaljenost između tog atoma i promatranog krajnjeg atoma, za kojeg definiramo torziju oko neke veze, će se razlikovati ovisno o tome u koju stranu se veza rotira. Pri tome kontinuiranost varijable neće biti izgubljena a dobivena je i orijentacija u prostoru. Rezultat PCA dane su u tablici i grafički.

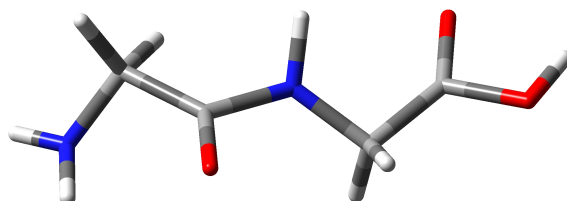


Slika 33. Raspodjela faktorskih bodova 4. glavne komponente.

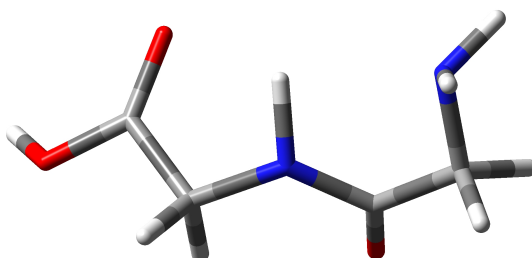
Glavna razlika će sada biti u odabiru glavnih komponenti dovoljnih za opis glavnog kostura. Treba uzeti u obzir da su vibracije sada istog reda veličine kao i ostale promjene. Ovo znači da će biti potreban manji ukupni udio varijancije za opis glavnog kostura za razliku od internih koordinata. Iz tablice i grafa vidi se da je znatan skok između 3. i 4. glavne komponente. Normalna raspodjela 4. glavne komponente također ide u prilog činjenici da uvrštavanje više od 3 glavne komponente nema smisla.



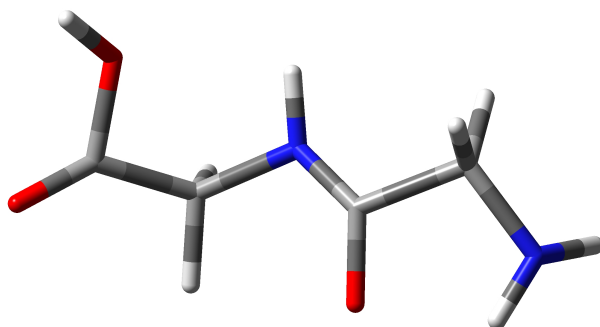
Slika 34. Rezultati PCA za glicin dipeptid u koordinatama dobivenim računanjem svih mogućih kombinacija udaljenosti među atomima.



Slika 35. Geometrija najvećeg maksimuma za analizu svih udaljenosti.

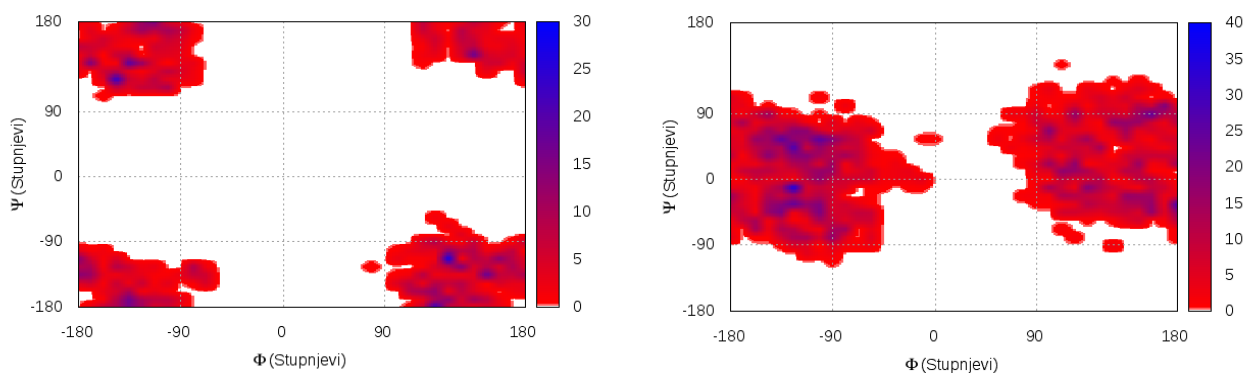


Slika 36. Geometrija drugog najvećeg maksimuma za analizu svih udaljenosti.



Slika 37. Geometrija trećeg najvećeg maksimuma za analizu svih udaljenosti.

Prvi maksimum odgovara ravnolačnoj molekuli. Karbonilni kisik iz karboksilne skupine je orijentiran prema vodiku vezanom na dušik u peptidnoj vezi, dok je jedan od vodikovih atoma orijentiran tako da gleda prema karbonilnom kisiku iz peptidne veze. U drugom maksimumu su dušik iz terminalne amino skupine i karbonilni kisik iz terminalne karboksilne skupine orijentirani prema vodiku vezanom na dušik iz peptidne veze. Tu dolazi do nastanka vodikove veze koja stabilizira ovu strukturu. Kod geometrije iz trećeg maksimuma vodikovi atomi koji se nalaze na terminalnoj amino skupini su orijentirani prema karbonilnom kisikovom atomu iz peptidne veze, dok je hidroksilni kisik iz terminalne karboksilne skupine orijentiran na vodikov atom vezan na dušik iz peptidne veze. Sve 3 geometrije se dobro slažu s podacima iz literature za slične spojeve.<sup>11</sup>



Slika 38. Ramachandrovi dijagrami za najveći maksimum (lijevo) i drugi najveći maksimum (desno).

Usporedba s Ramachandrovim dijagramom za glicin daje jako dobra slaganja.<sup>12</sup> Prvi i treći maksimum odgovaraju istom dijelu dijagrama. To je zato što u pravom polipeptidnom lancu hidroksilna skupina na karbonilnom terminalnom kraju bi bila zamijenjena s dušikom i dala peptidnu vezu. Tada ta geometrija nikad ne bi bila optimalna jer bi steričke smetnje bile prevelike. Međutim u našem slučaju kisikov atom iz hidroksilne skupine može poslužiti kao dobar akceptor u vodikovoj vezi te je ovakav rasplet očekivan.

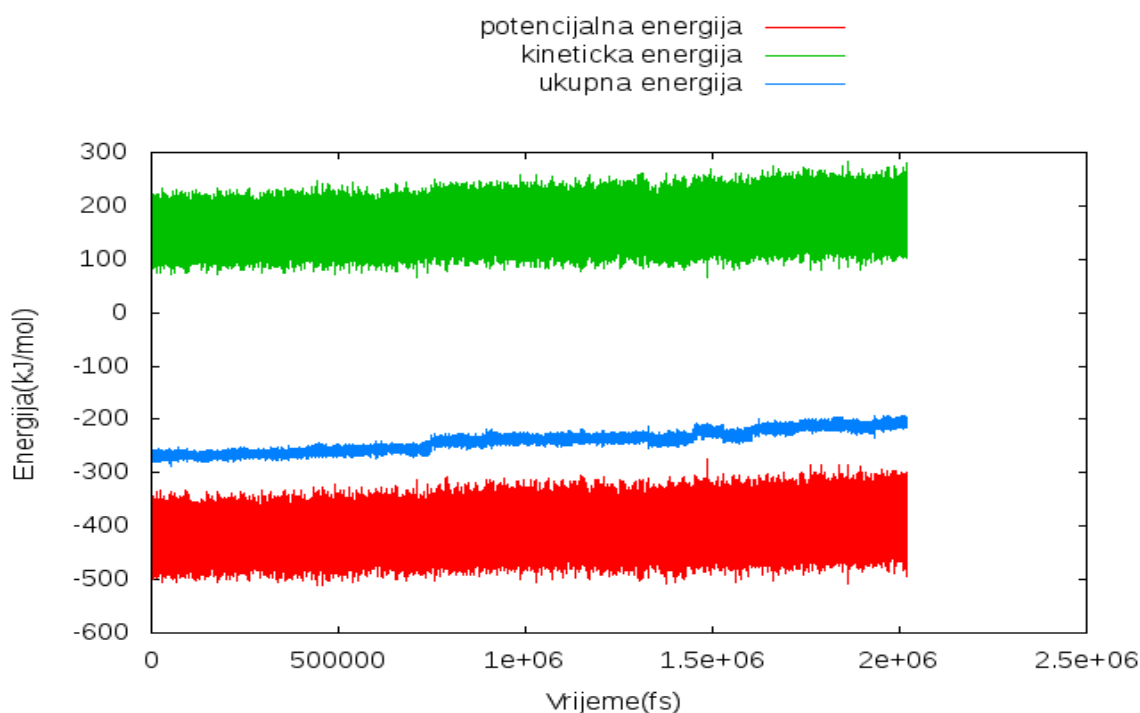
## 5.4. Alanilglicin

Statistički podaci vezani uz trajektoriju dani su u tablici.

Tablica 12. Prosječne energije i temperaturu za trajektoriju alanilglicina

	Srednja vrijednost	Standardna devijacija
Temperatura/K	665.9	98.0
Potencijalna energija/kJ mol <sup>-1</sup>	-406.2	26.3
Kinetička energija/kJ mol <sup>-1</sup>	166.1	24.4
Ukupna energija/kJ mol <sup>-1</sup>	-240.4	19.5

Iz grafa energijskog profila možemo zaključiti da je očuvanje energije bilo zadovoljavajuće.



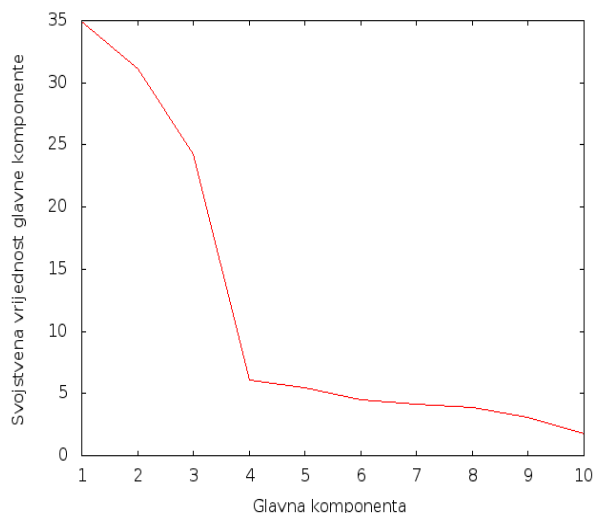
Slika 39. Energijski profil trajektorije alanilglicina.

### 5.4.1. Analiza u kartezijevim koordinatama

Tablica 13. Rezultati PCA za alanilglicin u kartezijevim koordinatama

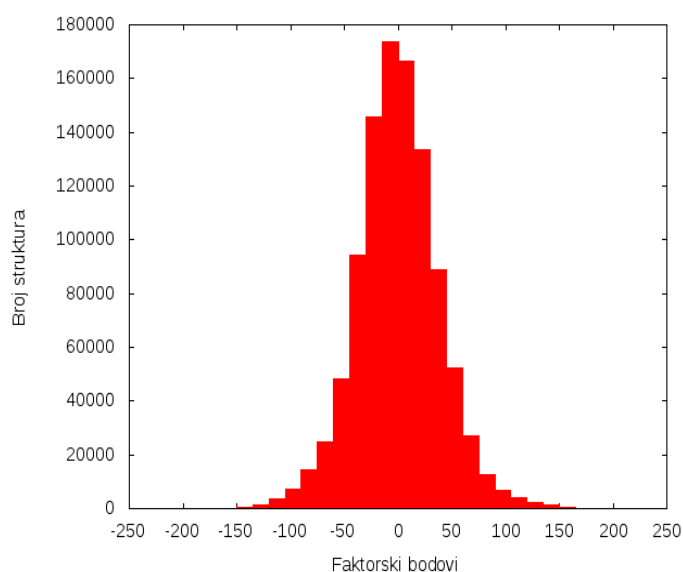
Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojevna vrijednost	34.91	31.09	24.26	6.10	5.45	4.48	4.15	3.89	3.07	1.76
Udio opisane varijancije /%	25.49	22.70	17.71	4.45	3.98	3.27	3.03	2.84	2.24	1.29
Kumulativna varijancija / %	25.49	48.19	65.90	70.35	74.34	77.61	80.64	83.48	85.72	87.01

Rezultat PCA dane su u tablici i grafički. Za daljnju analizu je odabrano prvih 5 glavnih komponenti. Ovo je opravdano činjenicom da je raspodjela 6. glavne komponente normalna, uz već postignuti visoki postotak opisa ukupne varijancije.

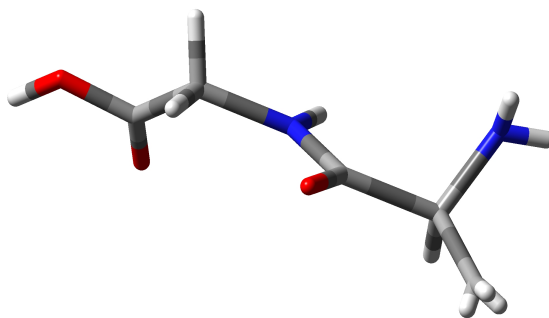


Slika 40. Rezultati PCA za alanilglicin u kartezijevim koordinatama.

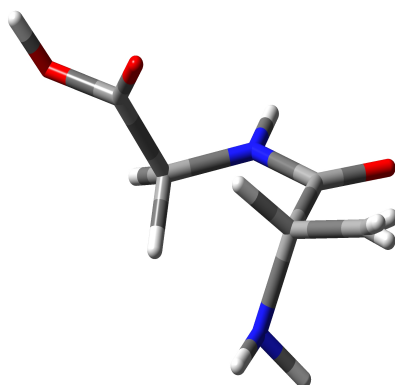
Geometrija prvog maksimuma predstavlja ravnolančanu molekulu u nezasjenjenoj konformaciji. Sljedeća dva maksimuma imaju geometriju sličnu onoj iz najvećeg maksimuma glicin dipeptida. Posljednja struktura je slična najvećem maksimumu s iznimkom da je karboksilna terminalna skupina sada više orijentirana izvan ravnine glavnog kostura molekule, dok je terminalna amino skupina orijentirana u suprotnom smjeru.



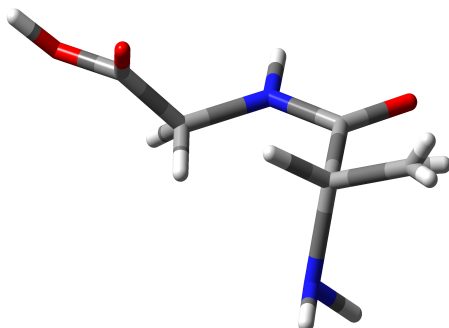
Slika 41. Raspodjela faktorskih bodova 6. glavne komponente.



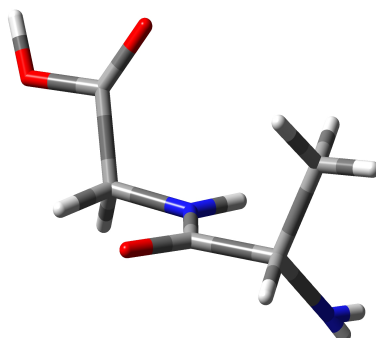
Slika 42. Geometrija najvećeg maksimuma za alanilglicin u kartezijskim koordinatama.



Slika 43. Geometrija 2. najvećeg maksimuma za alanilglicin u kartezijskim koordinatama.



Slika 44. Geometrija 3. najvećeg maksimuma za alanilglicin u kartezijskim koordinatama.



Slika 45. Geometrija 4. najvećeg maksimuma za alanilglicin u kartezijskim koordinatama.

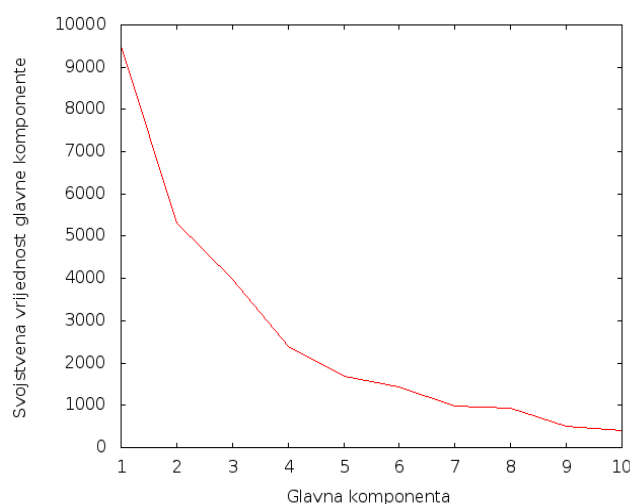
## 5.4.2. Analiza u internim koordinatama dobivenim funkcijom acos

Rezultat PCA dane su u tablici i grafički.

Tablica 14. Rezultati PCA za alanilglicin u internim koordinatama dobivenim funkcijom acos

Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojstvena vrijednost	9485	5319	3984	2398	1696	1434	990	921	496	398
Udio opisane varijancije / %	31.92	17.90	13.41	8.07	5.71	4.83	3.33	3.10	1.67	1.34
Kumulativna varijancija / %	31.92	49.83	63.24	71.31	77.02	81.84	85.17	88.27	89.94	91.28

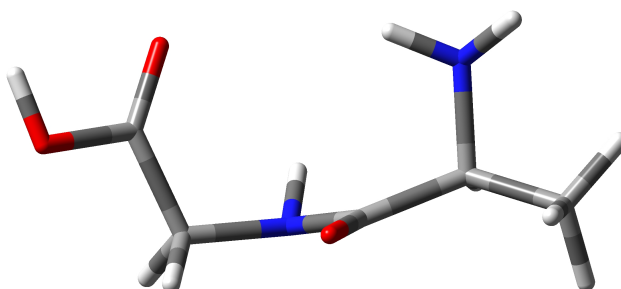
Nakon 5. glavne komponente opisano je već 77% varijancije što bi trebalo biti dovoljno za adekvatan opis sustava.



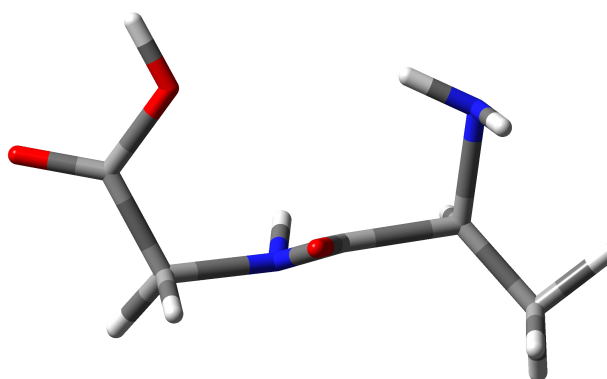
Slika 46. Rezultati PCA za alanilglicin u internim koordinatama dobivenim funkcijom acos.

Sva četiri promatrana maksimuma odgovaraju geometrijama gdje su obje terminalne skupine orijentirane izvan ravnine ostatka kostura molekule u istim smjerovima, za razliku od analize u kartezijevim koordinatama. Strukture u prvom i trećem maksimumu su nešto polegnutije dok kod struktura u drugom i četvrtom maksimumu čak nastaje nastaje vodikova veza između terminalnih skupina. Razlika između drugog i četvrtog maksimuma je to što je kod drugog maksimuma u vodikovoj vezi sudjeluje karbonilni kisikov atom iz terminalne karboksilne skupine, dok u četvrtom maksimumu je to hidroksilni kisikov atom iz terminalne karboksilne skupine.





Slika 47. Geometrija 2. najvećeg maksimuma za alanilglicin u internim koordinatama dobivenim funkcijom acos.



Slika 48. Geometrija 4. najvećeg maksimuma za alanilglicin u internim koordinatama dobivenim funkcijom acos.

### 5.4.3. Analiza u internim koordinatama dobivenim funkcijom atan2

Rezultat PCA dane su u tablici i grafički.

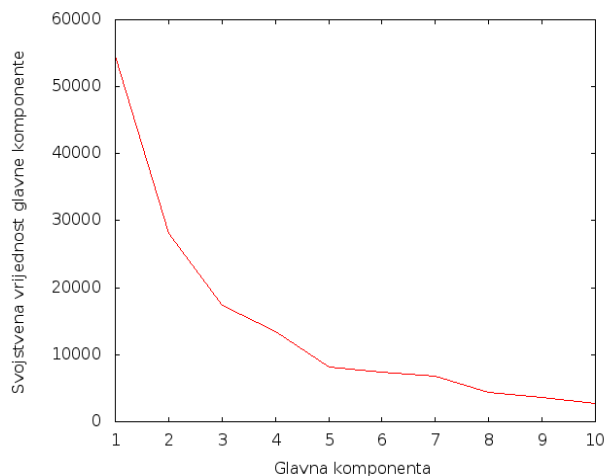
Tablica 15. Rezultati PCA za alanilglicin u internim koordinatama dobivenim funkcijom atan2

Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojevna vrijednost	54585	28062	17330	13482	8126	7366	6756	4361	3630	2745
Udio opisane varijancije / %	34.69	17.83	11.01	8.57	5.16	4.68	4.29	2.77	2.31	1.74
Kumulativna varijancija / %	34.69	52.52	63.53	72.10	77.26	81.94	86.24	89.01	91.31	93.06

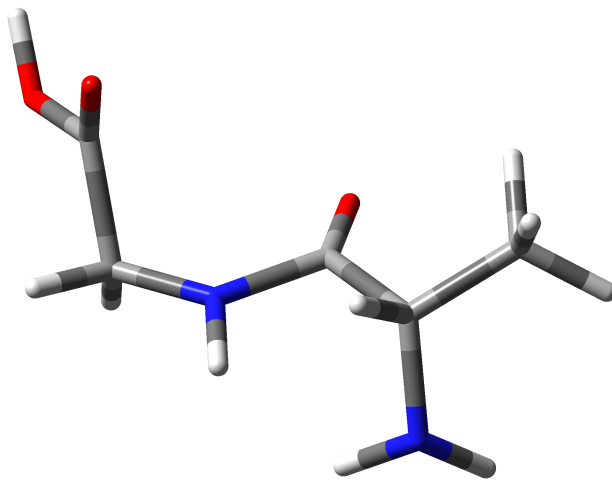
Analogno prethodnom slučaju odabrano je prvih 5 glavnih komponenti za daljnju analizu. PCA je dala vrlo slične rezultate u slučaju atan2 i cos funkcija za interne koordinate.

Prvi i treći maksimum odgovaraju vrlo sličnim strukturama. Terminalne skupine su orijentirani okomito na ravninu ostatka kostura molekule i to u suprotnim smjerovima. Drugi i četvrti maksimum odgovaraju strukturama u kojima su terminalne skupine orijentirane

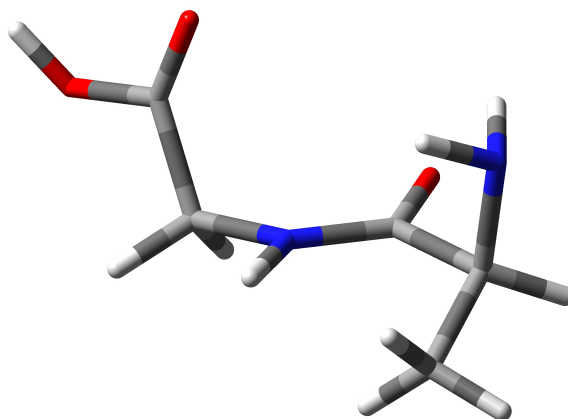
okomito na ostatak molekule ali u istim smjerovima.



Slika 49. Rezultati PCA za alanilglicin u internim koordinatama dobivenim funkcijom atan2.



Slika 50. Geometrija 1. i 3. najvećeg maksimuma za alanilglicin u internim koordinatama dobivenim funkcijom atan2.



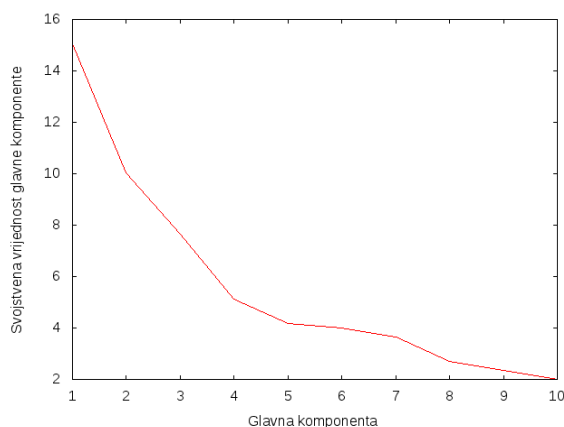
Slika 51. Geometrija 2. i 4. najvećeg maksimuma za alanilglicin u internim koordinatama dobivenim funkcijom atan2.

## 5.4.4. Analiza za sve moguće udaljenosti

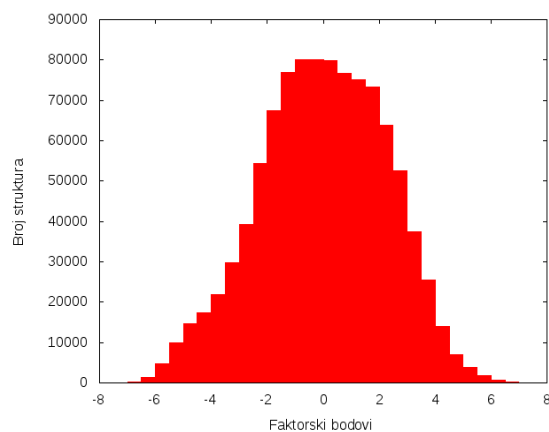
Rezultat PCA dane su u tablici i grafički.

Tablica 16. Rezultati PCA za alanilglicin za sve moguće udaljenosti

Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojtvena vrijednost	15.06	10.04	7.66	5.12	4.19	4.01	3.65	2.70	2.36	2.00
Udio opisane varijancije / %	22.01	14.68	11.19	7.48	6.13	5.87	5.33	3.94	3.44	2.92
Kumulativna varijancija / %	22.01	36.68	47.88	55.36	61.49	67.35	72.69	76.63	80.07	83.00



Slika 52. Rezultati PCA za alanilglicin za sve moguće udaljenosti

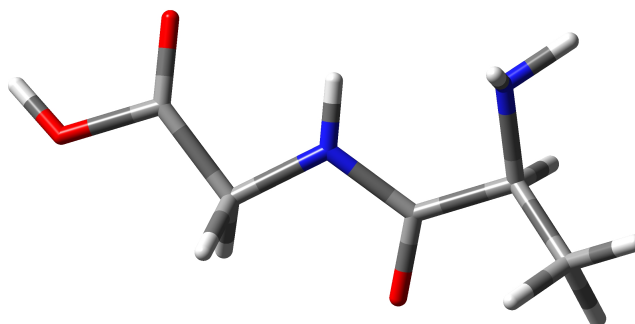


Slika 53. Raspodjela faktorskih bodova 4. glavne komponente.

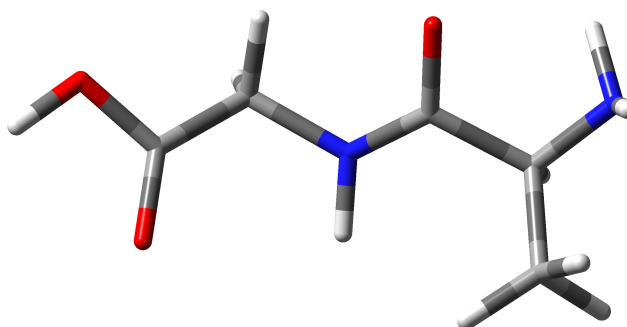
Četvrta glavna komponenta ima već približno normalnu distribuciju pa je opravdano odabrati prve 3 glavne komponente za daljnju analizu.

Prvi i zadnji maksimum odgovaraju strukturama gdje su terminalne skupine orijentirane na istu stranu u prostoru. Kod prvog maksimuma može se uočiti vodikova veza između karbonilnog kisikovog atoma iz terminalne karboksilne skupine i vodika vezanog na dušik u peptidnoj vezi, te istog vodika sa dušikom iz terminalne amino skupine. Treći maksimum

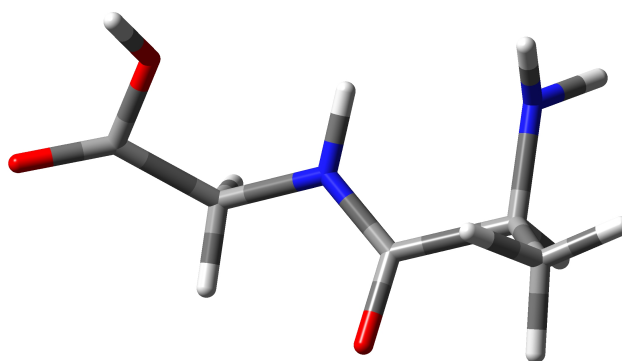
ima vrlo sličnu strukturu. Jedina je razlika što kod terminalne karboksilne skupine u vodikovu vezu stupa hidroksilni kisik. Za očekivati je da je ta vodikova veza slabija te da je struktura 3. maksimuma očekivano više energije od one iz 1. maksimuma. U geometriji drugog maksimuma terminalne funkcionalne grupe su orijentirane u suprotne strane. Ponovno je prisutna vodikova veza između karbonilnog kisika iz karboksilne skupine i vodika vezanog na dušik u peptidnoj vezi, te između karbonilnog kisika iz peptidne veze i jednog od vodika vezanih na dušik u terminalnoj amino skupini.



Slika 54. Geometrija najvećeg maksimuma za analizu svih udaljenosti.



Slika 55. Geometrija drugog najvećeg maksimuma za analizu svih udaljenosti.



Slika 56. Geometrija trećeg najvećeg maksimuma za analizu svih udaljenosti.

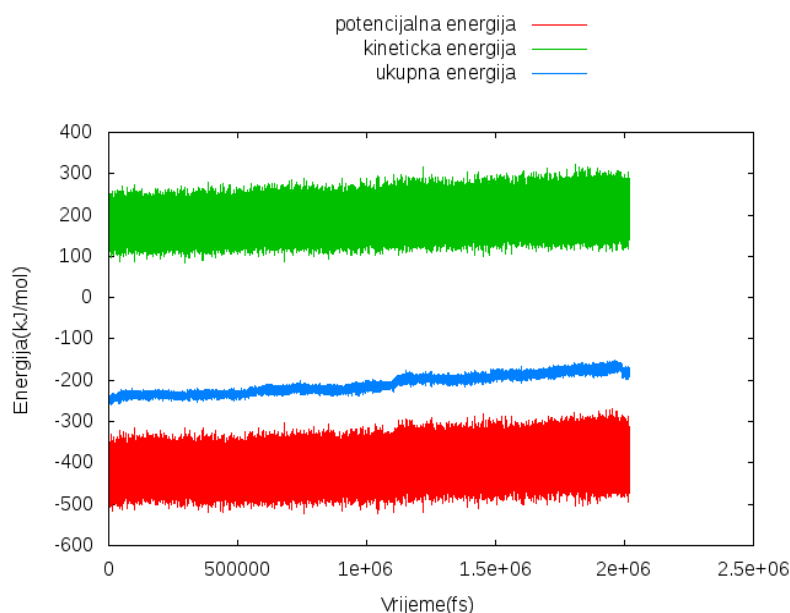
## 5.5. Alanin dipeptid

Statistički podaci vezani uz trajektoriji dani su u tablici.

Tablica 17. Prosječne energije i temperaturu za trajektoriju alanin dipeptid

	Srednja vrijednost	Standardna devijacija
Temperatura/K	681.0	93.7
Potencijalna energija/kJ mol <sup>-1</sup>	-404.4	29.2
Kinetička energija/kJ mol <sup>-1</sup>	195.4	26.9
Ukupna energija/kJ mol <sup>-1</sup>	-208.9	22.8

Iz grafa energijskog profila možemo zaključiti da je očuvanje energije bilo zadovoljavajuće.



Slika 57. Energijski profil trajektorije alanin dipeptid

### 5.5.1. Analiza u kartezijevim koordinatama

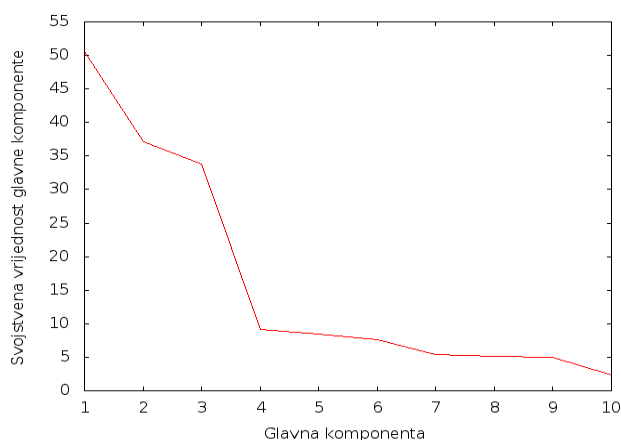
Rezultat PCA dane su u tablici i grafički.

Tablica 18. Rezultati PCA za alanin dipeptid u kartezijevim koordinatama

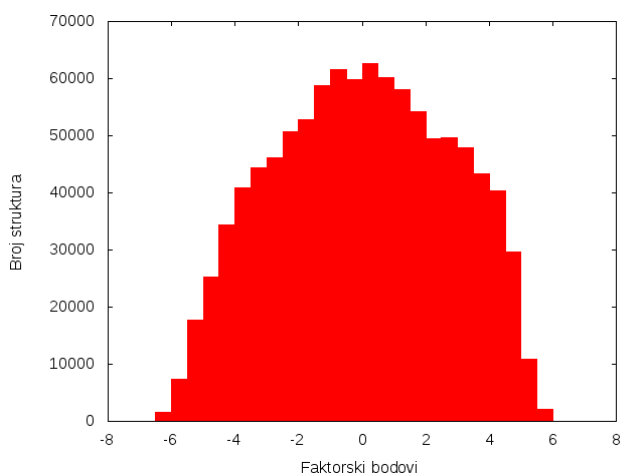
Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojevna vrijednost	50.55	37.15	33.77	9.08	8.52	7.64	5.46	5.16	4.95	2.33
Udio opisane varijancije / %	27.02	19.86	18.05	4.86	4.55	4.08	2.92	2.76	2.64	1.25
Kumulativna varijancija / %	27.02	46.87	64.93	69.78	74.33	78.42	81.33	84.09	86.74	87.98

Nagli pad u postotku opisane varijancije događa se već nakon prve 3 glavne komponente.

Međutim normalna distribucija faktorskih bodova se tek primjećuje na šestoj glavnoj komponenti.

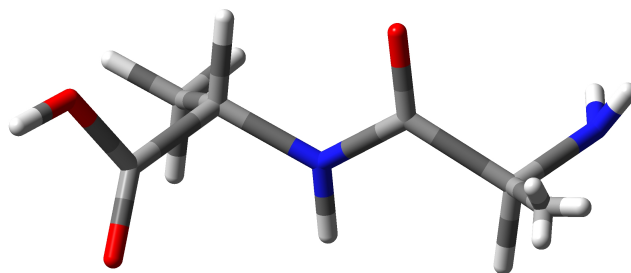


Slika 58. Rezultati PCA za alanin dipeptid u kartezijevim koordinatama.

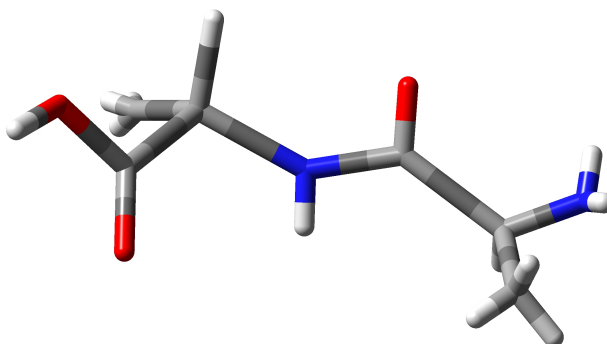


Slika 59. Raspodjela faktorskih bodova 6. glavne komponente.

Uprosječene geometrije za 4 najveća maksimuma dane su na slikama. Prva 3 maksimuma odgovaraju istim geometrijama. Terminalna karboksilna i amino skupina su orijentirane na suprotne strane. Ovakva konformacija je nezasjenjena duž svih veza i očekivano je najpovoljnija. U geometriji 4. najvećeg maksimuma terminalna amino skupina je u zasjenjenoj konfiguraciji sprem karbonilnog kisika iz peptidne veze. Ovo se može objasniti vodikovom vezom između jednog od vodika iz terminalne amino skupine i karbonilnog kisika iz peptidne veze. U obje geometrije karbonilni atom iz terminalne karboksilne skupine je orijentiran prema vodikovom atomu vezanom na dušikov atom iz peptidne veze.



Slika 60. Geometrije prva 3 najveća maksimuma u kartezijevim koordinatama.

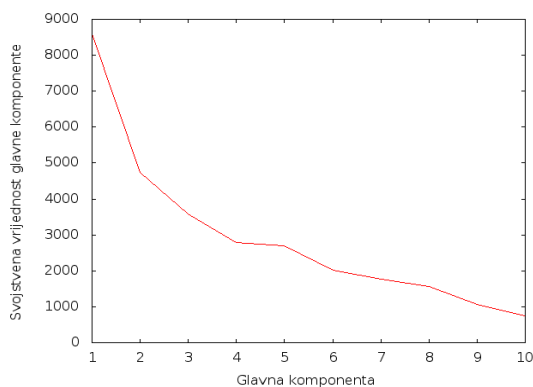


Slika 61. Geometrija 4. najvećeg maksimuma u kartezijevim koordinatama.

### 5.5.2. Analiza u internim koordinatama dobivenim funkcijom acos

Tablica 19. Rezultati PCA za alanin dipeptid u internim koordinatama dobivenim funkcijom acos

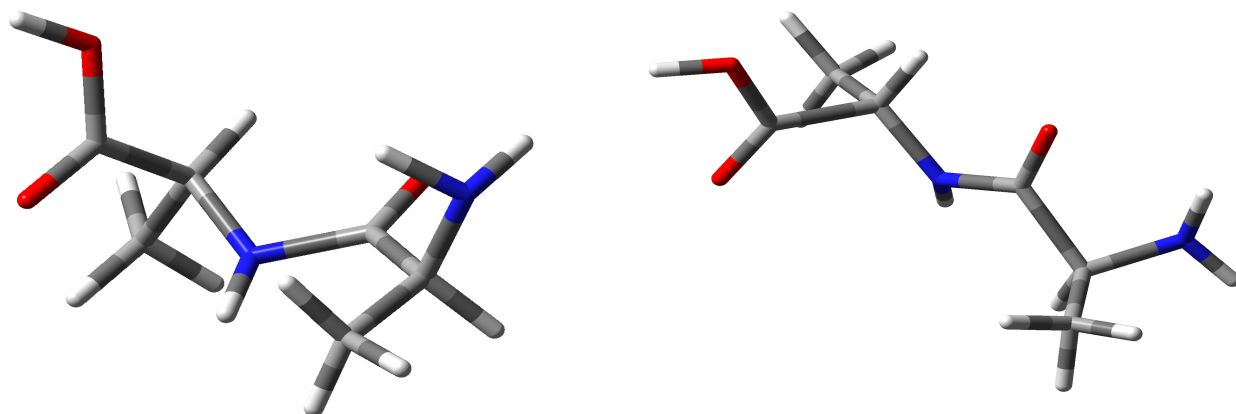
Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojstvena vrijednost	8584	4728	3587	2789	2702	2021	1769	1571	1065	748
Udio opisane varijancije / %	25.92	14.28	10.83	8.42	8.16	6.10	5.34	4.74	3.22	2.26
Kumulativna varijancija / %	25.92	40.19	51.02	59.45	67.60	73.71	79.05	83.79	87.01	89.27



Slika 62. Rezultati PCA za alanin dipeptid u internim koordinatama dobivenim funkcijom acos.

Rezultat PCA dane su u tablici i grafički.

Za opis sustava uzeto je prvih 5 glavnih komponenti. Interne koordinate opisane funkcijom acos ponovno daju dvosmislene rezultate što je sad već očekivano. U maksimumima dominiraju strukture analogne onima dobivenim kartezijevim koordinatama.



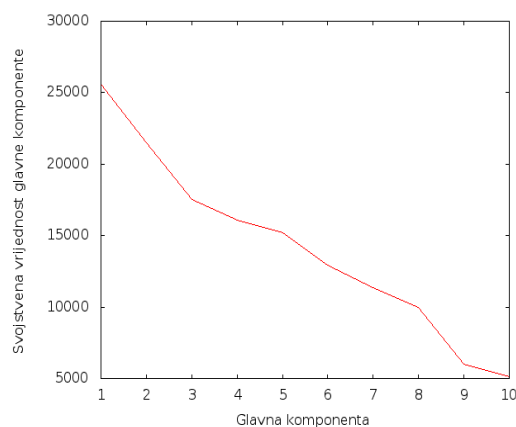
Slika 63. Geometrije iz najvećeg maksimuma za funkciju acos.

### 5.5.3. Analiza u internim koordinatama dobivenim funkcijom atan2

Rezultat PCA dane su u tablici i grafički.

Tablica 20. Rezultati PCA za alanin dipeptid u internim koordinatama dobivenim funkcijom atan2

Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojevstvena vrijednost	25571	21502	17538	16059	15177	12950	11343	9967	6019	5111
Udio opisane varijancije / %	16.58	13.94	11.37	10.41	9.84	8.39	7.35	6.46	3.90	3.31
Kumulativna varijancija / %	16.58	30.51	41.88	52.29	62.13	70.53	77.88	84.34	88.24	91.56



Slika 64. Rezultati PCA za alanin dipeptid u internim koordinatama dobivenim funkcijom atan2.



Analogno prethodnom slučaju, odabrano je prvih 5 glavnih komponenti za analizu raspodjele vjerojatnosti. Pošto je alanin dipeptid u potpunosti akiralna molekula simetričnost raspodjele faktorskih bodova neće raditi probleme pri analizi.

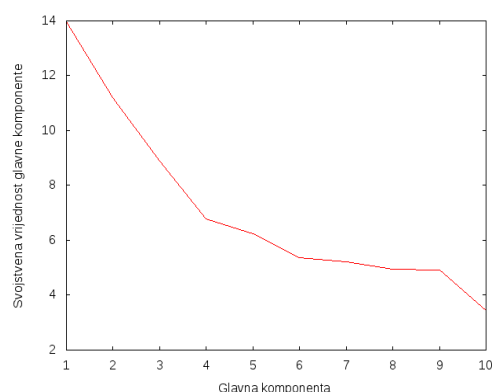
Rezultati su jako slični onima za kartezijske. Prva 3 maksimuma odgovaraju geometrijama gdje su terminalne skupine orijentirane na suprotne strane, dok 4. maksimum odgovara više zasjenjenoj strukturi 4. maksimuma kod kartezijskih koordinata.

#### 5.5.4. Analiza za sve moguće udaljenosti

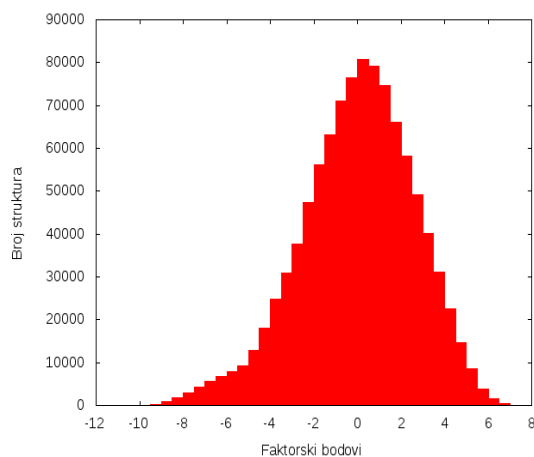
Rezultati PCA dani su u tablici i grafički.

Tablica 21. Rezultati PCA za alanin dipeptid dobiveni računom svih udaljenosti

Glavna komponenta	1	2	3	4	5	6	7	8	9	10
Svojsvena vrijednost	13.96	11.20	8.88	6.79	6.23	5.36	5.21	4.95	4.92	3.43
Udio opisane varijancije / %	14.95	11.99	9.50	7.27	6.67	5.74	5.58	5.30	5.26	3.67
Kumulativna varijancija / %	14.95	26.94	36.45	43.71	50.38	56.12	61.70	67.01	72.27	75.94

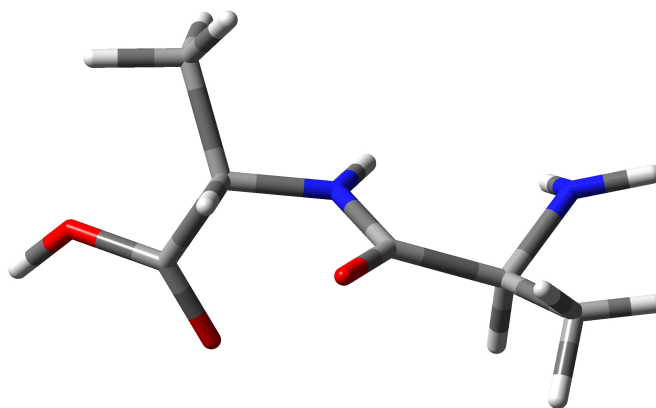


Slika 65. Rezultati PCA za alanin dipeptid dobiveni računom svih udaljenosti.

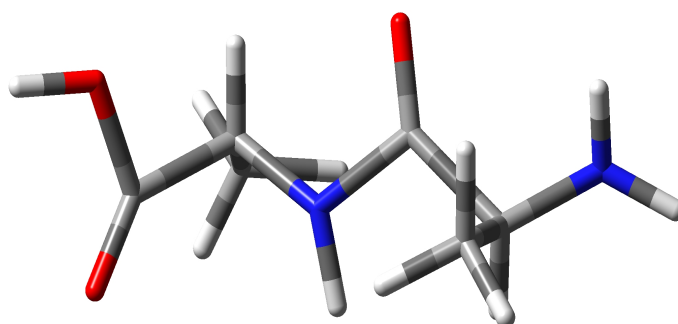


Slika 66. Raspodjela faktorskih bodova 4. glavne komponente.

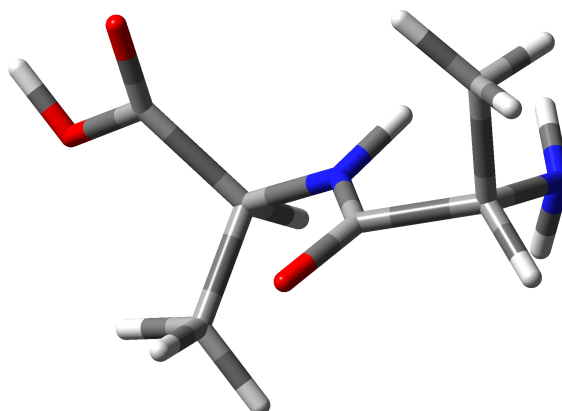
Za daljnju analizu odabrane su prve 3 glavne komponente slično kao u prethodnim slučajevima. Distribucija faktorskih bodova za 4. komponentu je već normalna stoga nema potrebe uzimati više glavnih komponenti.



Slika 67. Geometrija najvećeg maksimuma za analizu svih udaljenosti.



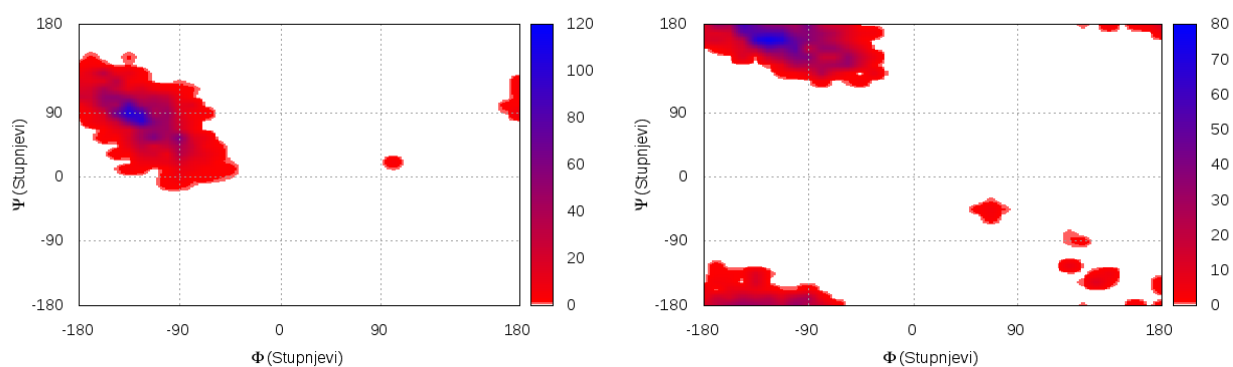
Slika 68. Geometrija drugog najvećeg maksimuma za analizu svih udaljenosti.



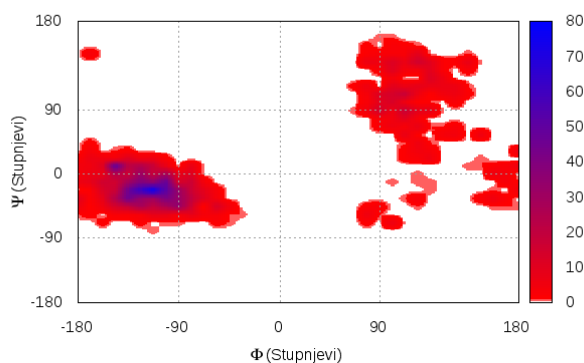
Slika 69. Geometrija trećeg najvećeg maksimuma za analizu svih udaljenosti.

Kod geometrije prvog i trećeg maksimuma aminokiselinski ostatak terminalnog amino dijela nalazi se u ekvatorijalnom odnosno aksijalnom položaju. Drugi maksimum je pak definiran vodikovim vezama između vodika sa terminalne amino skupine i karbonilnog kisika iz peptidne veze, te između karbonilnog kisika iz terminalne karboksilne skupine i vodika vezanog na dušik u peptidnoj vezi. Ovi rezultati su u odličnom slaganju s literaturom što se tiče geometrija<sup>13</sup> te relativnih odnosa energija.<sup>14</sup>

Napravljeni su dijagrami slični Ramachandrovim dijagramima i uspoređeni s literaturom.<sup>15,16</sup> Dobivena su odlična slaganja te su opažene sve važnije strukture.



Slika 70. Ramachandrovi dijagrami za najveći maksimum (lijevo) i drugi najveći maksimum (desno).



Slika 71. Ramachandrov dijagram za 3. najveći maksimum.

## 6. Zaključak

Analiza alifata u internim koordinatama dala je jako dobre rezultate, pogotovo za cikloheksan. Možemo zaključiti da u trajektoriji cikloheksan poprima geometrije svih glavnih struktura te da dolazi do izvrtnja prstena. Metoda radi jako dobro za male alkane, te cikloalkane. Rezultati u kartezijevim koordinatama su nešto slabiji no mogu biti od koristi.

Tek se kod peptida vide glavne slabosti internih koordinata. Novi pristup generiranju internih koordinata u vidu svih mogućih udaljenosti između atoma daje odlične rezultate za sva 3 peptida. Slaganje s podacima iz literature je odlično, a strukturama dominiraju vodikove veze koje očito imaju veliku ulogu u stabilizaciji malih peptida. Ovo je očekivano jer i u proteinima su vodikove veze glavni faktor kod nastajanja native konfiguracije.

Odabir koordinatnog sustava za PCA igra ključnu ulogu u procesu analize. Za standardni minimalni skup internih koordinata opis ponekad zna biti dosta dobar (etan), ali ovo jako ovisi o kiralnosti molekule (glicin dipeptid) i elementima simetrije. Čim se upotrijebe nestandardne interne koordinate rezultati postaju bolji (alanin dipeptid). Iako se dobri rezultati mogu dobiti kartezijevim koordinatama najbolja opcija su sve moguće udaljenosti između atoma. Broj glavnih osi potreban za opis sustava se smanjuje gotovo za pola a rezultati se jako dobro slažu s literaturom. Broj mogućih udaljenosti raste eksponencijalno što za veće molekule može pružiti veliki problem u vidu memorije potrebne za račun. Ovaj problem može se prebroditi pametnim odabirom udaljenosti koje će biti uzete za analizu. Udaljenosti duž veza opisuju samo vibracije i one nisu korisne za opisivanje konformacije molekule tako da mogu biti izbačene. Također mnoge kombinacije udaljenosti će uvijek imati isti trend, odnosno bit će jako korelirane i uključivanje više takvih koordinata u analizu je nepotrebno.

Metoda konformacijske analize i molekulske dinamike pomoću PCA i distribucije vjerojatnosti te njene optimizacije pokazala se jako uspješnom na nekoliko manjih i lakših, ali i nešto kompliciranijih molekula. Uz nekoliko manjih modifikacija metoda se može koristiti i za mnogo veće molekule te ima dobru perspektivu.

## 7. Sažetak

Domagoj Fijan

Teorijska konformacijska analiza alifatskih spojeva i peptida

Razvijena je i testirana metoda konformacijske analize za proučavanje malih i srednje velikih molekula koja iz molekulske dinamike uz pomoć analize glavnih komponenti i računa distribucije vjerojatnosti pronalazi minimume na plohi potencijalne energije. Metoda je testirana na nekoliko alifatskih spojeva i peptida i dobiveno je odlično slaganje s literaturom. Proučen je utjecaj odabira različitih koordinatnih sustava na analizu, te je pronađeno da najbolje rezultate daju koordinate opisane svim mogućim udaljenostima između atoma.

**Ključne riječi** : konformacijska analiza, peptidi, alifati, molekulska dinamika, analiza glavnih komponenti

## 8. Summary

Domagoj Fijan

Theoretical conformational analysis of aliphatic compounds and peptides

New method of conformational analysis has been developed and tested for small and medium sized molecules. Data from molecular dynamics was analyzed with principal component analysis and probability distribution was calculated and optimized. Probability distribution maximums were found to correspond to minima on potential energy surface. Method was tested on several aliphatic compounds and peptides with very good agreement with literature data. Impact of coordinate system chosen in analysis was studied and it was found that best results were obtained when distances between all atoms were used.

**Keywords:** Conformational analysis, peptides, aliphatic compounds, molecular dynamics, principal component analysis

# Literatura

- 1 Y. Mo, W. Wu, L. Song, M. Lin, Q. Zhang, J. Gao, (2004). 'The magnitude of hyperconjugation in ethane: a perspective from ab initio valence bond theory.' *Angewandte Chemie (International Ed. in English)*, **43**(15), 1986–1990.
- 2 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Ha-segawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, (2009) *Gaussian 09*, Revision D.02, Gaussian, Inc., Wallingford CT.
- 3 L. Verlet, (1968) 'Computer "Experiments" on Classical fluids. I. Thermodynamical Properties of Lennard-Jones Molecules.' *Phys. Rev.* **159**(1), 98-103.
- 4 W.C. Swope, H.C. Andersen, P.H. Berens, K.R. Wilson (1982) 'A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters.' *J. Chem. Phys.* **76** (1), 648(Appendix).
- 5 J.V. Neumann (1932) 'Proof of quasi-ergodic hypothesis' *Proc Natl Acad Sci USA* **18**(1): 70-82.
- 6 W. Svante, K I M Esbensen, P. Geladi, (1987) 'Principal Component Analysis' *Chemometr. Intell. Lab.* **2** : 37–52.
- 7 L.S.D. Caves, J.D. Evanseck, M. Karplus (1997) 'Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin' *Protein Sci* **7** : 649-666.
- 8 P. Geladi, B. R. Kowalski, (1986) 'Partial Least-Squares Regression: A Tutorial', *Analytica Chimica Acta* **185** : 1–17.
- 9 J.P. Stewart, (2007) 'Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application on 70 elements.' *J Mol Model* **13**:1173-1213.
- 10 G.N. Ramachandran, C. Ramakrishnan, V. Sasisekharan (1963) 'Stereochemistry of Polypeptide Chain Configuration' *J Mol Biol* **7** :95 -99.
- 11 L. Schäfer, C. Van Alsenoy, J.N. Scarsdale (1982) 'Ab Initio Studies of Structural Features Not Easily Amenable to Experiment. 23. Molecular Structures and Conformational Analysis of the Dipeptide N-acetyl-N'-methyl Glycyl Amide and the Significance of Local Geometries for Peptide Structures' *J Chem Phys* **76** :1439-1444.
- 12 S.C. Lovell; I.W. Davis, W.B. Arendall, P.I.W. De Bakker, J.M Word, M.G. Prisant, J.S. Richardson, D.C. Richardson (2003). 'Structure validation by C $\alpha$  geometry:  $\phi$ ,  $\psi$  and C $\beta$  deviation' *Proteins: Structure, Function, and Genetics* **50** (3): 437–450.
- 13 T.C. Cheam, S. Krimm (1988) 'Ab initio force fields of alanine dipeptide in C<sub>5</sub> and C<sub>7</sub> configurations' *J Mol Struct* , **188** :15-43.
- 14 I.R. Gould, W.D. Cornell, I.H. Hillier (1994) 'A Quantum Mechanical Investigation of the Conformational Energetics of the Alanine and Glycine Dipeptides in the Gas Phase and in Aqueous Solution', *J Am Chem Soc* **116** : 9250-9256.
- 15 M. Feig, (2008) 'Is Alanine Dipeptide a Good Model for Representing the Torsional Preferences of Protein Backbones?', *J Chem Theory Comput* **4** : 1555-1564.
- 16 G.J. Kleywegt, T.A. Jones (1996) 'Phi/psi-chology: Ramachandran revisited.' *Structure* **4** (12):1395-1400.