VILNIAUS UNIVERSITETAS

ANDRIUS MERKYS

KRISTALOGRAFINĖS INFORMACIJOS IŠGAVIMAS BEI PANAUDOJIMAS MOLEKULIŲ MODELIŲ TIKSLINIMUI IR TIKRINIMUI

Daktaro disertacijos santrauka
Technologiniai mokslai, chemijos inžinerija (05T)

Vilnius, 2018

Disertacija rengta 2013–2017 metais Vilniaus universitete.

Mokslinis vadovas – prof. dr. Saulius Gražulis (Vilniaus universitetas, technologiniai mokslai, chemijos inžinerija – 05T).

Disertacija ginama viešame disertacijos Gynimo tarybos posėdyje:

Pirmininkas – prof. dr. Rolandas Meškys (Vilniaus universitetas, technologiniai mokslai, chemijos inžinerija – 05T)
Nariai:
prof. habil. dr. Mindaugas Bloznelis (Vilniaus universitetas, fiziniai mokslai, matematika – 01P);
dr. Mindaugas Margelevičius (Vilniaus universitetas, technologiniai mokslai, chemijos inžinerija – 05T);
prof. dr. Peter Murray-Rust (Kembridžo universitetas, technologiniai mokslai, chemijos inžinerija – 05T);
dr. Česlovas Venclovas (Vilniaus universitetas, technologiniai mokslai, chemijos inžinerija – 05T).

Disertacija bus ginama viešame Gynimo tarybos posėdyje 2018 m. rugsėjo mėn. 18 d. 15 val. Vilniaus universiteto Gyvybės mokslų centro R 402 auditorijoje.

Adresas: Saulėtekio al. 7, Vilnius, Lietuva

Disertacijos santrauka išsiuntinėta 2018 m. rugpjūčio mėn. 18 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU internetinėje svetainėje adresu:
`https://www.vu.lt/naujienos/ivykiu-kalendorius`

VILNIUS UNIVERSITY

ANDRIUS MERKYS

EXTRACTION AND USAGE OF CRYSTALLOGRAPHIC KNOWLEDGE FOR REFINEMENT AND VALIDATION OF MOLECULAR MODELS

Summary of doctoral dissertation
Technological sciences, chemical engineering (05T)

Vilnius, 2018

The dissertation work was carried out at Vilnius University from 2013 to 2017.

Scientific supervisor – prof. dr. Saulius Gražulis (Vilnius University, technical sciences, chemical engineering – 05T).

The dissertation is defended at the public hearing of the Defence Board:

Chairman – prof. dr. Rolandas Meškys (Vilnius University, technical sciences, chemical engineering – 05T)
Members:
prof. habil. dr. Mindaugas Bloznelis (Vilnius University, physical sciences, mathematics – 01P);
dr. Mindaugas Margelevičius (Vilnius University, technical sciences, chemical engineering – 05T);
prof. dr. Peter Murray-Rust (University of Cambridge, technical sciences, chemical engineering – 05T);
dr. Česlovas Venclovas (Vilnius University, technical sciences, chemical engineering – 05T).

The dissertation will be defended at the public hearing of the Defence Board on the 18th of September, 2018 at 3 PM in room R 402 of Vilnius University Life Sciences Center.

Address: Saulėtekio al. 7, Vilnius, Lithuania

The summary of the dissertation is distributed on 18th of August, 2018.

The dissertation is available at the Vilnius University Library and at the VU website:
`https://www.vu.lt/naujienos/ivykiu-kalendorius`

Padėka

Esu labai dėkingas savo vadovui Sauliui Gražuliui už vadovavimą bei neįkainojamas įžvalgas bioinformatikos ir chemoinformatikos srityse bei pačiame moksle apskritai.

Dėkoju Vilniaus universiteto Biotechnologijos instituto Baltymų–nukleorūgščių sąveikos tyrimų skyriaus vedėjui Virginijui Šikšniui ir esamiems bei buvusiems skyriaus kolegoms, ypatingai Elenai Manakovai, Justui Butkui, Antanui Vaitkui ir Algirdui Grybauskui. Taip pat esu dėkingas Nicola Marzari, Giovanni Pizzi, Nicolas Mounet, Ivano E. Castelli, Andrea Cepellotti, Marco Gibertini, Philippe Schwaller, Miguel Quirós Olozábal, Aleksandrui Konovalovui, Garib N. Murshudov, Peter Murray-Rust, Fei Long bei Robert A. Nicholls. Taip pat dėkoju disertacijos recenzentams Mindaugui Blozneliui ir Kliment Olechnovič, pateikusiems labai naudingų patarimų ir komentarų.

Ypatingai dėkoju savo žmonai Miglei, tėvams, seserims ir seneliams už jų begalinę kantrybę bei palaikymą. Taip pat ačiū draugams, kurie visad buvo pasiruošę padėti.

Šis tyrimas buvo iš dalies finansuotas Lietuvos mokslo tarybos granto Nr. MIP-025/2013, SCIEX mokslinių mainų programos paramos stažuotei Nr. 13.169, Šveicarijos nacionalinio mokslo fondo paramos MARVEL Nacionaliniam tyrimų kompetencijos centrui bei Europos Sąjungos Horizon 2020 tyrimų bei inovacijos programos granto Nr. 689868.

Turinys

Skyrius 1

Įvadas

Žinios apie erdvines atomines kristalų struktūras nuo XX a. pradžios lėmė precedento neturinčius atradimus. Pirmųjų organinių junginių (XX a. ketvirtasis dešimtmetis), mioglobino (Kendrew, 1958 m., apdovanotas 1962 m. Nobelio premija), DNR (Franklin, Wilkins, Watson & Crick, 1952-1954 m., apdovanoti 1962 m. Nobelio premija) ir ribosomos (Ramakrishnan, Steitz & Yonath, XXI a. pirmasis dešimtmetis, apdovanoti 2009 m. Nobelio premija) erdvinių struktūrų nustatymai suteikė naujos informacijos apie pagrindinių gyvybės egzistavimui būtinų molekulių struktūrą bei funkcijas. Visi šie proveržį sukėlę tyrimai buvo atlikti remiantis nagrinėjamų molekulių kristalografija, kuri pateikia matematinę metodologiją, leidžiančią susieti nuo kristalų atspindėtus atspindžius su kristalų atomų erdviniu išsidėstymu [1].

Plačiausiai erdvinėms kristalų struktūroms nustatyti naudojama Rentgeno spindulių kristalografija. Rentgenostruktūrinės analizės eksperimento metu gaunamų duomenų dažniausiai nepakanka nepriklausomai nustatyti visiems erdvinės struktūros parametrams, įskaitant erdvines atomų koordinates. Todėl makromolekulių, turinčių keliomis dydžio eilėmis daugiau atomų nei mažos molekulės, struktūrų nustatymui naudojamos papildomos geometrijos žinios, dažniausiai gaunamos iš analogiškų mažų molekulių. Šios geometrinės žinios įtraukiamos kaip papildomi stebiniai, arba pasinaudojant jomis yra nustatomi sąryšiai tarp modelio dalių, tokiu būdu sumažinant modelio parametrų skaičių. Tiek papildomi stebiniai, tiek sąryšiai dažniausiai pritaikomi tarpatominių jungčių ilgiams, jungčių kampams arba dvisieniams kampams. Be to, tam tikra atomų grupė gali būti aprašoma kaip esanti vienoje plokštumoje arba kaip kietas kūnas, išlaikantis nekintamus tarpatominius atstumus ir kampus tarp grupės narių. Be abejo, tokiais būdais įvedama papildoma geometrinė informacija turi būti išvesta iš aukščiausios kokybės [2, 3, 4] bei didelio panašumo cheminių junginių [5, 6, 7, 3, 8, p. 221–250]. Abiejų reikalavimų įgyvendinimas nėra trivialus.

Nustatomų mažų molekulių kiekis auga kiekvienais metais [9]. Įrašų skaičius Kembridžo kristalografinių duomenų centro (CCDC) vystomoje Kembridžo struktūrinėje duomenų bazėje (CSD), didžiausiame mažų molekulių struktūrų archyve, per pastarąjį dešimtmetį padvigubėjo. Skaičiuojama, jog ši duomenų bazė kasmet pasipildo virš 50 000 naujų struktūrų[1]. Kadangi šis skaičius gerokai viršija srities specialistų bei žurnalų recenzentų kiekį [10], pasitaiko atvejų, kai prastos kokybės ar klaidingos struktūros paskelbiamos mokslinėje

[1] `https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/`, pasiekta 2017-07-12

literatūroje [11]. Tokioms struktūroms automatiškai aptikti pasitelkiama programinė įranga, leidžianti lyginamuoju būdu nustatyti neįprastas nagrinėjamų modelių savybes. Lyginimui dažniausiai naudojamos ekspertų sudarytos geometrinių žinių bibliotekos. Pastaruoju metu mokslinėje literatūroje skelbta apie bandymus bibliotekas generuoti automatiškai [7, 12], dažniausiai duomenų šaltiniu pasirenkant CSD [13]. Tačiau rezultatams, gautiems naudojantis CSD, yra taikomi CSD licenzijos apribojimai, draudžiantys rezultatus atvirai naudoti ir platinti[2] [14].

Darbo tikslai

- Sukurti metodiką ir programinę įrangą automatiniam geometrijos parametrų išgavimui iš mažų molekulių kristalų struktūrų. Panaudojus šią programinę įrangą surinkti geometrinius parametrus iš Atviros mažų molekulių kristalografinės duomenų bazės (COD[3] [15]) struktūrų.
- Sukurti metodiką bei programinę įrangą automatiniam geometrinių žinių bibliotekos konstravimui. Panaudoti programinę įrangą surinktų geometrinių duomenų organizavimui bei aprašymui.
- Sukurti metodiką ir programinę įrangą mažų molekulių kristalų struktūrų modeliams validuoti panaudojant sukonstruotą geometrinių žinių biblioteką.

Mokslinis naujumas, rezultatai bei jų reikšmė

Dažniausiai mažų molekulių geometrinių žinių bibliotekos yra konstruojamos panaudojant CSD duomenis bei programinę įrangą. Kadangi CSD yra komercinė duomenų bazė, jos duomenims, programinei įrangai bei išvestiniams rezultatams taikoma licenzija, draudžianti laisvą naudojimą bei viešą skelbimą. Šiame tyrime sukūrėme atvirą programinę įrangą (išleista GNU GPL2 ar kitomis suderinamomis licenzijomis) bei ja apdorojome duomenis iš viešo naudojimo COD duomenų bazės, kuri savo turiniui netaiko jokių apribojimų. CSD pakeitimas COD leidžia neribotą išgautos informacijos bei sudarytos geometrinių žinių bibliotekos duomenų naudojimą.

Daugumoje geometrijos bibliotekų laikomasi prielaidos, jog kiekvienas geometrinis parametras yra statistiškai pasiskirstęs pagal normalųjį skirstinį. Tačiau beveik visi srities tyrėjai pažymi, jog yra stebėję asimetrinius, daugiamodalinius bei kitokius nuo normaliojo modelio nutolusius pasiskirstymus. Šiame tyrime normalusis modelis pakeistas modos-skalės šeimos (angl. *location-scale family*) skirstinių mišiniu. Šis pakeitimas mums leidžia lanksčiai aprašyti visas minėtas pasiskirstymų rūšis.

Vadovaujantis požiūriu, jog geometrinis parametras pasiskirstęs pagal normalųjį skirstinį, neįprastų jo reikšmių, arba išskirčių (angl. *outlier*), paieškai (validavimui) dažniausiai naudojamas nuotolio nuo vidurkio matas, pavyzdžiui, Z įvertis [16]. Turėdami skirstinių mišiniais aprašytus geometrinių parametrų pasiskirstymus išskirčių paieškai pritaikėme

[2] PURY licensing policy. `http://pury.ijs.si/beta_servers.html`, pasiekta 2017-07-12
[3] `http://www.crystallography.net/cod`

Bajesinius metodus, kurie, be kita ko, yra tinkami naudoti ir su parametrais, pasiskirsčiusiais pagal normalųjį skirstinį.

Sukurta programinė įranga tiek geometrijos išgavimui, tiek apibendrinimui, tiek išskirčių paieškai yra pilnai automatinė ir gali veikti neprižiūrima. Geometrijos išgavimo bei apibendrinimo programinė įranga yra paruošta automatiniam periodiniam geometrijos žinių bibliotekos atnaujinimui. Geometrijos validavimo sąsaja yra pateikta viešam naudojimui bei gali būti integruota į COD duomenų įkėlimo sąsają naujiems į duomenų bazę keliamiems duomenims tikrinti.

Ginamieji teiginiai

- Atvira kristalografinė mažų molekulių duomenų bazė gali būti naudojama kaip informacijos šaltinis mažų molekulių geometrijos žinių bibliotekos kūrimui.
- Sukurti metodai neprižiūrimam, pilnai automatiniam mažų molekulių duomenų išgavimui bei organizavimui yra tinkami kristalų geometrijos įvairovei bei ypatybėms aprašyti.
- Sukurta geometrijos biblioteka yra tinkama išskirčių aptikimui naudojant Bajesinius metodus.

Skyrius 2

Tyrimų metodika

2.1 Kristalografinės informacijos išgavimas

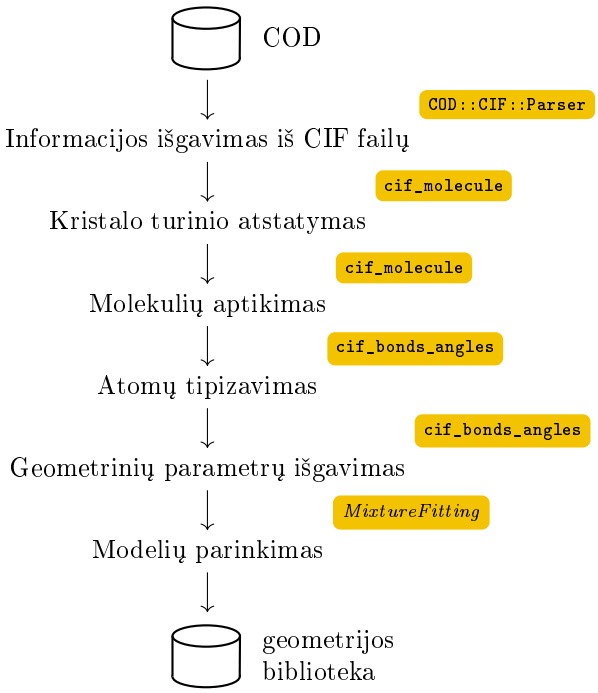
Pradinis geometrijos žinių bibliotekos konstravimo žingsnis (viso proceso schema pateikiama 2.1 pav.) yra informacijos išgavimas iš COD duomenų failų. Mažų molekulių kristalografinei informacijai saugoti ir platinti dažniausiai naudojamas kristalografinės informacijos formatas CIF 1.1 [17]. Šio formato gramatikai reikalingas specifinis sintaksinis analizatorius. Kadangi universalaus, aktyviai vystomo ir pagal nemokamo kodo licenzijas platinamo analizatoriaus CIF formatui nebuvo, sukūrėme `COD::CIF::Parser` sintaksinį analizatorių `Perl` [18] programavimo kalba, pasižymintį minėtomis savybėmis. Be to, mūsų sukurtas analizatorius geba aptikti bei pataisyti dažniausiai CIF failuose pasitaikančias klaidas [19].

2.2 Kristalo atstatymas

2.2.1 Elementaraus narvelio turinio atstatymas

Molekulių kristalai yra sudaryti iš pasikartojančių (bendru atveju) vienodų pasviręjų gretasienių, todėl viename CIF faile iš esmės aprašoma vieno tokio gretasienio, vadinamo elementariuoju narveliu, sudėtis: išvardinami narvelio atomai, pažymint kiekvieno atomo cheminį elementą bei erdvines koordinates. CIF failuose elementarieji narveliai aprašomi jų simetriją redukavus iki asimetrinio vieneto: mažiausio įmanomo atomų rinkinio, iš kurio kristalo simetrijos operatoriais galima atstatyti visą elementariojo narvelio turinį [8, p. 20]. Failuose taip pat pateikiami atstatymui reikalingi simetrijos operatoriai arba juos vienareikšmiškai nusakanti kristalo simetrijos grupė. Iš šio aprašo atstatyti visoms kristalo molekulėms, t.y., jungiems atomų dariniams, `cod-tools` paketo `cif_molecule` programoje įgyvendintas toks algoritmas:

1. Kiekvienas simetrijos grupės operatorius pritaikomas kiekvienam asimetrinio vieneto atomui.
   Sugeneruoti atomų atvaizdai perkeliami į elementarųjį narvelį (trupmeninės koordinatės $[0..1), [0..1), [0..1)$) atimant jų trupmeninių koordinačių sveikąsias dalis. Kiekvienam atomo atvaizdui priskiriamas unikalus identifikatorius „cell_label“. Sukuriamas iš pradžių tuščias molekulių konstravimui jau panaudotų „cell_label“ sąrašas.



Pav. 2.1: Geometrijos žinių bibliotekos konstravimo iš COD duomenų schema. Oranžiniame fone ties kiekvienu procesu nurodyta jį atliekanti programinė įranga.

2. Kaimynų (atomų porų, susietų cheminėmis jungtimis) paieškai atlikti visi praeitame žingsnyje sugeneruoti atomai pastumiami sukuriant $3 \times 3 \times 3$ dydžio supernarvelį (angl. *supercell*), kuris reikalingas cheminėms jungtims, kertančioms elementariojo narvelio ribas, nustatyti. Kaimynų paieškai pagreitinti supernarvelis suskaidomas į kubines dėžutes, kurių kraštinės lygios ilgiausioms galimoms kovalentinėms jungtims kristale. Tokiu būdu kiekvieno atomo kaimynų ieškoma tik aplinkinėse 27 dėžutėse, tad algoritmo sudėtingumas sumažėja iki tiesinio tuo atveju, kai atomų tankis yra pastovus [20]. Laikoma, kad visos cheminės jungtys kristale yra trumpesnės už kiekvieną elementariojo narvelio kraštinę.

3. Atomo atvaizdas, kurio „cell_label“ dar nebuvo panaudotas molekulių konstravime, yra paimamas naujai molekulei generuoti. Prie šio atomo prijungiami jo kaimynai, kaimynų kaimynai ir t.t. Gautą jungų grafą vadiname molekule. Kaimynų paieškai priėjus atomą už elementariojo narvelio krašto šio atomo koordinatės yra perkeliamos į elementarųjį narvelį atimant atomo trupmeninių koordinačių sveikąsias dalis. Kai molekulės grafo nebeįmanoma papildyti naujais atomais, laikoma, kad molekulė baigta, ir šis žingsnis kartojamas jau pradedant nuo kito nepanaudoto atomo. Tokių atomų nebelikus pereinama į kitą algoritmo etapą.

4. Praėjusiame etape atstatomos visos molekulės, tarp jų ir simetriškai ekvivalenčios (susietos kristalo simetrijos operatoriais), turinčios bent po vieną atomą atstatytame elementariajame narvelyje. Gautas molekules galima suskirstyti į grupes, į kurias pateks visos molekulės, esančios simetriškai ekvivalenčios viena kitai. Kadangi visos vienos grupės

molekulės yra sudarytos iš tų pačių asimetrinio vieneto atomų atvaizdų, visi šių molekulių atomai turės tuos pačius originalius identifikatorius (`_atom_site_label` CIF duomenų vardo reikšmes). Šiuos identifikatorius surūšiavus ir sujungus gauname grupės raktą K, kurį panaudodami kiekvieną molekulę priskiriame kuriai nors grupei. Padalinę kiekvienos grupės molekulių skaičių iš didžiausio bendro visų grupių dydžių daliklio D gauname stechiometriškai tikslią kristalo molekulių reprezentaciją.

5. Sukurtas aprašas nebūtinai yra minimalus, nes kristalo asimetriniame vienete gali būti daugiau nei viena chemiškai identiška, tačiau simetriškai nepriklausoma molekulė. Tokių dublikatų paieškai ir pašalinimui gali būti naudojami kiti K raktų sudarymo būdai, pavyzdžiui, Morgan metodas [21], kuris leidžia aptikti izomorfinius grafus. Šį metodą įgyvendinome `cif_molecule` kaip galimą pasirinkimą, tačiau šiame tyrime naudojome aukščiau aprašytus iš identifikatorių sujungtus raktus.

Molekules, cheminėmis jungtimis susijungusias su savo postūmio atvaizdais kituose kristalo narveliuose, vadiname polimerinėmis. Tokios molekulės teoriškai yra begalinės ir pasižymi periodiškumu. Paprastumo dėlei mūsų algoritmas polimerines molekules atkuria „perkirptas“ ties elementariųjų narvelių kraštais.

2.2.2 Cheminių jungčių nustatymas

Cheminių jungčių nustatymui panaudojome 2008 m. CCDC atnaujintą [22] 1979 m. paskelbtą kovalentinių spindulių lentelę [23, 24]. Šis parametrų rinkinys pasirinktas vietoje naujesnių rinkinių (pavyzdžiui, Pyykkö & Atsumi [25]) dėl leidžiamų ilgesnių cheminių jungčių ilgių periodinės elementų lentelės d ir s blokų elementams. Jungtims polimerinių molekulių turinčiuose kristaluose nustatyti mūsų programinė įranga sukuria supernarvelį iš $9 \times 9 \times 9$ elementarių kristalo narvelių.

Fiziškai neįmanomi tarpatominiai atstumai kristalų struktūrose pasitaiko dėl nesužymėtų alternatyvių atomų pozicijų arba kitų modeliavimo klaidų. `cif_molecule` praneša apie visus tarpatominius atstumus, trumpesnius už 0,75 kovalentinių spindulių sumos. Tokius atstumus nusprendėme visgi laikyti cheminėmis jungtimis ir jų įtaką rezultatams įvertinti vėliau.

2.2.3 Alternatyvios atomų pozicijos

Kartais molekulės ar jų dalys skirtinguose to paties kristalo narveliuose užima skirtingas diskrečias pozicijas, dar vadinamas alternatyviomis pozicijomis. Tokios molekulės ar jų dalys dažniausiai aprašomos molekulės fragmentais (angl. *disorder assembly*), turinčiais alternatyvias pozicijas (angl. *disorder group*). Atomų pora, kurios atomai priklauso to paties fragmento alternatyvioms pozicijoms, mūsų algoritmo nelaikoma sujungta chemine jungtimi, net jei atstumas tarp atomų ir būtų pripažintas tinkamu jungčiai.

Alternatyvomis taip pat aprašomos pozicijos kristaluose, kurias gali užimti daugiau nei vieno cheminio elemento atomai. Tačiau dažnai tokiuose aprašuose tą pačią erdvės poziciją užimantys atomai nebūna pažymėti kaip alternatyvūs. Tokioms situacijoms aptikti ir pažymėti sukūrėme programą `cif_mark_disorder`. Galimi cheminio elemento pasikeitimai lemia aplinkinių atomų

tipus, ko dabartinė mūsų sistema nėra pajėgi aprašyti. Šiame tyrime tokios kristalų struktūros yra automatiškai aptinkamos ir praleidžiamos.

2.2.4 Apribojimai

Kadangi polimerinių arba dideliu jungumu pasižyminčių kristalų struktūrų apdorojimas reikalauja daug laiko ir operatyviosios atminties, nutarėme kristalų atstatymo procesą apriboti. `cif_molecule` leidžia nustatyti didžiausią leidžiamą atomo pasikartojimo polimere skaičių, jį mes apribojome iki 100. Kai šis skaičius pasiekiamas, mažinamas polimero supernarvelis (plačiau 2.3.3 skyriuje). Taip pat apribojome `cif_molecule` proceso trukmę iki 600 sekundžių procesoriaus darbo laiko bei virtualią atmintį iki 1 GB. Užsiblokavę procesai, nenaudojantys nei procesoriaus laiko, nei atminties, yra šalinami po valandos nuo paleidimo pradžios.

2.2.5 Stebiniai iš simetrinių ekvivalentų

Mažų molekulių kristalai dažnai pasižymi aukšta simetrija. Dėl to vieną kartą nepriklausomai pamatuotas geometrinis stebinys molekulės modelyje gali pasikartoti daug kartų. Stebinius, kurie yra vienas kito atvaizdai siejami kristalo simetrijos operatorių, vadiname simetriniais ekvivalentais arba priklausomais stebiniais. Simetriniams ekvivalentams aptikti ir atmesti pritaikėme tokį algoritmą:

1. Kiekvienam kristalo asimetrinio vieneto atomui x_i pritaikomas kiekvienas iš kristalo simetrijos operatorių $s_1, s_2, ..., s_m$, tokiu būdu gaunant atomo atvaizdą $x_{i,j}$ su panaudotų simetrijos operatorių aibe $S_{i,j} = \{s_j\}$.

2. Atomo x_i vaizdai x_{i,j_1} ir x_{i,j_2}, kurie užima tą patį erdvės tašką (užima „specialiąją poziciją“), yra sujungiami į vieną; jų simetrijos operatorių aibės apjungiamos: $S_{i,\{j_1,j_2\}} = S_{i,j_1} \cup S_{i,j_2}$.

3. Sukuriama tuščia aibė B_{i_1,i_2} jungtims tarp atomų x_{i_1} ir x_{i_2} atvaizdų laikyti.

4. Aptikus jungtį tarp atvaizdų x_{i_1,j_1} ir x_{i_2,j_2}, aibėse S_{i_1,j_1} ir S_{i_2,j_2} ieškoma bendro simetrijos operatoriaus, kuris susietų kurios nors iš B_{i_1,i_2} jungčių atomų atvaizdus su x_{i_1,j_1} ir x_{i_2,j_2}. Jei toks operatorius surandamas, jungtis tarp x_{i_1,j_1} ir x_{i_2,j_2} traktuojama kaip jau stebėta. Jei ne, ji laikoma nauja ir įtraukiama į B_{i_1,i_2}.

Šis algoritmas išplečiamas jungčių (trys atomai) bei dvisieniams (keturi atomai) kampams.

2.3 Atomų tipai

Atomų cheminei apsupčiai klasifikuoti sukūrėme atomų tipizavimo sistemą, panašią į kitas sistemas, naudojančias kaimyninių ryšių, plokštumo bei dalyvavimo žieduose informaciją [26, 27, 28, 29]. Sistema kiekvienam atomui priskiria tipą – tekstinę eilutę, kurioje prefiksine tvarka rekursiškai išvardinami atomo kaimynai bei kaimynų kaimynai. Šiame tyrime naudojamas apribojimas, vadinamas klasifikavimo gyliu, kuris nusako, jog kaimynų vardinimo rekursija

sustoja pasiekusi nagrinėjamo atomo kaimynų kaimynus, tačiau klasifikavimo gylis gali būti padidintas.

Atomo tipo identifikatorius pradedamas klasifikuojamojo atomo cheminio elemento žymeniu. Jei atomas yra plokščioje apsuptyje (plačiau 2.3.1 skyriuje), jo cheminis elementas rašomas pradedant mažąja raide, jei ne – pradedant didžiąja. Jei atomas dalyvauja bent viename žiede, toliau laužtiniuose skliaustuose įrašoma narystės žieduose informacija (plačiau 2.3.2 skyriuje), ši informacija įtraukiama tik nagrinėjamam atomui bei jo tiesioginiams kaimynams. Toliau identifikatoriuje rekursiškai išvardinami tiesioginių kaimynų atomų tipai. Vienodi kaimynų tipai sutraukiami.

Jungtys, kampai bei dvisieniai kampai skirstomi į klases pagal juos sudarančių atomų tipus. Šioms klasėms nusakyti atitinkamai naudojami du, trys bei keturi atomų tipai.

2.3.1 Plokštùmo nustatymas

Tris ar daugiau chemines jungtis turintis atomas šiame tyrime laikomas esantis plokščioje aplinkoje, jei kiekvieno šių jungčių vektorių trejeto $\vec{v}_1$, $\vec{v}_2$ ir $\vec{v}_3$ aprašomo gretasienio tūrio santykis su vektorių ilgių sandauga yra mažesnis už 0,1:

$$\frac{|\vec{v}_1(\vec{v}_2 \times \vec{v}_3)|}{||\vec{v}_1|| \times ||\vec{v}_2|| \times ||\vec{v}_3||} < 0{,}1. \tag{2.1}$$

Jungčių tarpusavio kampams esant vienodiems, ši sąlyga yra tenkinama, kai kampų dydžiai yra lygūs $\sim 119{,}9^\circ$ arba didesni.

Cheminių apsupčių, kurių plokštùmas priklauso nuo atomų įgyjamų alternatyvių konformacijų, mūsų sukurta programinė įranga apdoroti negali. Tokias apsuptis turinčios kristalinės struktūros yra praleidžiamos.

2.3.2 Žiedų paieška

Kiekvieno atomo tipe laužtiniuose skliaustuose įrašomas žiedų, kuriuose atomas dalyvauja, dydis ir kiekis. Žiedais laikomos bestygės (angl. *chordless*) ciklinės atomų grandinės be pasikartojimų, turinčios ne daugiau nei 7 atomus. Žiedų paieška atliekama modifikuotu Downs ir k.t. [30] bei kitų autorių aprašomu algoritmu [31, 32, 33]. Paieškos į gylį būdu surandami visi keliai, jungiantys nagrinėjamą atomą su juo pačiu. Algoritmas iš šios kelių aibės pašalina kelius, kurie apima visus kurio nors kito, trumpesnio kelio atomus. Tokiu būdu gaunamas mažiausias nagrinėjamo atomo mažiausių žiedų rinkinys (angl. *smallest set of smallest rings*) [34].

2.3.3 Polimerinių struktūrų apdorojimas

Kaip minėta, mūsų algoritmas polimerines molekules atkuria „perkirptas“ ties elementariųjų narvelių kraštais. Tam, kad būtų teisingai nustatomi ties kirpimo vietomis esančių atomų tipai, konstruojamas $9 \times 9 \times 9$ supernarvelis. Jei supernarvelio konstravimo metu viršijamas maksimalus polimerinių atomų skaičius (plačiau 2.2.4 skyriuje), supernarvelio dydis mažinamas dviem narveliais (nuo $9 \times 9 \times 9$ iki $7 \times 7 \times 7$). Procedūra gali būti kartojama iki tol, kol supernarvelis susilygina su elementariuoju narveliu.

2.4 Geometriniai matavimai

Šiame tyrime jungties ilgiu laikomas atstumas tarp atomų Dekarto koordinačių. Kampas α tarp jungčių vektorių $\vec{a}$ ir $\vec{b}$ skaičiuojamas taip:

$$c = \cos\alpha = \frac{\vec{a}\cdot\vec{b}}{||\vec{a}||\ ||\vec{b}||} \tag{2.2}$$

$$\alpha = \text{atan2}(\sqrt{1-c^2}, c), \tag{2.3}$$

kur atan2 funkcija atitinka `C` ir `Perl` programavimo kalbų apibrėžimą. A–B–C–D atomų sekoje dvisienis kampas ϕ skaičiuojamas tarp plokštumų (A, B, C) ir (B, C, D), ir įgyja reikšmes iš intervalo $[0, 360°)$. Kampui tarp jungčių vektorių $\vec{a}$, $\vec{b}$ ir $\vec{c}$, $\vec{d}$ naudojamos šios formulės:

$$c = \cos\phi = \frac{(\vec{a}\times\vec{b})\cdot(\vec{c}\times\vec{d})}{||\vec{a}\times\vec{b}||\ ||\vec{c}\times\vec{d}||} \tag{2.4}$$

$$\phi = \text{atan2}(\sqrt{1-c^2}, c). \tag{2.5}$$

Jei minėti vektoriai sudaro kairės rankos sistemą, kampo dydis ϕ atimamas iš 360° [8, p. 205–219].

2.5 Statistiniai modeliai

2.5.1 Skirstiniai

Tyrimo metu laikome, jog stebiniai – tarpatominių jungčių ilgiai, kampai bei dvisieniai kampai – yra nepriklausomi ir vienodai pasiskirstę. Todėl imties vektoriaus $\vec{x}$ tikėtinumas pagal modelį M su parametrų vektoriumi $\vec{\theta}$ yra

$$p(\vec{x}|\vec{\theta}, M) = \prod_{j=1}^{n} p(x_j|\vec{\theta}, M), \tag{2.6}$$

kur n yra $\vec{x}$ ilgis. Jungčių bei kampų imtims aprašyti pasirinkome normalųjį bei Koši (pranc. *Cauchy*) skirstinius, kadangi preliminarios analizės metu pastebėjome kelis leptokurtinius pasiskirstymus (t.y., stebėtas imties ekscesas buvo didesnis už normaliojo skirstinio). Dvisieniams kampams aprašyti pasirinkome von Mises skirstinį, kuris yra skirtas perteikti pasiskirstymams intervale $[0, 2\pi)$.

2.5.2 Modelio parinkimas

Naudodami tikėtinumo maksimizavimo algoritmą kiekvienai stebinių imčiai parinkome po dešimt kiekvieno skirstinio modelių su komponentų skaičiumi nuo 1 iki 10. Tinkamiausiu imčiai aprašyti parinkome modelį su mažiausia Bajesinio informacijos koeficiento (BIC [35]) verte:

$$BIC = -2\log p(\vec{x}|\vec{\theta}) + (3m-1)\log n, \tag{2.7}$$

kur m yra modelio komponentų skaičius.

Tikėtinumo maksimizavimo algoritmui parinkome šias pradines reikšmes pagal Bohning ir k.t. [36]:

$$A_j(0) = \frac{1}{m} \tag{2.8}$$

$$\mu_j(0) = \min(x) + j\frac{\max(x) - \min(x)}{m+1} \tag{2.9}$$

$$\sigma_j(0) = \frac{\max(x) - \min(x)}{6(m+1)}, \forall i \in [1, ..., m], \tag{2.10}$$

kur $A_j(0)$ – j-ojo komponento dalis mišinyje ($\sum_{i=1}^{m} A_i = 1$), $\mu_j(0)$ – komponento moda, $\sigma_j(0)$ – koncentraciją reguliuojantis parametras. Tikėtinumo maksimizavimo algoritmui Koši skirstinio atveju pritaikėme iteratyvų parametrų parinkimą pagal Nagy [37] su 20 iteracijų. Algoritmo konvergencijos normaliajam bei Koši skirstiniams kriterijumi pasirinkome skirtumą tarp dviejų iš eilės einančių iteracijų parametrų. Algoritmas stabdomas kai $\forall i : |\theta_i^{t+1} - \theta_i^t| < \epsilon$, $\epsilon = 10^{-6}$. Konvergencijos kriterijumi von Mises atveju pasirinkome skirtumą tarp imties logaritmuoto tikėtinumo pagal modelį su parinktais parametrais, kaip yra siūloma Hornik & Grün [38].

Modelių parinkimo algoritmai įgyvendinti `R` programavimo kalba [39] programinės įrangos pakete `MixtureFitting`[1]. Algoritmų paspartinimui dalis paprogramių perrašyta `C` programavimo kalba.

2.5.3 Išskirčių paieška

Parinkti modeliai naudojami naujų stebinių vertinimui siekiant nustatyti ar stebinys atitinka modelį (nulinė hipotezė arba H_0), ar yra išskirtis, t.y., yra kilęs iš tolydžiojo skirstinio (alternatyvi hipotezė arba H_1) [40]. Stebinio atitikimas kiekvienai iš hipotezių yra vertinamas pagal Bajeso faktorių, kuris atitinka tikėtinumų santykį

$$K = \frac{P(H_0|x)}{P(H_1|x)}, \tag{2.11}$$

kai abiejų hipotezių tikimybė yra vienoda. Jeffreys [41, p. 432] K reikšmę siūlo vertinti pagal šias kategorijas:

- 0 laipsnis. $K > 1$. Duomenys atitinka H_0
- 1 laipsnis. $1 > K > 10^{-0,5}$. Įrodymų prieš H_0 yra, tačiau jie verti tik paminėjimo.
- 2 laipsnis. $10^{-0,5} > K > 10^{-1}$. Įrodymai prieš H_0 tvirti.
- 3 laipsnis. $10^{-1} > K > 10^{-1,5}$. Įrodymai prieš H_0 stiprūs.
- 4 laipsnis. $10^{-1,5} > K > 10^{-2}$. Įrodymai prieš H_0 labai stiprūs.
- 5 laipsnis. $10^{-2} > K$. Įrodymai prieš H_0 neabejotini.

[1] Platinama su GNU GPL2 atviro kodo licenzija `https://github.com/merkys/MixtureFitting`, šiame tyrime naudota 0.1.0 paketo versija (132 revizija)

Šiame tyrime laikome, jog tikėtinumų santykis $K < 0{,}1$ yra pakankamas H_0 hipotezei atmesti. Tokiu atveju laikoma, jog nagrinėjamas stebinys yra išskirtis.

Bajeso faktoriaus apskaičiavimui naudojame Schwarz kriterijų:

$$S = \log P(\text{duomenys}|\hat{\theta}_0, H_0) - \log P(\text{duomenys}|\hat{\theta}_1, H_1) - \frac{1}{2}(d_0 - d_1)\log n, \tag{2.12}$$

kur $\hat{\theta}_k$ yra didžiausio tikėtinumo H_k parametrai, d_k yra θ_k dimensija, o n yra imties dydis [42].

2.6 Tinklo sąsaja

Molekulių geometrijos naršymui buvo sukurta tinklinė naudotojo sąsaja, paremta `Perl` kalba sukurtomis programomis[2]. Sąsaja susideda iš paieškos[3], geometrijos peržiūros bei molekulių validavimo[4].

2.6.1 Paieškos sąsaja

Įgyvendinti du cheminės apsupties paieškos būdai, grafinis ir tekstinis. Grafinėje sąsajoje pateikiamas `JSME` [43] `JavaScript` įskiepis, kuriame naudotojo nupieštai molekulei vėliau yra priskiriami atomų tipai. Įskiepis įvestą struktūrinę formulę verčia molekulių grafams tekstinėse eilutėse atvaizduoti skirtu SMILES formatu [44], pagal kurį `Open Babel` [45] paketas konstruoja molekulinį grafą, iš kurio mūsų programinė įranga nustato atomų tipus. Aromatiniai atomai, turintys tris ar daugiau kaimynų, laikomi plokščiais. Pasirinkęs du, tris ar keturis atomų tipus naudotojas nukreipiamas į atitinkamai jungčių ilgių, kampų bei dvisienių kampų dydžių pasiskirstymo peržvalgą. Tekstinė paieška leidžia paiešką pagal įvestus atomų tipus. Taip pat galima pamatyti visų įvestos molekulės parametrų pasiskirstymus.

2.6.2 Molekulinės geometrijos naršyklė

Kiekvieno parametro pasiskirstymo peržiūra įgyvendinta interaktyvia histograma. Sąsajoje taip pat pateikiami visi pasiskirstymui 2.5.2 skyriuje aprašyta metodika parinkti mišinių modeliai. Šiuos modelius galima pavaizduoti ant histogramos. Pasirinkus histogramos stulpelį sąsajoje pateikiamas jį sudarančių stebinių COD sąrašas, čia pat `Jmol` [46] įskiepyje galima pamatyti ir paženklintą struktūros dalį, iš kurios stebinys yra kilęs.

2.6.3 Validavimo sąsaja

Geometrijos validavimui bei tuštumų paieškai sukurta sąsaja, į kurią įkeltuose kristalų struktūrų CIF failuose atliekama geometrinių parametrų išskirčių bei tuštumų paieška. Neįprasti parametrai bei tuštumos pavaizduojamos `Jmol` įskiepyje, taip pat pateikiamos kiekvieno iš neįprastų parametrų histogramos. Sąsaja taip pat praneša apie nematytus atomų tipus, fragmentų klases bei pernelyg trumpus tarpatominius atstumus.

[2] Išeities kodas platinamas su GNU GPL2 atviro kodo licenzija `svn://www.crystallography.net/molecules-in-COD/trunk`, šiame tyrime aprašoma 1499 revizija

[3] `http://www.crystallography.net/geometry/`

[4] `http://www.crystallography.net/geometry/cgi-bin/check_geometry.pl`

Skyrius 3

Rezultatai ir jų aptarimas

3.1 CIF sintaksinis analizatorius

3.1.1 Apžvalga

CIF 1.1 formatui įskaityti sukūrėme sintaksinį analizatorių `COD::CIF::Parser` `Perl` programavimo kalba. Tam, kad analizatorius gebėtų kai kuriuos netaisyklingus CIF failus pataisyti, CIF formato gramatiką papildėme keliomis klaidas taisančiomis euristikomis, tačiau šios yra taikomos tik naudotojui išreikštai nurodžius **fix_errors** opciją (toliau šiame tekste analizatorių su įjungtu klaidų taisymo režimu žymėsime $\texttt{COD::CIF::Parser}_{\text{fix}}$). Be šio nurodymo sintaksinis analizatorius laikosi nustatytos CIF 1.1 formato gramatikos. Tiesa, `COD::CIF::Parser` reikalauja vienu aspektu mažiau, nei CIF 1.1 gramatika: mūsų analizatorius neriboja teksto eilučių ilgio. Pranešimai apie pernelyg ilgas eilutes gali būti įjungti opcija. Šį savo sprendimą grindžiame tuo, kad dauguma modernių programavimo kalbų gali apdoroti neriboto ilgio tekstines eilutes (mes naudojame `Perl` ir `C`, tačiau šia savybe pasižymi bent `Python`, `Java`, `Julia` programavimo kalbos). Be to, ilgio ribojimo įgyvendinimas reikalautų daugiau pastangų. Kadangi ilgio patikrinimas gali praversti pavyzdžiui prieš CIF failus apdorojant `Fortran` programomis, kurios turi fiksuoto ilgio skaitymo buferius, `COD::CIF::Parser` patikrinimo metu aptiktas pernelyg ilgas eilutes praneša. Nepaisant ribojimo nebuvimo failus įskaitant, mūsų programinė įranga paiso eilučių ilgio apribojimo rašydama CIF formato išvestį.

3.1.2 Formato atitikimas

Siekdami patikrinti, kaip mūsų sukurtas CIF formato sintaksinis analizatorius atitinka formato aprašymą, palyginome jo elgseną su kitais plačiai naudojamais atviro kodo analizatoriais naudodami tiek teisingas, tiek klaidingas įvestis[1]. Lyginimui naudoti sintaksiniai analizatoriai išvardinti 3.1 lentelėje. Kadangi CIF 1.1 formatas išvestas iš STAR 1 [58] formato, į palyginimą taip pat įtraukėme du pastarojo formato analizatorius. Palyginimo metu nenaudotos kitos nei numatytosios programinės įrangos opcijos, išskyrus mūsų analizatorius, kuriems išreikštai

[1] Testų įvestys ir rezultatai paskelbti `https://github.com/cod-developers/CIF-parsers`, šiame tyrime aprašoma `416da44` revizija

Analizatorius	versija	programavimo kalba	šaltinis
`ASE`	3.14.1	`Python`	[47]
`cif2cif`	2.0.0	`Fortran`	[48]
`cif_linguist`	0.4.2	`C`	[49]
`COD::CIF::Parser`	rev. 5518	`C`	[19]
$\texttt{COD::CIF::Parser}_{\text{fix}}$	rev. 5518	`C`	[19]
`gemmi`	rev. 860d285	`C++`	[50]
`PyCIFRW`	4.2	`Python`	[51]
`STAR::Parser`†	0.59	`Perl`	[52]
`StarTools`†	0.2.0	`Java`	[53]
`ucif`	rev. 23314	`C++`	[54]
`vcif`	1.2	`C`	[55]
`vcif2`	0.9.3.1	`C`	[56]
`ZINC`	1.12	`C`	[57]

Lentelė 3.1: Lyginimui naudoti CIF 1.1 sintaksiniai analizatoriai. Simboliu † pažymėti STAR formato sintaksiniai analizatoriai.

nustatyta opcija pranešti duomenų vardus bei eilutes, viršijančius ilgio apribojimus. Palyginimo rezultatai pateikiami 3.2 lentelėje. Nustatėme keturias galimas CIF failo įskaitymo baigtis: analizatoriaus klaida (programa nutraukė darbą arba aptikusi klaidingą sintaksę/semantiką, arba patekusi į būseną, iš kurios nežinota kaip išeiti); išvestas įspėjimas (aptikta klaida, tačiau programa darbą pratęsė); programa pakliuvo į begalinį ciklą; sėkminga failo analizė. Šiame tyrime nenagrinėjome analizatorių sukurtų vidinių duomenų struktūrų, tad negalime teigti, jog visi sintaksiniai analizatoriai teisingai įskaitė duomenis net ir pirmaisiais dviem atvejais.

Iš palyginimo matyti, jog dauguma analizatorių sugeba įskaityti korektiškus CIF failus. `ase` ir `ZINC` pasirodė nereiklūs, klaidas taisantys analizatoriai, ignoruojantys draudžiamus simbolius, trūkstamas uždarančias kabutes ar duomenų bloko pradžios žymes. Tačiau `ase` analizatorius nesugebėjo įskaityti dviejų teisingų testinių failų, o `ZINC` skaitydamas tekstinį lauką be uždarančio kabliataškio pakliuvo į begalinį ciklą. `cif2cif` aptinka ir praneša dalį sintaksės ir semantikos klaidų (tokių kaip per ilgos eilutės ar besikartojantys duomenų vardai), tačiau leidžia kai kuriuos vertėse draudžiamus simbolius. `cif_linguist` praneša visas neteisingas CIF konstrukcijas išskyrus `^Z` simbolį bei per ilgus duomenų vardus. Be to, keli sintaksiškai teisingi testiniai CIF failai programos buvo identifikuoti kaip klaidingi. `gemmi` analizatorius savo elgsena labai panašus į `cif_linguist`. `gemmi` aptinka `^Z`, apdoroja DOS operacinės sistemos naujos eilutės simbolius bei neriboja duomenų vardo bei eilutės ilgio. Be to, šis analizatorius netaiko kai kurių apribojimų CIF naudojamų simbolių aibei. `PyCIFRW` taip pat mažiau riboja simbolių aibę, tačiau praneša apie pernelyg ilgus duomenų vardus. Analizatorius leidžia CIF lenteles (angl. *loops*) be duomenų vardų ar verčių, taip pat nereikalauja tarpais nuo kitų duomenų vardų skirti tekstinius laukus. `ucif` aptinka daugumą klaidingų įvesčių, tačiau nepraneša apie tarpais neatskirtus duomenų laukus, pasikartojančius duomenų vardus pernelyg ir ilgus duomenų vardus bei eilutes. `vcif` analizatorius pasirodė jautresnis ribiniams įvesties atvejams už daugumą kitų analizatorių. Be daugelio klaidų `vcif` taip pat praneša apie tuščius failus ir duomenų blokus, per ilgas eilutes ir duomenų vardus. Duomenų vardai besiskiriantys registru analizatoriaus nelaikomi klaida, nors to reikalauja CIF standartas. Taip pat neaptinkami trūkstami tarpai po tekstinių

Testas	Atitinka reikalavimus?	ase	cif2cif	cif_linguist	COD::CIF::Parser	COD::CIF::Parser$_{\text{fix}}$	gemmi	PyCIFRW	STAR::Parser	StarTools	ucif	vcif	vcif2	zinc
ascii-127.cif	×			×	×	/	×				×	/	/	
byte-order-mark.cif	×			×			×	/			×	×	×	×
closing-bracket.cif	×			×	×			/		/	×			
comment-only.cif												/	×	
dos-ctrl-z.cif	×		×		×	/	×					×	×	×
duplicate-tags-different-cases.cif	×		/	×	×	×	×	×						
duplicate-tags-different-values.cif	×		/	×	×	×	×	×				×		
duplicate-tags-same-values.cif	×		/	×	×	/	×	×				×		
empty-datablock.cif												/	/	×
empty-datablock-name.cif	×			×	×	/	×	×		/	×	×	/	
empty-file.cif										×		/	×	
form-feed.cif	×		×	×			×	×					/	
global.cif	×		×	×	×	×	×	×			×			×
long-line.cif	×		/	×	/							/	/	
loop-without-tags.cif	×	/	×	×	×	×	×		–		×	×	×	×
loop-without-values.cif	×	/	×	×	×	×	×		–		×	×	×	×
missing-closing-quote.cif	×		/	×	×	/	×	×	/	/	×	×	/	
missing-data-header.cif	×		/	×	×	/	×	×			×	×	/	
non-ascii.cif	×			×	×	/	×				×	/	/	
non-ascii-in-comment.cif	×			×	/	/					×		/	
null-symbol.cif	×			×	×	×	×				×	/	×	
_refine_ls_extinction_expression.cif				×										
single-quote-in-value.cif														
stray-values-at-start.cif	×			×	×	/	×	×			×	×	×	×
tag-immediately-following-textfield.cif	×		×	×	×	×	×			/		×	×	
textfield-in-loop.cif		/												
textfield-no-closing-semicolon.cif	×			×	×	×	×	×	×	/	×	×	×	–
unquoted-loop-prefix.cif			×				×		×	/			×	×
value-immediately-following-textfield.cif	×			×	×	×	×			/			×	
value-starting-with-bracket.cif	×			×	×			/		/	×			
value-starting-with-closing-bracket.cif	×			×	×			/		/	×			
value-starting-with-dollar.cif	×			×	×	×	×	×			×			
vertical-tab.cif	×		×	×			×	×			×		/	
whitespace-placement.cif		/												
wrong-number-of-loop-values.cif	×	/	×	×	×	×	×	×	×		×	×	/	

Lentelė 3.2: CIF 1.1 sintaksinių analizatorių palyginimas. Kryželiais („×“) žymimi atvejai, kai analizatorius dėl klaidos nutraukė darbą, pasviraisiais brūkšniais („/“) – kai analizatorius išvedė pranešimą, brūkšniais („–“) – kai analizatorius pateko į begalinį ciklą.

laukų bei kai kurie vertėse draudžiami simboliai. `vcif2` įvesčiai kelia mažesnius reikalavimus nei `vcif`: pasikartojantys duomenų vardai išvis nelaikomi klaida, kaip ir tekstiniai laukai be uždarančiųjų kabliataškių ar po jų einančių tarpų. Nepaisant to, `vcif2` už savo pirmtaką geriau aptinka draudžiamus simbolius. Abu STAR formato sintaksiniai analizatoriai, `STAR::Parser` ir `StarTools`, yra mažiau linkę aptikti klaidas, tačiau tai gali būti susiję su tuo, kad kai kurie CIF formato apribojimai nėra taikomi STAR formatui. Nepaisant to, `STAR::Parser` pasirodė mažiausiai patikimas, kadangi pakliuvo į begalinius ciklus skaitydamas nepilnai aprašytas CIF lenteles.

Mūsų sintaksiniai analizatoriai savo elgsena panašūs į `cif_linguist` ir `gemmi`. Veikdamas griežtu režimu `COD::CIF::Parser` aptinka visas sintaksės ir semantikos klaidas išskyrus UTF-8 baitų tvarkos žymę (angl. *BOM*) bei porą CIF formate draudžiamų tuščių simbolių (angl. *white space*). Nors šie simboliai CIF 1.1 formate yra draudžiami, mes manome, jog dažniausiai jie bus per klaidą įterpiami teksto rengimo programine įranga, todėl juos praleisdami CIF failų prasmės neiškreipsime. Mažiau reiklus $\texttt{COD::CIF::Parser}_{\text{fix}}$ režimas apdoroja bei pataiso dalį neteisingų CIF failų ir praneša apie atliktus pataisymus. Šios savybės pasirodė labai naudingos įskaitant CIF formato failus su nedideliais nukrypimais nuo standarto, kurie pasitaiko net ir recenzuotų publikacijų prieduose. Galima daryti išvadą, kad CIF analizatorių elgsenos įvairovė išduoda skirtingus jų kūrėjų poreikius, o šios įvairovės buvimas leidžia lyginti skirtingus analizatorius ieškant klaidų bei pasirinkti analizatorių su norimu funkcionalumu bei savybėmis, tokiomis kaip programavimo kalba.

3.1.3 Našumas

Savo bei kitų sintaksinių analizatorių našumą palyginome analizuodami 382 807 CIF failų iš COD (visi duomenų bazės įrašai iš 199925 revizijos, iš viso ∼49 GB duomenų). Bandymai atlikti mažai apkrautame kompiuteryje su 31 GB operatyviosios atminties ir 16 × Intel(R) Xeon(R) CPU E5-2450 v2 @ 2,50 GHz procesorių, naudojančiame Debian GNU/Linux 8.6 (jessie) operacinę sistemą, `gcc` 4.9.2, `Perl` 5.20.2 bei `Python` 2.7.9 versijas. Bendros programų veikimo trukmės palyginimui pateiktos 3.3 lentelėje. Tyrime nenaudojome `PyCIFRW` ir `vcif2` sintaksinių analizatorių, kadangi pastebėjome, jog jų veikimo laikas turi kvadratinę priklausomybę nuo CIF tekstinių laukų apimties, veikiausiai dėl neefektyvaus atminties valdymo. Mūsų tyrimo rezultatai parodė, jog sukurtas CIF analizatorius yra vienas greičiausių bei yra tinkamas visoms formato konstrukcijoms perskaityti.

3.1.4 Išvados

CIF formato sintaksinis analizatorius buvo sukurtas `Perl` programavimo kalba, o vėliau optimizuotas `C` su sąsajomis `Perl` ir `Python` programavimo kalboms. Mūsų atlikti našumo testai parodė, jog sukurtas `COD::CIF::Parser` sintaksinis analizatorius yra vienas iš greičiausių CIF formato analizatorių. `COD::CIF::Parser` vystymui esminę įtaką turėjo kitų sintaksinių analizatorių lyginimas pasinaudojant mūsų sukurtu testų rinkiniu. Būtina pabrėžti, jog skirtingos analizatorių elgsenos nebūtinai kyla dėl programavimo klaidų, kadangi kiti analizatoriai gali įgyvendinti kitas kūrėjų norimas funkcijas. Mūsų sukurtas klaidas taisantis `COD::CIF::Parser` sintaksinis analizatorius ypač pravertė sintaksiškai nekorektiškų CIF failų

Analizatorius	Laikas (min.)
`ase`	90,69
`cif2cif`	31,54
`cif_linguist`	27,05
`COD::CIF::Parser`	25,53
`COD::CIF::Parser`$_{\text{fix}}$	16,07
`gemmi`	12,25
`ucif`	16,61
`vcif`	15,77
`zinc`	16,16

Lentelė 3.3: COD CIF failų įskaitymo skirtingais analizatoriais laikas (minutėmis)

iš išorinių šaltinių įskaitymui bei taisymui. Tokiu būdu mūsų sintaksinis analizatorius buvo naudingas didelių kristalografinės informacijos kiekių apdorojimui. `COD::CIF::Parser` pasižymi dideliu darbo našumu bei suderinamumu su programomis, sukurtomis `Perl`, `C` bei `Python` programavimo kalbomis, todėl tikimės, jog mūsų CIF sintaksinis analizatorius palengvins kristalografinių duomenų apykaitą tarp tyrėjų.

Lygiagrečiai įgyvendinome CIF 1.1 ir CIF 2.0 formatų sintaksinius analizatorius. Kiekvienai CIF formato versijai sukūrėme atskirą leksinį analizatorių bei gramatiką. Kadangi programinės įrangos vystymo bei palaikymo kaštai daugiau nei tiesiškai priklauso nuo jos apimties [59], dviejų sintaksinių analizatorių išlaikymas pareikalaus ženkliai daugiau pastangų nei vieno. Tačiau dabartinė situacija yra neišvengiama, kadangi CIF 1.1 ir CIF 2.0 formatai nėra tarpusavyje suderinami.

3.2 Geometrijos biblioteka

3.2.1 Apžvalga

Tyrimo metu apdorota 382 807 kristalų struktūrų iš COD duomenų bazės 199925 revizijos. Skaičiavimai atlikti 140 branduolių 240 × Intel(R) Xeon(R) CPU E5-4650 v2 @ 2,40 GHz procesorių bendros atminties kompiuteryje su 1,1 TB RAM, naudojančiame CentOS 6.8 operacinę sistemą, ir užtruko beveik dvi paras. Bent po vieną chemine jungtimi susietą atomų porą buvo išgauta iš maždaug 320 000 kristalų struktūrų. Likusios struktūros buvo arba be atomų, turėjo nejungius atomus, ar buvo praleistos. Siekdami geriau išnagrinėti duomenų apdorojimo dėsningumus, kiekvienam apdorotam COD įrašui automatiškai priskyrėme tam tikras žymes pagal `cif_molecule` ir `cif_bonds_angles` išvestis. Buvo apibrėžtos tokios žymės:

- `APDOROTA` – kristalo struktūra yra apdorota; ši žymė suteikiama kiekvienai apdorotai struktūrai nepriklausomai nuo jos savybių ar apdorojimo rezultatų;
- `TRUMPI ATSTUMAI` – struktūra turi bent vieną porą per arti esančių atomų;
- `CIF TUŠČIAS` – `cif_molecule` išvestis yra tuščia;
- `CIF NUTRAUKTAS` – `cif_molecule` procesas buvo nutrauktas dėl paskirto laiko ar operatyviosios atminties perviršijimo;

- `ALTERNATYVIOS KLASĖS` – struktūra turi bent vieną atomą, kurio tipas priklauso nuo alternatyvios pozicijos; šių struktūrų dabartinis mūsų algoritmas apdoroti negali, todėl jos yra praleidžiamos;
- `NETVARKA` – struktūra turi bent vieną atomą, esantį alternatyviose pozicijose;
- `NETVARKA SPEC. POZIC.` – struktūra turi bent vieną atomą, kurio alternatyvios pozicijos yra išsidėsčiusios aplink specialiąją poziciją;
- `DUBLIKATAS` – struktūros įrašas yra dublikatas; dublikatais pažymėtas COD struktūras mūsų metodai praleidžia;
- `NETVARKA PAŽYMĖTA` – struktūra turi bent vieną `cif_mark_disorder` programa nustatytą ir pažymėtą atomą, esantį alternatyviose pozicijose;
- `NĖRA ATOMŲ` – struktūra neturi nė vieno atomo;
- `POLIMERAS` – struktūra yra polimeras;
- `TAB TUŠČIAS` – `cif_bonds_angles` išvestis tuščia, priežastis žinoma;
- `TAB NUTRAUKTAS` – `cif_bonds_angles` išvestis tuščia, priežastis nenurodyta;
- `NEŽINOMAS ELEMENTAS` – struktūra turi bent vieną nežinomo cheminio elemento atomą; tokias struktūras mūsų metodas praleidžia.

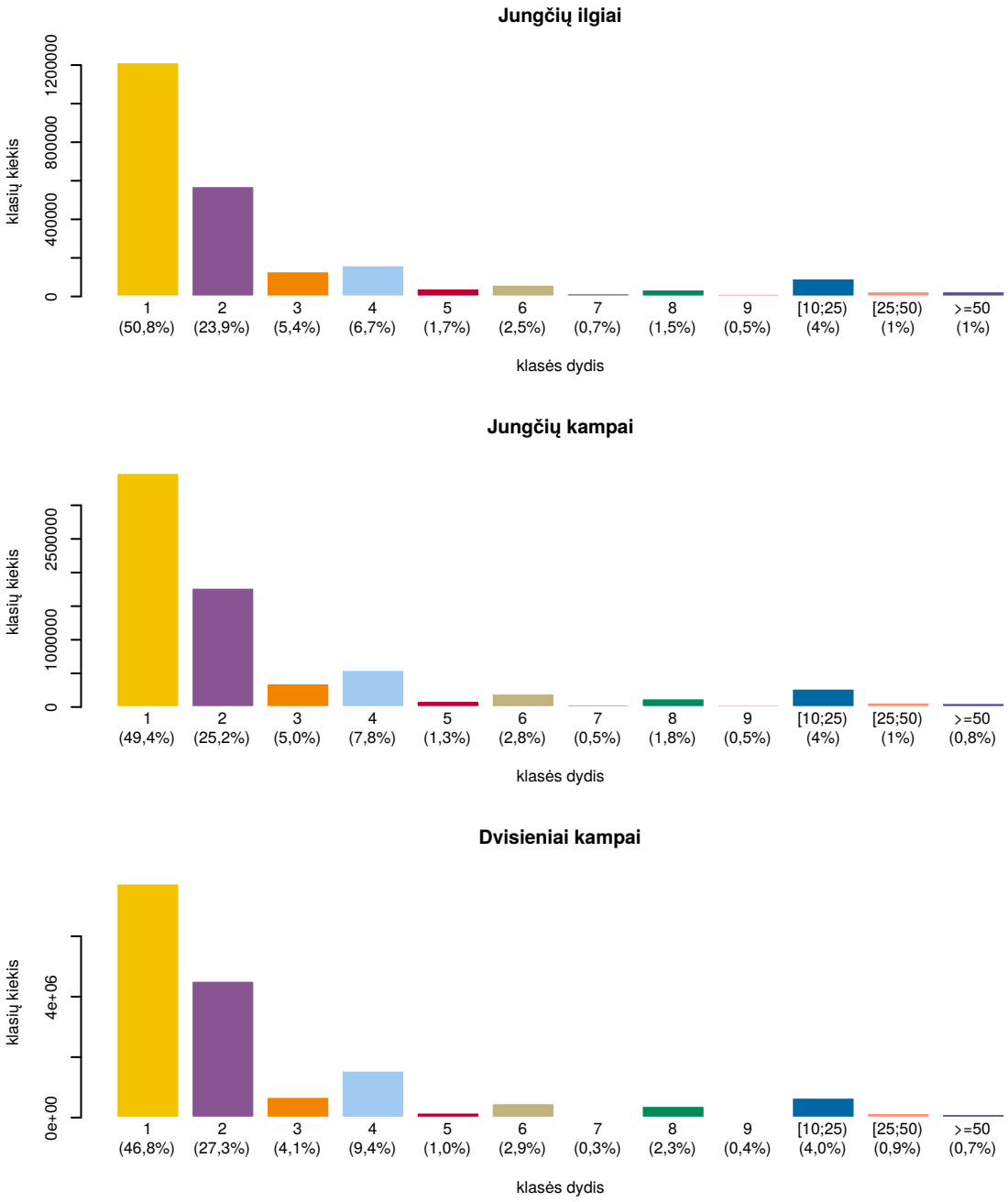
Automatiškai priskirtų žymių koreliacijos pateiktos 3.4 lentelėje. Joje galima pastebėti, jog nė vieno geometrinio parametro nebuvo išgauta iš pusės kristalų struktūrų su alternatyviomis atomų pozicijomis, dažniausiai dėl atomų tipų priklausomybės nuo pasirenkamos alternatyvos. Geometriniai parametrai taip pat nebuvo išgauti iš trečdalio polimerinių struktūrų bei struktūrų su pernelyg trumpais tarpatominiais atstumais.

3.2.2 Atomų tipai

Iš viso apdorotuose duomenyse identifikuota 1 073 426 atomų tipų, tai yra keturis kartus daugiau nei stebėta anksčiau [27]. Dauguma tipų, 822 860 priklauso „organiniam poaibiui“ (pagal SMILES specifikaciją, B, C, N, O, P, S, F, Cl, Br, I cheminiai elementai). Iš jų vien 374 303 yra anglies atomų tipai, o tai sudaro 35 % visų identifikuotų tipų, kas yra ženkliai daugiau už anksčiau PURY kūrėjų stebėto 14 % [7].

Vienoje struktūroje vidutiniškai stebėta apie 20 skirtingų atomų tipų. Struktūrose su pernelyg trumpais atstumais šis skaičius yra didesnis – 26 tipai. Iš to galima daryti išvadą, jog pernelyg trumpi atstumai iškreipia „tvarkingą“ struktūros jungumą, įvesdami atomų tipus, turinčius vos po vieną stebinį.

Automatiškai nustatyti atomų tipai galėtų būti validuojami pagal jų atitikimą chemijos dėsniams. Validacijos kriterijai, išreikšti kaip taisyklių sistema, padėtų aptikti struktūras su neįmanomais atomų tipais. Pavyzdžiui, apie 20 000 stebėtų vandenilio atomų tipų aprašo chemines aplinkas, kuriose vandenilio atomai yra susijungę su dviem kitais atomais. Dauguma šių tipų yra aiškūs artefaktai. 1500 atomų tipų aprašo atomus, turinčius penkis ar daugiau



Pav. 3.1: Klasių pasiskirstymas pagal stebinių kiekį.

	APDOROTA	TRUMPI ATSTUMAI	CIF TUŠČIAS	CIF NUTRAUKTAS	ALTERNATYVIOS KLASĖS	NETVARKA	NETVARKA SPEC. POZIC.	DUBLIKATAS	NETVARKA PAŽYMĖTA	NĖRA ATOMŲ	POLIMERAS	TAB TUŠČIAS	TAB NUTRAUKTAS	NEŽINOMAS ELEMENTAS
APDOROTA		11	2	1	9	21	2	1	6	0	23	16	4	1
TRUMPI ATSTUMAI	*		0	0	26	44	5	1	10	0	34	34	7	0
CIF TUŠČIAS	*	0		44	0	0	0	0	26	15	0	*	0	37
CIF NUTRAUKTAS	*	0	*		0	0	0	0	13	0	0	*	0	0
ALTERNATYVIOS KLASĖS	*	31	0	0		*	5	0	37	0	41	*	0	0
NETVARKA	*	24	0	0	45		9	1	24	0	30	51	5	0
NETVARKA SPEC. POZIC.	*	29	0	0	28	*		1	1	0	12	30	1	0
DUBLIKATAS	*	16	0	0	0	23	2		10	0	40	*	0	0
NETVARKA PAŽYMĖTA	*	21	9	2	62	91	0	1		0	83	91	19	7
NĖRA ATOMŲ	*	0	*	0	0	0	0	0	0		0	*	0	0
POLIMERAS	*	16	0	0	17	27	1	1	20	0		35	17	0
TAB TUŠČIAS	*	24	12	5	58	66	3	4	31	2	50		25	5
TAB NUTRAUKTAS	*	19	0	0	0	26	0	0	25	0	97	*		0
NEŽINOMAS ELEMENTAS	*	0	*	0	0	0	0	0	55	0	0	*	0	

Lentelė 3.4: Žymių koreliacija. Eilučių ir stulpelių susikirtimuose nurodyta, kiek procentų iš visų struktūrų, turinčių eilutės žymę, turi ir stulpelio žymę. Simbolis * atitinka 100 %.

kaimynų ir esančius vienoje plokštumoje su jais, kas taip pat yra gan neįprasta. Atomo dalyvavimas daugiau nei šimte žiedų taip pat gali liudyti apie klaidingą struktūrą. Tokių atomų tipų COD struktūrose pastebėta apie 350. Daugiausiai šie ir kiti neįprasti atomų tipai visoje duomenų imtyje yra sutinkami po vieną kartą, dažniausiai struktūrose, pasižyminčiose pernelyg trumpais tarpatominiais atstumais.

Dauguma jungčių, kampų bei dvisienių kampų klasių turi tik po kelis stebinius (3.1 pav. pateiktas klasių pasiskirstymas pagal stebinių kiekį). Vos 1 % klasių turi 50 ar daugiau stebinių. Tai galima paaiškinti cheminės apsupties įvairove mažų molekulių kristaluose.

Žinomi mūsų metodo ribotumai yra cheminės apsupties klasifikavimo gylis bei didžiausias žiedo dydis. Šiedu parametrai gali būti keičiami (didinami), tačiau šio tyrimo metu tai nebuvo daroma. Be abejo gylio bei didžiausio žiedo dydžio didinimas tik dar labiau susmulkintų atomų tipus, dėl ko prastai reprezentuojamų klasių tik padaugėtų. Dabartiniai mūsų pasirinkimai yra kompromisas tarp tikslumo ir stebinių klasėse skaitlingumo.

Į atomų tipus neįtraukdami aromatiškumo bei jungčių eilės (angl. *bond order*) informacijos stebime jungčių ilgių pasiskirstymus su keliomis smailėmis, tačiau automatinis jungčių eilės nustatymas nėra vienareikšmis. Tiesa, dalį minimos informacijos galima išgauti iš autorių pateikiamų cheminių vardų, tačiau juos turi tik kas trečias COD įrašas. Be to, apie 14 % atvejų ši informacija gali būti netiksli [60].

Taip pat problematiškas gali būti mūsų jungumo nustatymo taikymas už organinės chemijos ribų. Metalų koordinacija mūsų metodais atvaizduojama teisingai jeigu visos metalo ir jo koordinuojamų atomų poros yra atpažįstamos kaip jungtys. Tačiau taikomas metodas nesugeba atskirti koordinuojamų atomų nuo jų kaimynų, jei šie yra pakankamai arti prie koordinuojančio metalo atomo, kad patys būtų palaikomi jo kaimynais. Taip pat atomų tipizavimo algoritmas

Šaltinis	glicinas	be glicino ir prolino
EH 1991 [64]	112,5 ± 2,9	111,2 ± 2,8
LMT 1993 [65]	112,19 ± 3,64	110,77 ± 3,29
EH 2001 [66]	113,1 ± 2,5	111 ± 2,7
TV 2010 [62]	113,1 ± 3,4	111 ± 3
šis tyrimas	113,5 ± 1,9	111,3 ± 2,7

Lentelė 3.5: Baltymo karkaso dvisienio kampo τ palyginimas.

neatskiria skirtingų koordinavimo geometrijų, kas turi didelę įtaką geometrijos parametrų reikšmėms [61].

3.2.3 Dvisienis kampas τ

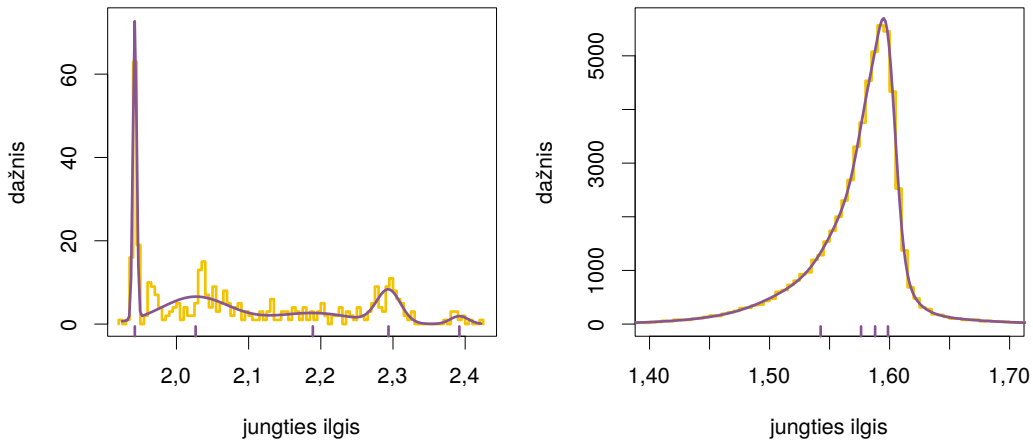
Savo tyrimo rezultatų palyginimui su anksčiau atliktais tyrimais pasirinkome dvisienį baltymo karkaso kampą τ [62]. Panaudodami atomų tipus, nusakančius į baltymo karkasą panašias chemines apsuptis, išrinkome τ kampo stebinius COD duomenų bazėje. Šių stebinių vidurkis pateiktas palyginimui su kitų tyrimų rezultatais 3.5 lentelėje. Matyti, jog mūsų tyrimo rezultatai yra panašūs į anksčiau gautus, nors stebimi τ kampai bendrai yra didesni. Palyginus su Balasco ir k.t. (2017) rezultatais (nagrinėtas τ kampas baltymuose ties aminorūgštimis išskyrus gliciną ir proliną), COD τ vertė yra labai panaši į vidutinę autorių stebėtą vertę ($\sim 111{,}3°$) [63]. Viena iš didesnio neatitikimo priežasčių gali būti mažas stebinių skaičius COD (11 glicino, 36 ne glicino ir prolino aminorūgščių aplinkas primenantys fragmentai). Tačiau mažesnis nei anksčiau stebėtas τ kampų standartinis nuokrypis ties glicino liekanomis liudija išskirčių nebuvimą. Mūsų metodika nagrinėtuose duomenyse identifikavo aštuonias netuščias τ kampo klases pagal C_β: glicino (`NCH`, 11 stebinių), alanino (`CH3`, 15 stebinių), linijinės alkano grandinės (`CCHH`, 10 stebinių), treonino (`CCHO`, 3 stebiniai), β-šakotų grandinių liekanų (valino ir izoleucino, `CCCH`, 4 stebiniai), serino (`CHHO`, du stebiniai), cisteino (`CHHS`) ir *tert*-leucino (`CC3`), turinčių po vieną stebinį.

3.2.4 Jahn–Teller efektas

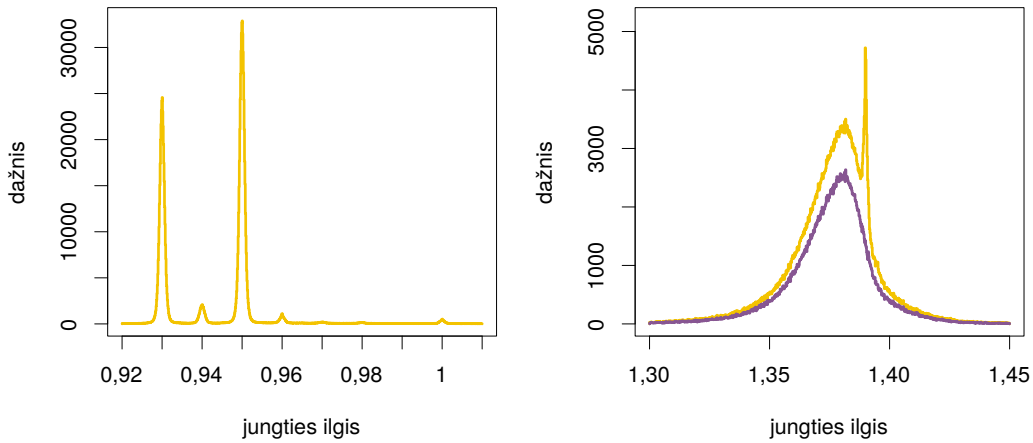
COD duomenyse stebimas Jahn–Teller efektas, pailginantis ašines jungtis šešianariuose vario–vandens kompleksuose. Efektas pasireiškia panašiai kaip ir Harding (1999) aprašytame CSD geometrijos tyrime [67] (3.2 pav.). Tačiau trumpų jungčių ilgiai COD yra $\sim 0{,}2$ Å trumpesni, o nemaža ilgesnių jungčių stebinių trūksta dėl naudojamo vario–deguonies jungties kovalentinio atstumo, lygaus 2,5 Å. Vario–deguonies jungčių ilgiams šiuose koordinaciniuose kompleksuose aprašyti mūsų algoritmas parinko penkių komponentų mišinį (mišinio tankis 3.2 pav. parodytas violetine spalva).

3.2.5 Patikslinimo priemonių poveikis

Laskowski ir k.t. (1993) pastebėjo, kad geometrijos bibliotekos bei patikslinimo programinė įranga dažnai neigiamai paveikia modelių geometriją. Kai kada įvedamas poveikis toks stiprus, jog vadovaujantis nesudėtingomis taisyklėmis tampa įmanoma 95 % atvejų iš geometrijos nustatyti naudotus metodus bei jų parametrus [65]. Galimi patikslinimo priemonių pėdsakai



Pav. 3.2: **kairėje)** vario–deguonies jungčių ilgiai šešianariuose vario–vandens kompleksuose. **dešinėje)** fosforo–fluoro jungčių ilgiai heksafluorofosfato molekulėse.



Pav. 3.3: Patikslinimo priemonių poveikis benzeno jungčių ilgiams: **kairėje)** C–H ir **dešinėje)** C–C.

taip pat sutinkami ir COD duomenyse. Pavyzdžiui, beveik diskretus benzeno C–H jungčių ilgių pasiskirstymas (3.3 pav. kairėje) turi penkias aiškias smailes ties 0,93, 0,94, 0,95, 0,96 ir 1 Å, iš ko galima spręsti apie struktūrų patikslinimo metu galimai naudotą papildomą geometrinę informaciją. Įtakos pasiskirstymams galima būtų išvengti iš imties pašalinus stebinius iš struktūrų, išreikštiniu būdu pažymėtų kaip patikslintų naudojant papildomą geometrinę informaciją (tokiam pažymėjimui dažniausiai naudojama CIF duomenų žymė `_atom_site_refinement_flags` su reikšme kita nei „.“). Tokio pasiskirstymo histogramoje (pateikiama 3.3 pav. dešinėje, violetinė spalva) anomalių smailių nebematoma. Šio tyrimo metu ieškota koreliacijų tarp C–H smailių bei patikslinimui naudotos programinės įrangos, tačiau sąsajų nerasta greičiausiai dėl to, kad su ta pačia programine įranga galėjo būti naudotos skirtingos geometrijos bibliotekos.

COD ID	d_C	α_C	ϕ_C	$\sum_C$	$\sum_P$	trumpi atstumai
1546823	11	17	0	28	1	
2020874	0	18	0	18	0	6
7229026	0	7	0	7	0	
4002839	0	6	0	6	0	
1546859	2	3	0	5	2	
4126340	5	0	0	5	0	
7228987	0	0	5	5	0	
1546887	0	4	0	4	4	
7056555	2	1	1	4	0	10
1546858	1	2	0	3	0	
4126332	1	2	0	3	0	
7229031	0	0	3	3	0	
1546862	0	2	0	2	6	
7044067	0	2	0	2	1	2
1546918	0	2	0	2	0	
7155853	0	2	0	2	0	
7228989	0	0	2	2	0	
7229032	0	2	0	2	0	
7120611	1	0	0	1	1	
1546820	0	1	0	1	0	
7044036	0	1	0	1	0	
7044045	0	1	0	1	0	
7228985	0	1	0	1	0	
1546884	0	0	0	0	4	
1546885	0	0	0	0	4	
1546889	0	0	0	0	4	
7229014	0	0	0	0	4	
7044095	0	0	0	0	3	3
7120612	0	0	0	0	3	
7056573	0	0	0	0	2	
1546848	0	0	0	0	1	
1546913	0	0	0	0	1	
7044089	0	0	0	0	1	
7229036	0	0	0	0	1	
7229042	0	0	0	0	1	

Lentelė 3.6: 100 naujų COD struktūrų validavimo rezultatai. Įrašai be neįprastos geometrijos pranešimų praleisti. d_C, α_C, ϕ_C yra mūsų metodo aptiktų neįprastų jungčių, kampų ir dvisienių kampų (atitinkamai) skaičiai; $\sum_C$ ir $\sum_P$ yra atitinkamai mūsų ir PLATON metodo pranešimų skaičiai.

COD ID	d_C	α_C	ϕ_C	$\sum_C$	$\sum_P$	trumpi atstumai
2214032	4	5	0	9	1	
2214266	7	0	0	7	1	
2214740	7	0	0	7	1	
2214731	6	0	0	6	1	
2215545	3	3	0	6	0	
2214494	4	1	0	5	0	
2215330	1	3	0	4	1	
2211060	1	3	0	4	0	
2215738	4	0	0	4	0	
2216003	2	2	0	4	0	
2216548	3	1	0	4	0	
2216555	3	1	0	4	0	
2216233	1	2	0	3	1	
2214035	3	0	0	3	0	
2214948	0	2	1	3	0	
2215165	1	1	1	3	0	
2215546	3	0	0	3	0	
2215724	3	0	0	3	0	
2215994	0	3	0	3	0	
2216004	3	0	0	3	0	
2215725	0	2	0	2	1	
2215733	1	1	0	2	1	
2214020	0	2	0	2	0	
2215167	2	0	0	2	0	
2215544	0	0	2	2	0	
2216567	1	1	0	2	0	
2214265	1	0	0	1	1	
2211463	0	1	0	1	0	
2211708	0	1	0	1	0	1
2211760	1	0	0	1	0	
2214728	0	1	0	1	0	
2214858	0	1	0	1	0	
2214947	1	0	0	1	0	
2214955	1	0	0	1	0	
2215999	0	1	0	1	0	
2216217	1	0	0	1	0	
2214262	0	0	0	0	2	
2215170	0	0	0	0	2	
2216576	0	0	0	0	2	
2214066	0	0	0	0	1	
2214261	0	0	0	0	1	
2214492	0	0	0	0	1	
2217640	0	0	0	0	1	285

Lentelė 3.7: 70 atšauktų struktūrų [68] validavimo rezultatai. Įrašai be neįprastos geometrijos pranešimų praleisti. Naudojami 3.6 lentelės žymėjimai.

3.2.6 Naujų struktūrų validavimas

Tikrindami savo metodą atlikome šimto atsitiktinių naujai į COD įkeltų struktūrų validavimą mūsų bei PLATON programine įranga. Pasirinktos naujos struktūros nebuvo panaudotos geometrijos bibliotekos konstravimui. Validavimo rezultatų suvestinė pateikiama 3.6 lentelėje. Iš tikrintų struktūrų devyniose aptikti pernelyg trumpi tarpatominiai atstumai, dvylika struktūrų negalėjo būti apdorotos dėl skaičiavimo resursų perviršijimo. Bent po vieną neįprastą geometrinį parametrą mūsų metodas aptiko 23 struktūrose, PLATON metodas – 18, šešios iš šių struktūrų neįprastomis pažymėtos abiejų metodų.

Dauguma nagrinėtų struktūrų turi neįprastų parametrų, kurių priežastis greičiausiai yra prastas vandenilio atomų modeliavimas. 11 jungčių ir 17 kampų, kuriuose dalyvauja vandenilio atomai, pažymėti kaip neįprasti COD 1546823 struktūroje tiek mūsų, tiek PLATON metodu. 1546859 struktūra neįprasta vėlgi laikoma abiejų metodų. Mūsų metodas pažymėjo neįprastus vandenilio atomų parametrus, PLATON pranešė apie netaisyklingus kampus ties sidabro

COD ID	d_C	α_C	ϕ_C	$\sum_C$	$\sum_P$	trumpi atstumai
1519776	5	27	0	32	0	66
4112502	22	9	0	31	0	
2005559	9	19	0	28	0	
4316914	7	9	8	24	0	
7152986	7	15	0	22	2	
1517225	9	10	0	19	41	4
7103775	9	8	0	17	1	
4083625	0	17	0	17	0	
2000089	11	3	0	14	0	2
4326131	3	9	2	14	0	
4028178	4	7	2	13	0	
4313003	6	4	3	13	0	4
2006345	6	6	0	12	0	
4308222	0	12	0	12	0	
7226719	0	3	9	12	0	
2101240	3	7	0	10	0	
4508401	0	8	2	10	0	
2100783	3	6	0	9	0	
4316029	0	9	0	9	0	879
4323192	4	5	0	9	0	1
4122124	6	2	0	8	1	6
1543698	2	6	0	8	0	
2001438	4	4	0	8	0	8
4061664	6	2	0	8	0	
4110701	4	3	0	7	4	
8102293	5	2	0	7	1	
1515403	7	0	0	7	0	
2008570	5	2	0	7	0	
4022241	0	7	0	7	0	
4068650	3	4	0	7	0	

Lentelė 3.8: 1000 atsitiktinių COD struktūrų validavimo rezultatai. Pateikti 30 įrašų, turinčių daugiausiai neįprastos geometrijos pagal mūsų metodą pranešimų. Naudojami 3.6 lentelės žymėjimai.

COD ID	d_C	α_C	ϕ_C	$\sum_C$	$\sum_P$	trumpi atstumai
4111438	0	0	0	0	15	1
4061731	0	0	0	0	12	
2019970	0	0	0	0	9	639
4077645	0	0	0	0	9	
4101695	0	0	0	0	6	
4309879	0	0	0	0	6	
7009707	0	0	0	0	6	
2203936	0	0	0	0	5	
4070531	0	0	0	0	5	
4074116	0	0	0	0	5	
4076457	0	0	0	0	4	
4104545	0	0	0	0	4	
4113595	0	0	0	0	4	
4317601	0	0	0	0	4	
4333210	0	0	0	0	4	
4502024	0	0	0	0	4	
7002930	0	0	0	0	4	
7041367	0	0	0	0	4	
7102241	0	0	0	0	4	1
1515649	0	0	0	0	3	4
2204549	0	0	0	0	3	
7003771	0	0	0	0	3	
7219216	0	0	0	0	3	
2201582	0	0	0	0	2	
2207064	0	0	0	0	2	
4001338	0	0	0	0	2	
4027467	0	0	0	0	2	
4077122	0	0	0	0	2	2
4077989	0	0	0	0	2	
4079811	0	0	0	0	2	

Lentelė 3.9: 1000 atsitiktinių COD struktūrų validavimo rezultatai. Pateikti 30 įrašų, turinčių daugiausiai neįprastos geometrijos pagal PLATON pranešimų. Naudojami 3.6 lentelės žymėjimai.

atomais. Neįprastos fragmentų su vandenilio atomais konformacijos pažymėtos 7228987 (5 dvisieniai kampai) ir 7229031 (3 dvisieniai kampai) struktūrose. Abiejose vandeniliai patikslinti naudojant geometrinius sąryšius. Veikiausiai nepakankamas 7229026 struktūros patikslinimas (kristalografinis struktūros kokybės faktorius visiems atspindžiams $R_w = 0{,}2966$ yra ganėtinai aukštas, kas ženklina prastesnį nei įprasta patikslinimą) pasireiškia 7 geometrinių parametrų pripažinimu įtartinais. Šie parametrai daugiausiai išsidėstę netvarkingose struktūros dalyse. Nepažymėti alternatyvūs atomų išsidėstymai aplink specialiąsias pozicijas pasireiškė pernelyg trumpais tarpatominiais atstumais dviejose struktūrose. Šešios poros pernelyg arti esančių atomų aptiktos 2020874 struktūros metilo grupėje. Šioje struktūroje mūsų programinė įranga aptiko 18 neįprastų parametrų vandenilio atomus turinčiuose fragmentuose. Dvi neįprastos jungtys 7056555 struktūroje stebimos dėl vandens molekulės, esančios ant simetrijos ašies, tačiau nepažymėtos kaip nepriklausančios nuo šio simetrijos operatoriaus. Be to, neįprastais pažymėti vienas kampas vario koordinaciniame komplekse bei dvisienis kampas, kuriame dalyvauja du vandenilio atomai, patikslinti naudojant sąryšius. Mūsų metodas neįprastais laiko 1546887 struktūros feroceno grupės kampus. Šią grupę sudarantys penkianariai žiedai struktūroje vienas kito atžvilgiu yra pasisukę labai artimu 36° kampu, kas COD yra retai stebima. PLATON metodas įtartinais nurodo šios struktūros boro ir germanio koordinacinių kompleksų kampus. Neįprastomis mūsų algoritmas laiko penkias anglies–deguonies jungtis 4126340 struktūros ciklodekstrino molekulėje. Ši molekulė yra sudėtinė didelio elementaraus narvelio dalis, o struktūrą aprašančiame straipsnyje ciklodekstrino jungčių ilgių ypatumai neminimi. Patikrinę

molekulę `olex2` [69] vaizdavimo programine įranga nustatėme, jog bent vienas kiekvienos jungties atomas turi didesnes nei įprasta šiluminio judėjimo parametrų vertes. 4002839 struktūra demonstruoja mūsų pasiskirstymų aprašymo metodo negebėjimą pakankamai gerai parinkti modelius pasiskirstymams su besikartojančiais vienodais stebiniais. Germanio–fosforo–sidabro kampų klasės, turinčios vienuolika stebinių iš kurių tik du yra unikalūs, pasiskirstymą mūsų metodas aprašė vieno Koši komponento mišiniu, atitinkančiu dažniausiai pasikartojantį stebinį. Šeši stebiniai iš 4002839 struktūros patenka per vidurį tarp minėtų unikalių stebinių, todėl algoritmo yra laikomi mažai tikėtinais.

Įdomu pastebėti, jog mūsų metodo bei `PLATON` neįprastos geometrijos traktavimai mažai persidengia. Tik šešios struktūros neįprastomis palaikytos abiejų metodų. Penkiose iš daugiausiai `PLATON` pranešimų turinčių struktūrų (trys ir daugiau pranešimai kiekvienai struktūrai) `PLATON` metodas neįprasta palaikė geometriją menkai COD reprezentuotuose fragmentuose. Galima daryti išvadą, jog mūsų metodas yra jautresnis už `PLATON`, tačiau pagal apibrėžimą negali validuoti anksčiau nestebėtų cheminių apsupčių. Todėl geresni validavimo rezultatai gali būti pasiekti derinant mūsų ir `PLATON` metodo pranašumus.

3.2.7 Atšauktų struktūrų validavimas

Taip pat išbandėme savo neįprastos geometrijos aptikimo algoritmą su 70 atšauktų struktūrų iš *Acta Crystallographica* žurnalo, kurios 2010 m. buvo paskelbtos falsifikuotomis [68]. Kadangi struktūrų atšaukimas įvyko anksčiau nei šis tyrimas, jų duomenys nebuvo panaudoti geometrijos bibliotekos konstravimui. Iš 70 tikrintų struktūrų dviejose aptikti pernelyg trumpi tarpatominiai atstumai, bent po vieną neįprastą geometrinį parametrą mūsų metodas aptiko 36 struktūrose, `PLATON` metodas – 16, devynios iš šių struktūrų neįprastomis pažymėtos abiejų metodų. Žemiau pateikiame daugiausiai validacijos pranešimų turinčių struktūrų apžvalgą.

Daugiausia neįprastų parametrų mūsų algoritmas aptiko koordinaciniuose junginiuose, publikuotuose Zhong ir k.t. (2007) [70]. Įtartina geometrija nustatyta iš anglies ir azoto atomų sudarytuose fragmentuose 2214032, 2214266, 2214494, 2214731, 2214740 ir 2215545 struktūrose. Daugumoje šių struktūrų neįprastus koordinacinių sąveikų kampus nustatė ir `PLATON`. Tos pačios tyrėjų grupės paskelbtose 2216548 ir 2216555 struktūrose įtartinais mūsų algoritmas laiko anglis–anglis jungčių ilgius acto rūgšties molekulėse bei kampus metalų koordinacinėse sferose. Keturi neįprasti geometriniai parametrai mūsų metodo aptikti ir 2216003 struktūros karboksifenilo grupėje. Keturi įtartini anglies–azoto jungčių ilgiai nustatyti 2215738 struktūroje. Daugumoje struktūrų neįprasta geometrija stebima fragmentuose, turinčiuose vandenilio atomų. Galima daryti išvadą, jog tokia geometrija pasitaiko gan dažnai lyginant su sunkesnių atomų fragmentais.

Šiuo atveju vėlgi matome, jog mūsų ir `PLATON` metodai mažai persidengia. Dvi struktūros, 2214262 ir 2216576, turinčios daugiausia `PLATON` validacijos pranešimų, iš tiesų yra tvarkingos: struktūrų hidroksigrupes `PLATON` laiko esančias prijungtas prie vario atomų ir šių jungčių kampus praneša kaip neįprastus. Vienas 2215170 struktūros fragmentas turi nepažymėtų alternatyvių atomų, `PLATON` praneša apie šio fragmento kampus. Mūsų metodas fragmento jungumą laiko teisingu, pranešdamas apie šešis anksčiau nestebėtus atomų tipus.

3.2.8 Atsitiktinių COD struktūrų validavimas

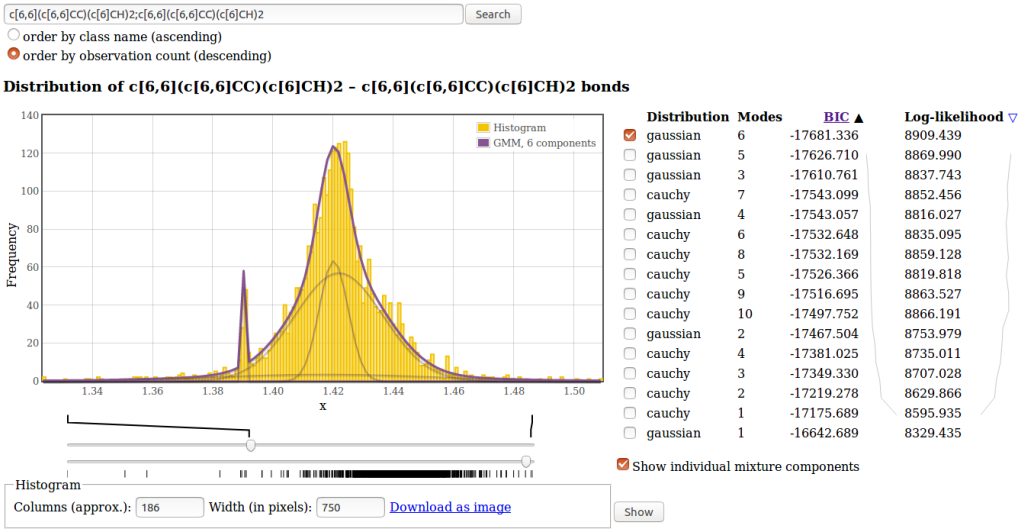
Validavimą atlikome ir tūkstančiui atsitiktinai parinktų iš geometrinei bibliotekai sukonstruoti naudotų COD struktūrų. Iš jų 125 aptikti pernelyg trumpi tarpatominiai atstumai, 159 struktūros negalėjo būti apdorotos dėl skaičiavimo resursų perviršijimo. Bent po vieną neįprastą geometrinį parametrą mūsų metodas aptiko 206 struktūrose, `PLATON` metodas – 109, 28 iš šių struktūrų neįprastomis pažymėtos abiejų metodų.

Keturių struktūrų geometrija laikoma neįprasta dėl vandenilio atomų, kurių pozicijos dažniausiai yra suskaičiuotos arba nepatikslintos. Visi atomų jungčių su vandeniliais ilgiai 4112502 struktūroje yra ilgesni nei įprasta, todėl galima daryti išvadą, jog šioms jungtims patikslinti buvo naudoti ne rentgenostruktūrinės analizės metu nustatyti matavimai. Tas pats pastebėta ir 2005559 bei 7103775 struktūrose. Vandenilio atomų pozicijos 2000089 struktūroje buvo nustatytos naudojant elektronų tankio žemėlapius, tačiau visi neįprasti struktūros geometriniai parametrai aptikti būtent fragmentuose su vandenilio atomais. Keturiose struktūrose, pasižyminčiose prastesniais nei įprasta patikslinimo kokybės įverčiais, mūsų metodas aptiko įtartinų geometrinių parametrų. 4316914 struktūroje, kurios kristalografiniai R ir R_w faktoriai dydžiu viršija įprastines vertes, keli fenilo žiedai sumodeliuoti iškraipyti. Neįprastos konformacijos benzeno žiedai bei sočiosios anglies grandinės stebimos 7152986 struktūroje, kuri taip pat pasižymi didelėmis R ir R_w vertėmis bei dideliu paskutinės patikslinimo iteracijos parametrų skirtumu, kas gali žymėti pernelyg anksti nutrauktą patikslinimo procesą. Labai didelio elementariojo narvelio 1517225 struktūroje, kurios R_w yra šiek tiek didesnis nei įprasta, mūsų metodas aptiko keliolika įtartinų jungčių ilgių, šiuo atveju neįprasta geometrija patvirtinta ir `PLATON` metodo. Keli įtartini geometrijos parametrai renio koordinacinėse sferose, imidazolo ir piridino žieduose aptikti 4326131 struktūroje. Ši struktūra taip pat atrodo ne iki galo patikslinta. Nepakankamai gerai patikslinta alternatyvias pozicijas įgyjanti 1519776 struktūros dalis mūsų metodo taip pat pažymėta kaip turinti neįprastų geometrijos parametrų. Tolueno molekulė 4083625 struktūroje yra stipriai iškreipta, nors pažymėta kaip patikslinta naudojant papildomą geometrinę informaciją. Gali būti, kad patikslinimo metu ši informacija buvo panaudota nekorektiškai.

3.2.9 Tipografinių klaidų aptikimas

Norėdami patikrinti kaip mūsų bei `PLATON` metodai aptinka tipografines klaidas struktūrų koordinatėse, atlikome testus su struktūromis, kurių koordinatėse atsitiktiniu būdu pakeitėme po vieną skaitmenį. Bandymą atlikome su 22 COD struktūromis, kurios mūsų metodo nebuvo pripažintos neįprastomis atsitiktinių struktūrų validavimo metu. Taip pat reikalavome, kad visi šių struktūrų parametrai turėtų parinktus aprašančius modelius. Kiekvienoje struktūroje atsitiktiniu būdu pakeitėme po vieną skaitmenį ketvirtoje ir trečioje mantisės pozicijoje.

Ketvirtoje pozicijoje įvesti klaidingi skaitmenys nesukėlė tokių pasikeitimų struktūrų geometrijoje, kad šie būtų aptikti bent vienu metodu. 9003119 struktūroje pakeista koordinatė išvedė chloro atomą iš specialiosios pozicijos, ką mūsų metodas aptiko, tačiau tikrindamas kristalo specialiųjų pozicijų simetrijos operatorių skaičius, o ne geometriją. Trečioje mantisės pozicijoje įvestos tipografinės klaidos mūsų metodo buvo aptiktos šešiose struktūrose. Penkiose iš jų pakeitimai sukėlė pernelyg trumpų tarpatominių atstumų atsiradimą, iš kurių trys pakeitė



Pav. 3.4: COD geometrijos naršyklė, vaizduojamas naftalino C_{4a}–C_{8a} jungčių ilgių pasiskirstymas. Histogramoje pateikiamas stebinių dažnis ruože 1,153–1,513 Å. Ant histogramos viršaus pavaizduotas pasiskirstymą aprašantis mišinio modelis, turintis mažiausią BIC įvertį (stora violetinė linija). Aiškiai išsiskiria keturi iš šešių normaliųjų skirstinių mišinio komponentų (plona violetinė linija). Kiti modeliai bei jų parametrai pateikiami histogramos dešinėje pusėje.

struktūrų jungumą – buvo stebimi anksčiau nematyti atomų tipai. Vienoje struktūroje mūsų metodas aptiko neįprastai trumpo ilgio jungtį, kitoje – neįprasto dydžio kampą. `PLATON` programa neįprastos geometrijos neaptiko.

3.3 COD ir TCOD kuravimas

Šio tyrimo metu COD įrašų skaičius išaugo nuo ~ 217 000 iki 390 000. Kadangi tyrimas priklausė nuo COD duomenų kokybės, daug pastangų įdėta duomenų bazės kuravimui. Daugiau nei 100 struktūrų su pasikartojančiais atomais buvo nustatyta ir pataisyta tikrinant COD įrašus, turinčius gausiausiai pernelyg trumpų atstumų. Dauguma šių struktūrų publikuota kartu su tyrimais, siekiančiais nustatyti trūkstamus simetrijos operatorius kristalų struktūrų modeliuose. Tuštumoms kristaluose aptikti sukūrėme programą `cif_voids`[2], paremtą `voronota` [71]. Sukurta programa panaudota 30 struktūrų su nurodyta pernelyg žema simetrija nustatymui. Dėl trūkstamų simetrijos operatorių šių struktūrų modeliams trūko atomų, vietoje jų buvo stebimos tuštumos. COD geometrijos bibliotekos naršyklė (ekranvaizdis pateikiamas 3.4 pav.) bei ja grįstas struktūrų validatorius buvo panaudotas aptikimui bei pataisymui 25 struktūrų, kurioms trūko pažymėtų nesumodeliuotų vandenilio atomų. Tyrimo metu rastos ir pataisytos klaidos dar 200 kitų COD įrašų. COD buvo aptikta ir pažymėta apie 450 teoretinių struktūrų, dauguma jų buvo įkeltos į Atvirą mažų molekulių teoretinę kristalografinę duomenų bazę TCOD[3]. Šioje duomenų bazėje įrašų skaičius išaugo iki 2600.

[2] Platinama su GNU GPL2 atviro kodo licenzija `svn://saulius-grazulis.lt/crystalvoids/trunk`, šiame tyrime naudota 64 revizija.

[3] `http://www.crystallography.net/tcod`

Skyrius 4

Išvados

- Atvira mažų molekulių kristalografinė duomenų bazė yra nuolat augantis ir besivystantis struktūrinės mažų molekulių informacijos šaltinis. Sukurta programinė įranga yra naudinga geometrinės informacijos išgavimui bei žinių bibliotekos kūrimui.
- Sukurta metodika yra pakankama pilnai automatizuotam ir neprižiūrimam šaltinyje stebimos geometrijos organizavimui į geometrijos žinių biblioteką.
- Sukurta biblioteka yra tinkama Bajesiniu metodu pagrįstam molekulių modelių validavimui.

Mokslinių darbų sąrašas

Disertacijoje pateikta medžiaga paskelbta šiuose moksliniuose straipsniuose

1. Saulius Gražulis, Andrius Merkys, Antanas Vaitkus, and Mykolas Okulič-Kazarinas. Computing stoichiometric molecular composition from crystal structures. *Journal of Applied Crystallography*, 48:85–91, 2015. URL: http://scripts.iucr.org/cgi-bin/paper?S1600576714025904

2. Andrius Merkys, Antanas Vaitkus, Justas Butkus, Mykolas Okulič-Kazarinas, Visvaldas Kairys, and Saulius Gražulis. COD::CIF::Parser: an error-correcting CIF parser for the Perl language. *Journal of Applied Crystallography*, 49(1):292–301, 2016. doi:10.1107/S1600576715022396

3. Fei Long, Robert A. Nicholls, Paul Emsley, Saulius Gražulis, Andrius Merkys, Antanas Vaitkus, and Garib N. Murshudov. Validation and extraction of stereochemical information from small molecular databases. *Acta Crystallographica Section D*, 73(2):103–111, 2017. doi:10.1107/S2059798317000079

4. Fei Long, Robert A. Nicholls, Paul Emsley, Saulius Gražulis, Andrius Merkys, Antanas Vaitkus, and Garib N. Murshudov. ACEDRG: A stereo-chemical description generator for ligands. *Acta Crystallographica Section D*, 73(2):112–122, 2017. doi:10.1107/S2059798317000067

5. Andrius Merkys, Nicolas Mounet, Andrea Cepellotti, Nicola Marzari, Saulius Gražulis, and Giovanni Pizzi. A posteriori metadata from automated provenance tracking: Integration of AiiDA and TCOD. *Journal of Cheminformatics*, 9(1), 2017. URL: https://jcheminf.springeropen.com/articles/10.1186/s13321-017-0242-y, arXiv:1706.08704v3, doi:10.1186/s13321-017-0242-y.

6. Nicolas Mounet, Marco Gibertini, Philippe Schwaller, Davide Campi, Andrius Merkys, Antimo Marrazzo, Thibault Sohier, Ivano Eligio Castelli, Andrea Cepellotti, Giovanni Pizzi, and Nicola Marzari. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nature Nanotechnology*, Feb 2018. doi:10.1038/s41565-017-0035-5

Kiti straipsniai

1. Saulius Gražulis, Adriana Daškevič, Andrius Merkys, Daniel Chateigner, Luca Lutterotti, Miguel Quirós, Nadezhda R. Serebryanaya, Peter Moeck, Robert T. Downs, and Armel Le Bail. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration *Nucleic Acids Research*, 40:D420–D427, 2012.

Disertacijoje pateikta medžiaga buvo pristatyta šiose konferencijose

1. CECAM/Psi-k Research Conference: Frontiers of first-principles simulations: materials design and discovery (2015 m. vasario 1–5 d., Berlynas, Vokietija). Stendinis pranešimas: *Theoretical Crystallography Open Database – open-access repository of theoretically computed crystal structures*

2. Platform for Advanced Scientific Computing Conference (2015 m. birželio 1–3 d., Ciurichas, Šveicarija). Stendinis pranešimas: *Developing Experimental & Theoretical Crystallography Open Databases*

3. 29th European Crystallographic Meeting (2015 m. rugpjūčio 23–28 d., Rovinis, Kroatija). Stendinis pranešimas: *Integration of TCOD (Theoretical Crystallography Open Database) and AiiDA (Automated Interactive Infrastructure and Database for Atomistic simulations)*

4. OpenReadings2016 (2016 m. kovo 16 d., Vilnius, Lietuva). Žodinis pranešimas: *Spotting the geometric properties in the Crystallography Open Database*

5. OpenReadings2017 (2017 m. kovo 14 d., Vilnius, Lietuva). Žodinis pranešimas: *Spotting the Unusual Geometry in Crystal Structures*

6. OpenReadings2018 (2018 m. kovo 20–23 d., Vilnius, Lietuva). Stendinis pranešimas: *Statistical Insights into the Chemical Bonding in Crystal Structures*

Curriculum Vitae

Asmeniniai duomenys

Vardas, pavardė:	Andrius Merkys
Gimimo data ir vieta:	1988-01-11, Vilnius, Lietuva
Telefonas:	+370 613 12191
El. paštas:	andrius.merkys@gmail.com

Išsilavinimas

2013 – 2017	Chemijos inžinerija, doktorantūros studijos, Vilniaus universitetas
2011 – 2013	Informatikos magistras (Magna Cum Laude), Vilniaus universitetas
2007 – 2011	Bioinformatikos bakalauras, Vilniaus universitetas
2000 – 2007	Vilniaus Karoliniškių gimnazija
1995 – 2000	Vilniaus Tuskulėnų vidurinė mokykla

Darbo patirtis

2010 – 2013	Asistentas Baltymų–nukleorūgščių sąveikos tyrimo skyriuje, Vilniaus universitetas
2013 – dabar	Jaunesnysis mokslo darbuotojas Baltymų–nukleorūgščių sąveikos tyrimo skyriuje, Vilniaus universitetas
2017 – dabar	Lektorius Matematikos ir informatikos fakultete, Vilniaus universitetas

Stažuotės

2014 – 2015	École Polytechnique Fédérale de Lausanne, Lozana, Šveicarija (13 mėnesių)
2012	Medical Research Council Laboratory of Molecular Biology, Kembridžas, Jungtinė Karalystė (2 mėnesiai)

Santrumpų sąrašas

BIC Bayesian Information Criterion (liet. *Bajesinis informacijos kriterijus*).
BOM byte order mark (liet. *baitų tvarkos žymė*).

CCDC Cambridge Crystallographic Data Centre (liet. *Kembridžo kristalografinių duomenų centras*).
CIF Crystallographic Information Framework/Format (liet. *Kristalografinės informacijos formatas*).
COD Crystallography Open Database (liet. *Atvira mažų molekulių kristalografinė duomenų bazė*).
CSD Cambridge Structural Database (liet. *Kembridžo struktūrinė duomenų bazė*).

SMILES Simplified Molecular-Input Line-Entry System (liet. *Supaprastina tekstinė molekulinės įvesties sistema*).
STAR Self-defining Text Archive and Retrieval (liet. *Save aprašantis teksto archyvavimo ir išgavimo metodas*).

TCOD Theoretical Crystallography Open Database (liet. *Atvira mažų molekulių teoretinė kristalografinė duomenų bazė*).

Santrauka anglų kalba (Abstract)

This dissertation describes fully automated means to extract geometric information – interatomic bond lengths, bond and dihedral angles – from small-molecule crystal structures, and to use this information for the validation of novel crystal structures. Crystallography Open Database (COD), regularly updated open-access resource of small-molecule crystal structures, has been chosen as the source of input data. Software has been developed to prefilter the records from the COD, transform them to a form appropriate for geometric analysis, extract and organise the geometric parameters. Statistical models chosen to describe the groups of chemically similar observations can be used for Bayesian method-based outlier detection: previously unseen, or seen relatively rarely, geometric observations in molecules in consideration are spotted and marked for further analysis. Software implementing this principle has been developed and a Web based user interface has been presented. The method for structure validation has been tested with novel, retracted and deliberately deformed small-molecule crystal structures. The main conclusions of this dissertation are that the COD is a proper resource for small-molecule geometric information, developed methods and software tools are sufficient to organise the data from the source database into a library of molecular geometry, which is in turn capable to spot unusual geometric features in small-molecule crystal structures.

Literatūros sąrašas

[1] Wayne A. Hendrickson. Stereochemically restrained refinement of macromolecular structures. *Methods in enzymology*, 115:252–270, 1985. doi:10.1016/0076-6879(85)15021-4.

[2] John Liebeschuetz, Jana Hennemann, Tjelvar Olsson, and Colin R. Groom. The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. *Journal of Computer-Aided Molecular Design*, 26(2):169–183, Jan 2012. URL: http://dx.doi.org/10.1007/s10822-011-9538-6, doi:10.1007/s10822-011-9538-6.

[3] Marc C. Deller and Bernhard Rupp. Models of protein–ligand crystal structures: trust, but verify. *Journal of Computer-Aided Molecular Design*, 29(9):817–836, Feb 2015. URL: http://dx.doi.org/10.1007/s10822-015-9833-8, doi:10.1007/s10822-015-9833-8.

[4] Roberto A. Steiner and Julie A. Tucker. Keep it together: restraints in crystallographic refinement of macromolecule–ligand complexes. *Acta Crystallographica Section D Structural Biology*, 73(2):93–102, Feb 2017. URL: http://dx.doi.org/10.1107/S2059798316017964, doi:10.1107/s2059798316017964.

[5] Gerard J. Kleywegt. Crystallographic refinement of ligand complexes. *Acta crystallographica. Section D, Biological crystallography*, 63:94–100, 2007. doi:10.1107/S0907444906022657.

[6] Gerard J. Kleywegt and Mark R. Harris. ValLigURL: a server for ligand-structure comparison and validation. *Acta Crystallographica Section D*, 63:935–938, 2007. URL: http://scripts.iucr.org/cgi-bin/paper?S090744490703315X, doi:10.1107/S090744490703315X.

[7] Miha Andrejašič, Jure Pražnikar, and Dušan Turk. PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures. *Acta Crystallographica Section D, Biological Crystallography*, 64:1093–109, 2008. doi:10.1107/S0907444908027388.

[8] Alexander J. Blake, William Clegg, Jacqueline M. Cole, John S. O. Evans, Peter Main, Simon Parsons, and David J. Watkin. *Crystal Structure Analysis: Principles and Practice*. Oxford University Press, 2nd edition, 2009.

[9] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The Open Quantum Materials Database

(OQMD): Assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(1):15010, Dec 2015. URL: http://dx.doi.org/10.1038/npjcompumats.2015.10, doi:10.1038/npjcompumats.2015.10.

[10] A. L. Spek. Single-crystal structure validation with the program PLATON. *Journal of Applied Crystallography*, 36:7–13, 2003.

[11] R.L. Harlow. Troublesome crystal structures. Prevention, detection, and resolution. *Journal of Research of the National Institute of Standards and Technology*, 101(3):327, May 1996. URL: http://dx.doi.org/10.6028/jres.101.034, doi:10.6028/jres.101.034.

[12] Jason C. Cole, Ilenia Giangreco, and Colin R. Groom. Using more than 801296 small-molecule crystal structures to aid in protein structure refinement and analysis. *Acta Crystallographica Section D*, 73(3):234–239, Mar 2017. URL: https://doi.org/10.1107/S2059798316014352, doi:10.1107/S2059798316014352.

[13] Robin Taylor, Jason Cole, Oliver Korb, and Patrick McCabe. Knowledge-based libraries for predicting the geometric preferences of druglike molecules. *Journal of Chemical Information and Modeling*, 54(9):2500–2514, Sep 2014. URL: http://dx.doi.org/10.1021/ci500358p, doi:10.1021/ci500358p.

[14] Pierre Baldi. Data-driven high-throughput prediction of the 3-D structure of small molecules: review and progress. A response to the letter by the Cambridge Crystallographic Data Centre. *Journal of chemical information and modeling*, 51:3029, 2011. URL: http://pubs.acs.org/doi/abs/10.1021/ci200460z, doi:10.1021/ci200460z.

[15] Saulius Gražulis, Adriana Daškevič, Andrius Merkys, Daniel Chateigner, Luca Lutterotti, Miguel Quirós, Nadezhda R. Serebryanaya, Peter Moeck, Robert T. Downs, and Armel Le Bail. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research*, 40(D1):D420–D427, Jan 2012. URL: http://nar.oxfordjournals.org/content/40/D1/D420.abstract, doi:10.1093/nar/gkr900.

[16] Denis Cousineau and Sylvain Chartier. Outliers detection and treatment: a review. *International Journal of Psychological Research*, 3(1):58–67, 2010. URL: http://revistas.usb.edu.co/index.php/IJPR/article/view/844.

[17] S. R. Hall, F. H. Allen, and I. D. Brown. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47(6):655–685, Nov 1991. URL: http://dx.doi.org/10.1107/S010876739101067X, doi:10.1107/S010876739101067X.

[18] Larry Wall, Tom Christiansen, and Jon Orwant. *Programming Perl*. O'Reilly Media, third edition, July 2000.

[19] Andrius Merkys, Antanas Vaitkus, Justas Butkus, Mykolas Okulič-Kazarinas, Visvaldas Kairys, and Saulius Gražulis. *COD::CIF::Parser*: an error-correcting CIF parser for the

Perl language. *Journal of Applied Crystallography*, 49(1):292–301, Feb 2016. URL: http://dx.doi.org/10.1107/S1600576715022396, doi:10.1107/S1600576715022396.

[20] C. Levinthal. Molecular model-building by computer. *Sci Am.*, 214:42–52, 1966.

[21] H. L. Morgan. The generation of a unique machine description for chemical structures – a technique developed at chemical abstracts service. *J. Chem. Doc.*, 5:107–113, 1965. doi:10.1021/c160017a018.

[22] The Cambridge Crystallographic Data Centre. Element data and radii [online]. 2008. URL: https://web.archive.org/web/20080701015237/http://www.ccdc.cam.ac.uk/products/csd/radii/table.php4.

[23] F. H. Allen, S. Bellard, M. D. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B. G. Hummelink-Peters, O. Kennard, W. D. S. Motherwell, J. R. Rodgers, and D. G. Watson. The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallographica Section B*, 35(10):2331–2339, Oct 1979. URL: http://dx.doi.org/10.1107/S0567740879009249, doi:10.1107/S0567740879009249.

[24] Elaine C. Meng and Richard A. Lewis. Determination of molecular topology and atomic hybridization states from heavy atom coordinates. *Journal of Computational Chemistry*, 12(7):891–898, sep 1991. doi:10.1002/jcc.540120716.

[25] Pekka Pyykkö and Michiko Atsumi. Molecular single-bond covalent radii for elements 1–118. *Chemistry – A European Journal*, 15:186–197, 2009. URL: http://dx.doi.org/10.1002/chem.200800987, doi:10.1002/chem.200800987.

[26] Fei Long, Robert A. Nicholls, Paul Emsley, Saulius Gražulis, Andrius Merkys, Antanas Vaitkus, and Garib N. Murshudov. ACEDRG: A stereo-chemical description generator for ligands. *Acta Crystallographica Section D*, 73(2):112–122, Feb 2017. doi:10.1107/S2059798317000067.

[27] Fei Long, Robert A. Nicholls, Paul Emsley, Saulius Gražulis, Andrius Merkys, Antanas Vaitkus, and Garib N. Murshudov. Validation and extraction of stereochemical information from small molecular databases. *Acta Crystallographica Section D*, 73(2):103–111, Feb 2017. doi:10.1107/S2059798317000079.

[28] B. Dittrich, C. B. Hübschle, K. Pröpper, F. Dietrich, T. Stolper, and J. J. Holstein. The generalized invariom database (GID). *Acta Crystallogr Sect B Struct Sci*, 69(2):91–104, Mar 2013. URL: http://dx.doi.org/10.1107/S2052519213002285, doi:10.1107/s2052519213002285.

[29] Sławomir Domagała, Bertrand Fournier, Dorothee Liebschner, Benoît Guillot, and Christian Jelsch. An improved experimental databank of transferable multipolar atom models – ELMAM2. construction details and applications. *Acta Crystallogr Sect A*, 68(3):337–351, Mar 2012. URL: http://dx.doi.org/10.1107/S0108767312008197, doi:10.1107/s0108767312008197.

[30] Geoffrey M. Downs, Valerie J. Gillet, John D. Holliday, and Michael F. Lynch. Review of ring perception algorithms for chemical graphs. *J. Chem. Inf. Comput. Sci.*, 29:172–187, 1989.

[31] John Figueras. Ring perception using breadth-first search. *J. Chem. Inf. Comput. Sci.*, 36:986–991, 1996.

[32] T. Hanser, P. Jauffret, and G. Kaufmann. A new algorithm for exhaustive ring perception in a molecular graph. *Journal of Chemical Information and Modeling*, 36:1146–1152, 1996. URL: `http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci960322f`, `doi:10.1021/ci960322f`.

[33] Andrew R. Leach, Daniel P. Dolata, and Keith Prout. Automated conformational analysis and structure generation: algorithms for molecular perception. *Journal of Chemical Information and Modeling*, 30(3):316–324, Aug 1990. URL: `http://dx.doi.org/10.1021/ci00067a017`, `doi:10.1021/ci00067a017`.

[34] Morris Plotkin. Mathematical basis of ring-finding algorithms in CIDS. *Journal of Chemical Documentation*, 11(1):60–63, feb 1971. `doi:10.1021/c160040a013`.

[35] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978. The PDF is also available at: http://www.andrew.cmu.edu/user/kk3n/simplicity/schwarzbic.pdf. URL: `http://projecteuclid.org/euclid.aos/1176344136`, `doi:10.1214/aos/1176344136`.

[36] Dankmar Bohning, Peter Schlattmann, and Bruce Lindsay. Computer-assisted analysis of mixtures (c.a.man): Statistical algorithms. *Biometrics*, 48(1):283–303, Mar 1992. URL: `http://dx.doi.org/10.2307/2532756`, `doi:10.2307/2532756`.

[37] Ferenc Nagy. Parameter estimation of the Cauchy distribution in information theory approach. *Journal of Universal Computer Science*, 12:1332–1344, 2006. URL: `http://www.jucs.org/jucs_12_9/parameter_estimation_of_the/jucs_12_09_1332_1344_nagy.pdf`, `doi:10.3217/jucs-012-09-1332`.

[38] Kurt Hornik and Bettina Grün. movMF: An R package for fitting mixtures of von Mises-Fisher distributions. *Journal of Statistical Software*, 58(10), July 2014. URL: `https://www.jstatsoft.org/article/view/v058i10/v58i10.pdf`.

[39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL: `https://www.R-project.org`.

[40] Daniel Pena and Irwin Guttman. Comparing probabilistic methods for outlier detection in linear models. *Biometrika*, 80(3):603, Sep 1993. URL: `http://www.jstor.org/stable/2337181`, `doi:10.2307/2337181`.

[41] Harold Jeffreys. *The Theory of Probability*. Oxford, 3 edition, 1961.

[42] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, Jun 1995. URL: http://dx.doi.org/10.1080/01621459.1995.10476572, doi:10.1080/01621459.1995.10476572.

[43] Bruno Bienfait and Peter Ertl. JSME: a free molecule editor in JavaScript. *Journal of Cheminformatics*, 5:24, 2013. URL: http://www.jcheminf.com/content/5/1/24, doi:10.1186/1758-2946-5-24.

[44] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, Feb 1988. URL: http://dx.doi.org/10.1021/ci00057a005, doi:10.1021/ci00057a005.

[45] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33, 2011. doi:10.1186/1758-2946-3-33.

[46] Robert M. Hanson. *Jmol* – a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography*, 43:1250–1260, 2010. URL: http://dx.doi.org/10.1107/S0021889810030256, doi:10.1107/S0021889810030256.

[47] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, Bjørk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W. Jacobsen. The atomic simulation environment – a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, jun 2017. doi:10.1088/1361-648x/aa680e.

[48] S. R. Hall and H. J. Bernstein. CIF Applications. V. *ciftbx*2: extended tool box for manipulating CIFs. *Journal of Applied Crystallography*, 29(5):598–603, Oct 1996. URL: http://dx.doi.org/10.1107/S0021889896006371, doi:10.1107/S0021889896006371.

[49] John C. Bollinger. A portable general-purpose application programming interface for CIF 2.0. *J Appl Crystallogr*, 49(1):285–291, Feb 2016. URL: http://dx.doi.org/10.1107/S1600576715021883, doi:10.1107/s1600576715021883.

[50] Marcin Wojdyr. *Gemmi - General MacroMolecular I/O*. Global Phasing Ltd., GIT commit 860d28508767752288ae74c3737ab602f444a896 edition, 2017.

[51] J. R. Hester. A validating CIF parser: *PyCIFRW*. *Journal of Applied Crystallography*, 39(4):621–625, Aug 2006. URL: http://dx.doi.org/10.1107/S0021889806015627, doi:10.1107/S0021889806015627.

[52] W. Bluhm. STAR (CIF) parser, 2000. URL: http://pdb.sdsc.edu/STAR/index.html.

[53] Peter A. Keller. A lexical analyser for STAR/CIF/mmCIF data, Sep 2013. URL: `http://www.globalphasing.com/startools/StarTools_article.pdf`.

[54] Richard J. Gildea, Luc J. Bourhis, Oleg V. Dolomanov, Ralf W. Grosse-Kunstleve, Horst Puschmann, Paul D. Adams, and Judith A. K. Howard. iotbx.cif: a comprehensive CIF toolbox. *Journal of Applied Crystallography*, 44(6):1259–1263, December 2011. `doi:10.1107/S0021889811041161`.

[55] Brian McMahon. *vcif*, volume G, chapter 5.3.2.1, pages 499–501. International Union of Crystallography, 2006.

[56] Georgi Todorov and Herbert J. Bernstein. *VCIF2*: extended CIF validation software. *Journal of Applied Crystallography*, 41(4):808–810, Aug 2008. URL: `http://dx.doi.org/10.1107/S002188980801385X`, `doi:10.1107/S002188980801385X`.

[57] David R. Stampf. ZINC – galvanizing CIF to work with UNIX, 2004. URL: `http://www.iucr.org/__data/iucr/cif/software/zinc/doc/zinc-paper.pdf`.

[58] Sydney R. Hall and Nick Spadaccini. The STAR file: Detailed specifications. *Journal of Chemical Information and Computer Sciences*, 34(3):505–508, 1994. URL: `http://dx.doi.org/10.1021/ci00019a005`, `doi:10.1021/ci00019a005`.

[59] Barry W. Boehm. *Software Engineering Economics*. Prentice Hall, 1981. URL: `http://csse.usc.edu/csse/research/COCOMOII/cocomo81.htm`.

[60] Miguel Quirós Olozábal, Saulius Gražulis, Saulė Girdzijauskaitė, Andrius Merkys, and Antanas Vaitkus. Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database. Accepted to Journal of Cheminformatics, 2018.

[61] Antanas Vaitkus. Asmeninis bendravimas.

[62] Wouter G. Touw and Gert Vriend. On the complexity of Engh and Huber refinement restraints: The angle τ as example. *Acta Crystallographica Section D*, 66:1341–1350, 2010. `doi:10.1107/S0907444910040928`.

[63] Nicole Balasco, Luciana Esposito, and Luigi Vitagliano. Factors affecting the amplitude of the τ angle in proteins: a revisitation. *Acta Crystallographica Section D*, 73(7):618–625, Jul 2017. URL: `https://doi.org/10.1107/S2059798317007793`, `doi:10.1107/S2059798317007793`.

[64] Richard A. Engh and Robert Huber. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica Section A*, 47:392–400, 1991.

[65] Roman A. Laskowski, David S. Moss, and Janet M. Thornton. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.*, 231:1049–1067, 1993.

[66] R. A. Engh and R. Huber. *International Tables for Crystallography, Vol. F*, pages 382–392. Kluwer Academic Publishers, 2001.

[67] Majorie M. Harding. The geometry of metal-ligand interactions relevant to proteins. *Acta Crystallographica Section D*, 55:1432–1443, 1999. URL: `http://dx.doi.org/10.1107/S0907444999007374`, `doi:10.1107/S0907444999007374`.

[68] William T. A. Harrison, Jim Simpson, and Matthias Weil. Editorial. *Acta Crystallographica Section E*, 66:e1–e2, 2010. `doi:10.1107/S1600536809051757`.

[69] Oleg V. Dolomanov, Luc J. Bourhis, Richard J. Gildea, Judith A. K. Howard, and Horst Puschmann. OLEX2: a complete structure solution, refinement and analysis program. *Journal of Applied Crystallography*, 42(2):339–341, jan 2009. `doi:10.1107/s0021889808042726`.

[70] H. Zhong, X.-M. Yang, Q.-Y. Luo, and Y.-P. Xu. (1,10-phenanthroline)tri(3-phenylpropanoato)lanthanum(III). *Acta Crystallographica Section E Structure Reports Online*, 63(7):m1909–m1909, jun 2007. `doi:10.1107/s1600536807028693`.

[71] Kliment Olechnovič and Česlovas Venclovas. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. *Journal of Computational Chemistry*, 35:672–681, 2014. `doi:10.1002/jcc.23538`.