

Üleminekumetalliühendi konformatsioonianalüüs andmeteadusest inspireeritud töövooga

Urmas Pitsi
Infotehnoloogia teaduskond

16.01.2023

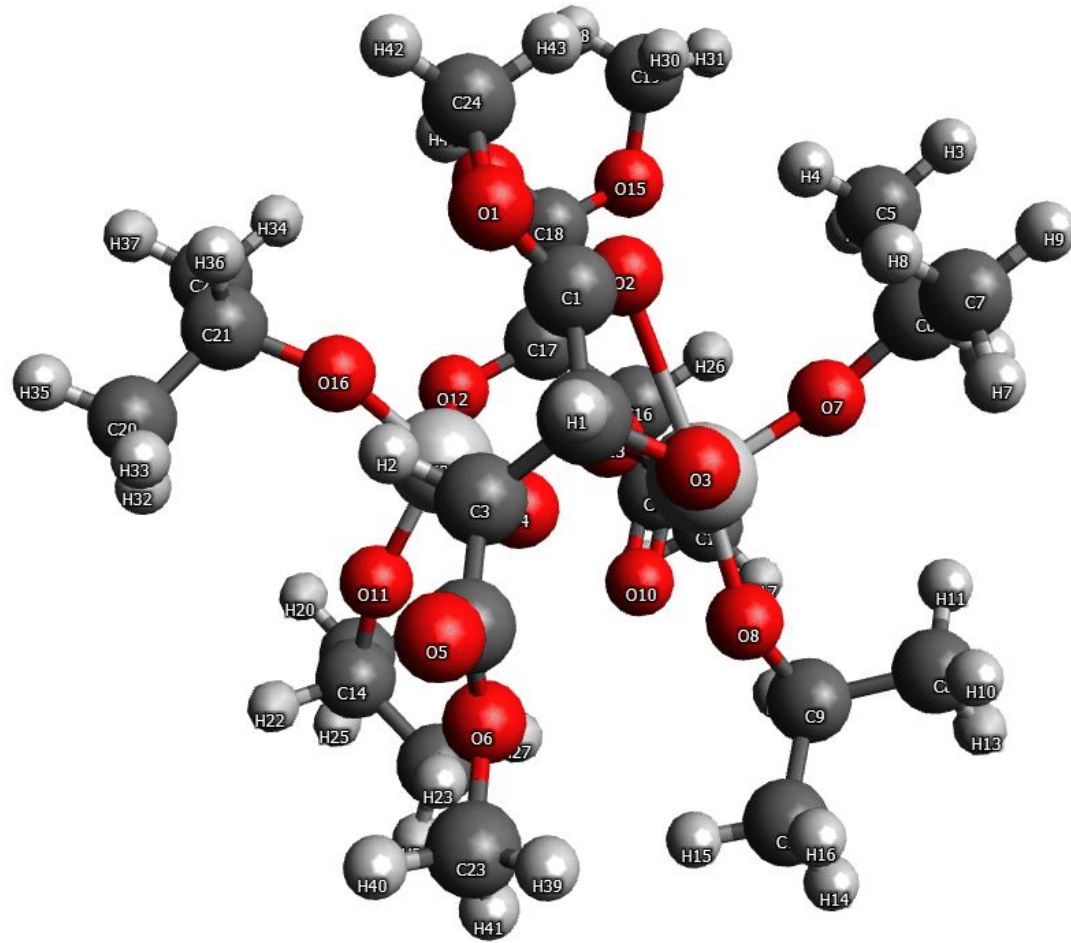
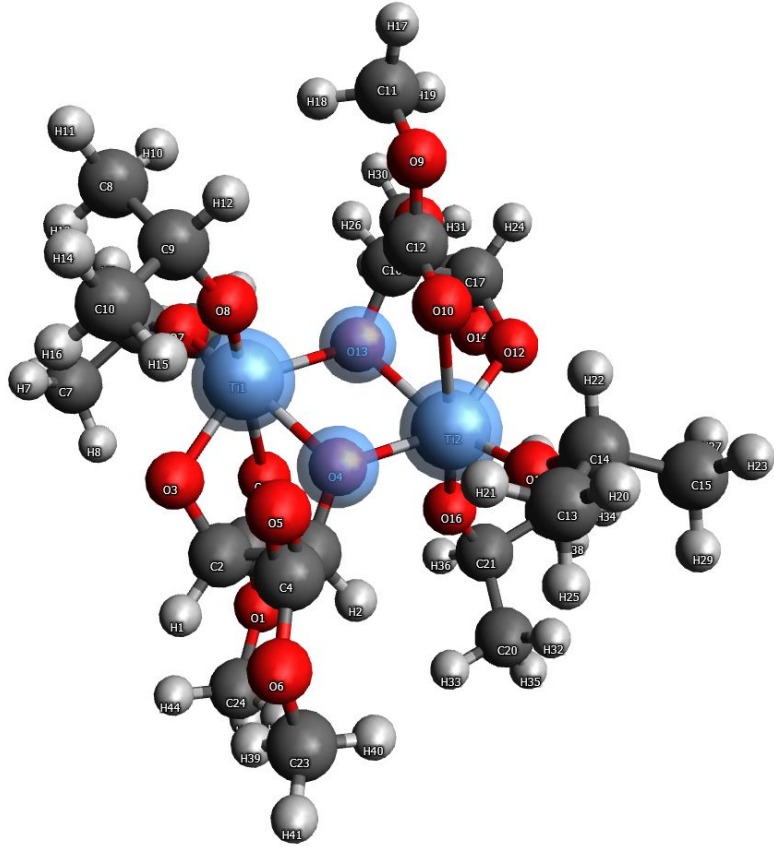
Sissejuhatus

- Teaduskondadevaheline koostöö: teeme midagi kasulikku keemikute jaoks
- Perioodilisustabel, üleminekumetallid sinine blokk

	1	2																	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		
1	H																																			He
2	Li	Be																												B	C	N	O	F	Ne	
3	Na	Mg																													Al	Si	P	S	Cl	Ar
4	K	Ca																	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr		
5	Rb	Sr																	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe		
6	Cs	Ba	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn				
7	Fr	Ra	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Nh	Fl	Mc	Lv	Ts	Og				
s-block			f-block														d-block										p-block									

Allikas: [https://en.wikipedia.org/wiki/Block_\(periodic_table\)#d-block](https://en.wikipedia.org/wiki/Block_(periodic_table)#d-block)

Otsime madalaima energiaga geomeetriat

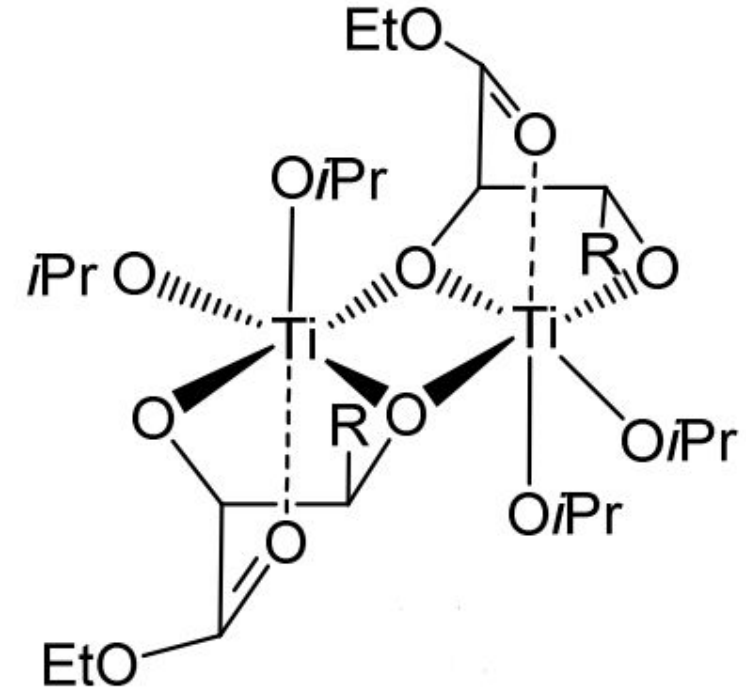


Arvutuskeemia meetodid ja terminid

- Molekulaarmehaanilised, poolempiirilised ja kvantmehaanika meetodid
- DFT: tihedusfunktsionaalide teooria (Density Functional Theory)
- DFT meetod (mudelkeemia): kvantmehaanikal põhinev DFT energiaarvutusmeetod
- Mudelkeemia = funktsionaal + baasikomplekt
- Näide (funktsionaal/baasikomplekt): PBE0/def2-SV(P), BP86/def2-SV(P), PBE0/cc-PVTZ
- Kuidas valida sobiv mudelkeemia uuritavale probleemile?
- Meie jaoks on meetodid nn “must kast”

Uurimisobjekt: titaantartraat (TargetMol)

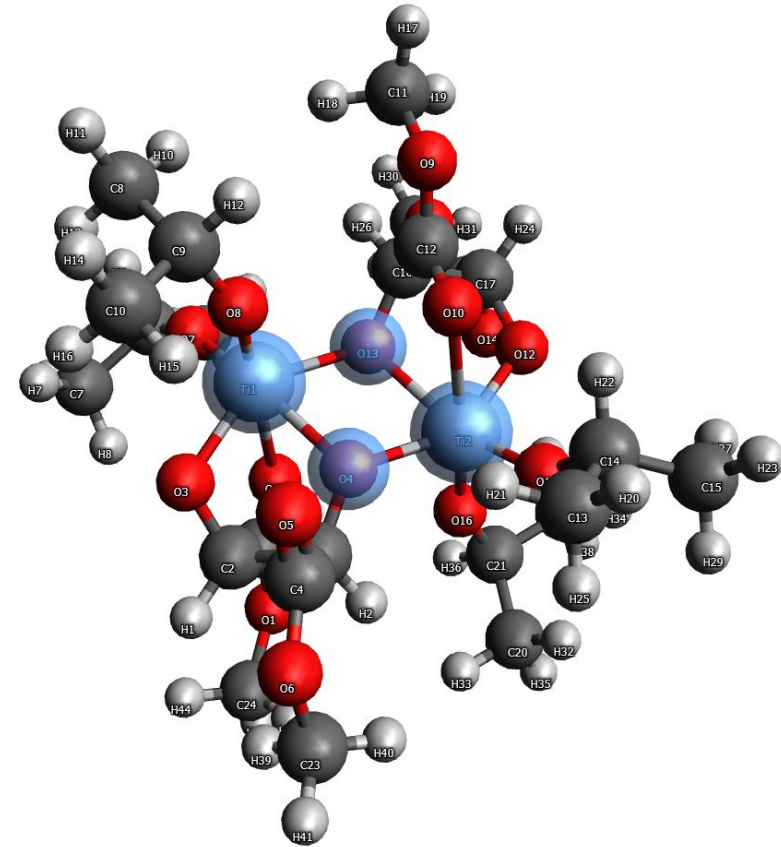
- 86 aatomit, sealhulgas 2 titaani aatomit
- Sharplessi epoksüdatsioon (1984), Nobeli preemia (2001)
- Käelised (kiraalsed) katalüsaatorid: uute orgaaniliste ühendite süntees, ravimid jne.
- Iseloomulik keskne aatomite nelik: Ti_2O_2 , moodustab peaaegu sümmeetrilise tasandilise rombi
- Motivatsioon: osa poolelilolevast suuremast projektist, käeliste ühendite süntees → toimimise mehhanism → vaja konformeere



Titaantartraadi keemiline struktuur

Konformatsioonianalüüsi töövoog

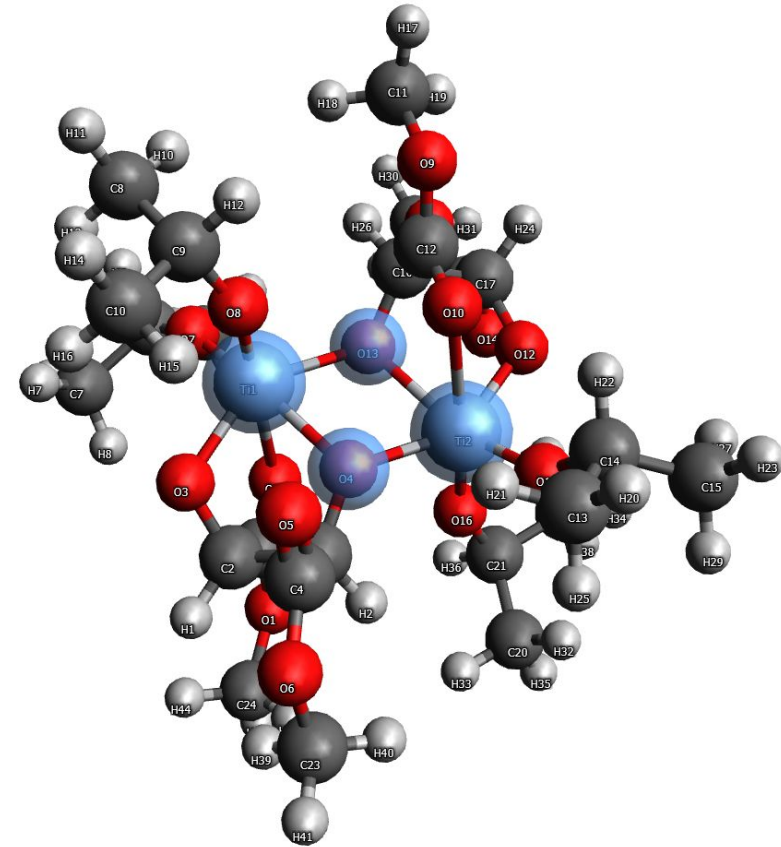
1. **Lähtegeomeetria:** juhuslikult genereeritud 3D koordinaadid, vastab soovitavale keemilisele struktuurile
2. **Kandidaatgeomeetriad:** liigutame lähtegeomeetria aatomeid ruumis
3. **Konformatsioonid:** optimeerime kandidaatgeomeetriad lokaalsesse energia miinimumi
4. **Konformeerid:** sorteerime konformatsioonid energia järgi, madalam energia on parem



TargetMol 3D-struktuur

Magistritöö eesmärgid

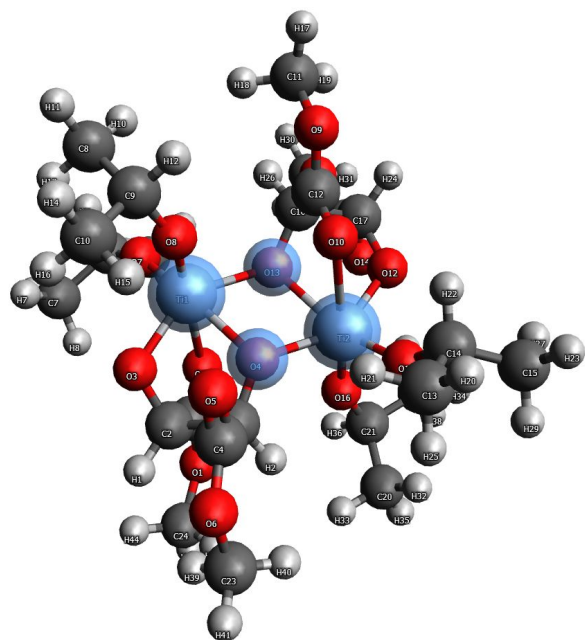
- **Leida konformeeride hulk** otsitavale titaantartraadi molekulile, kasutades vabavara CREST
- **Hinnata CREST-i sobivust** meie kontekstis: kas on üldse võimaline genereerima häid konformeere TargetMol-le
- **Hinnata** CREST-is kasutatavate energia-arvutusmeetodite **GFN2-xTB, GFN-FF sobivust** meie kontekstis. Kas genereeritud geomeetriad on korrektsed?
- **Leida võimalikke lühiteid**, mis aitaksid tulevikus hõlbustada konformatsioonianalüüsi töövoogu



TargetMol 3D-struktuur

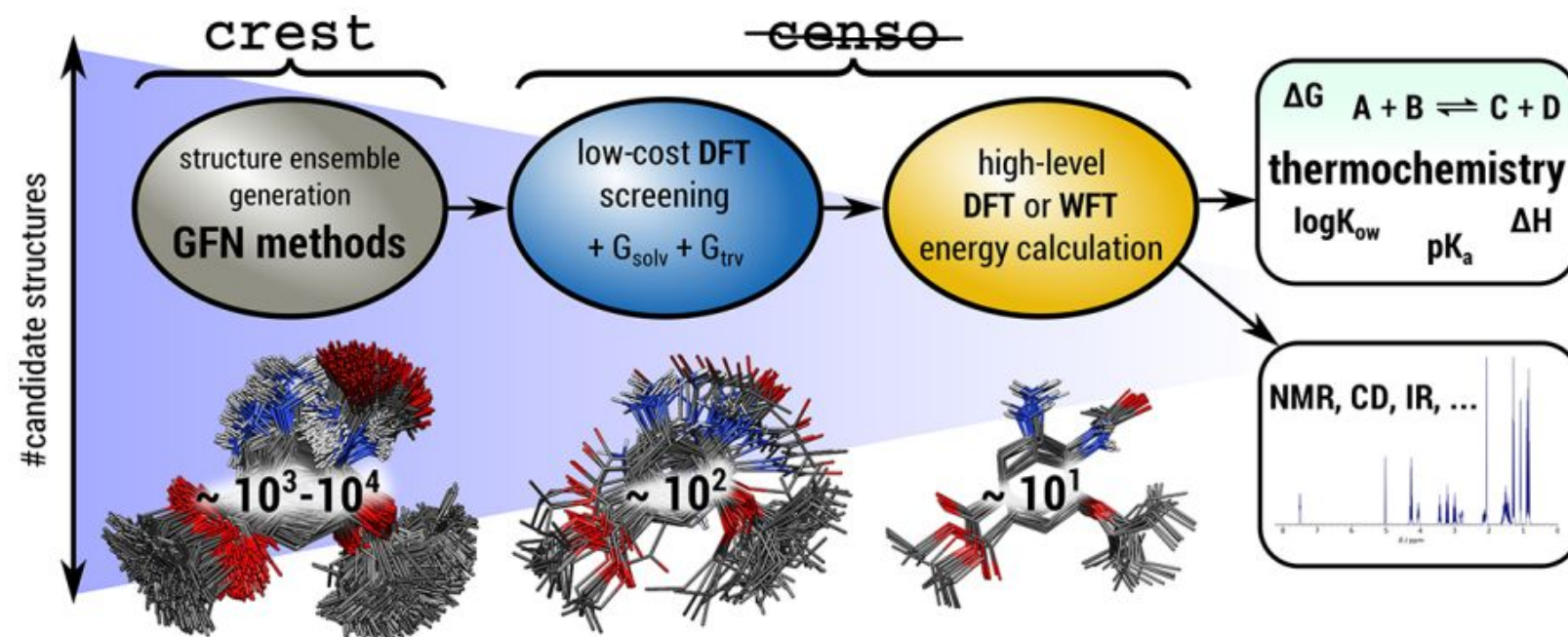
Konformatsioonianalüüsi töövoogi ülddiagramm

Algeomeetria (2011 a.)



CREST: 200 kandidaati

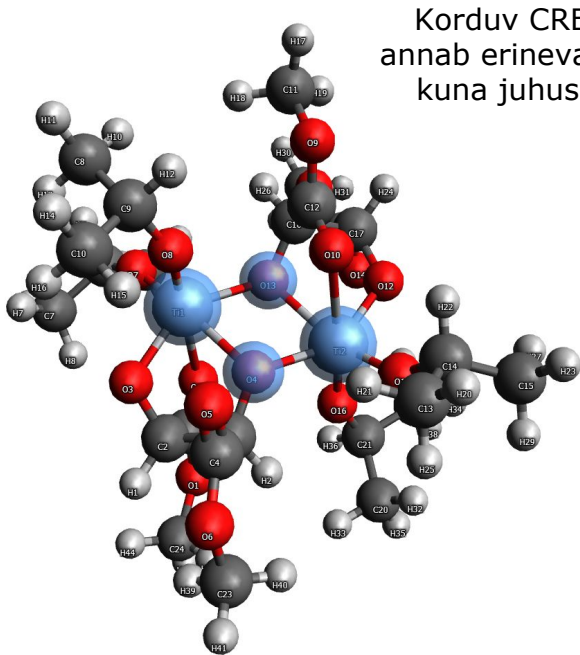
Magistritöös teostatud DFT arvutused



Allikas: Grimme et. al, Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules, J. Phys. Chem. A 2021, 125, 4039–4054

Geomeetria valik

Algeomeetria (2011 a.)



Uute kandidaatide iteratiivne loomine

CREST GFN2-xTB

CREST GFN-FF

Kandidaatgeomeetriad, 200 tk.
CREST-i tulemused

DFT osaline/jätkav
optimeerimine

DFT energiaarvutus

Järjestus energia järgi
Valida parimad (madalam on parem)

Edasine analüüs:
arvutada veel täpsema
DFT-ga omadused,
statistikud jms

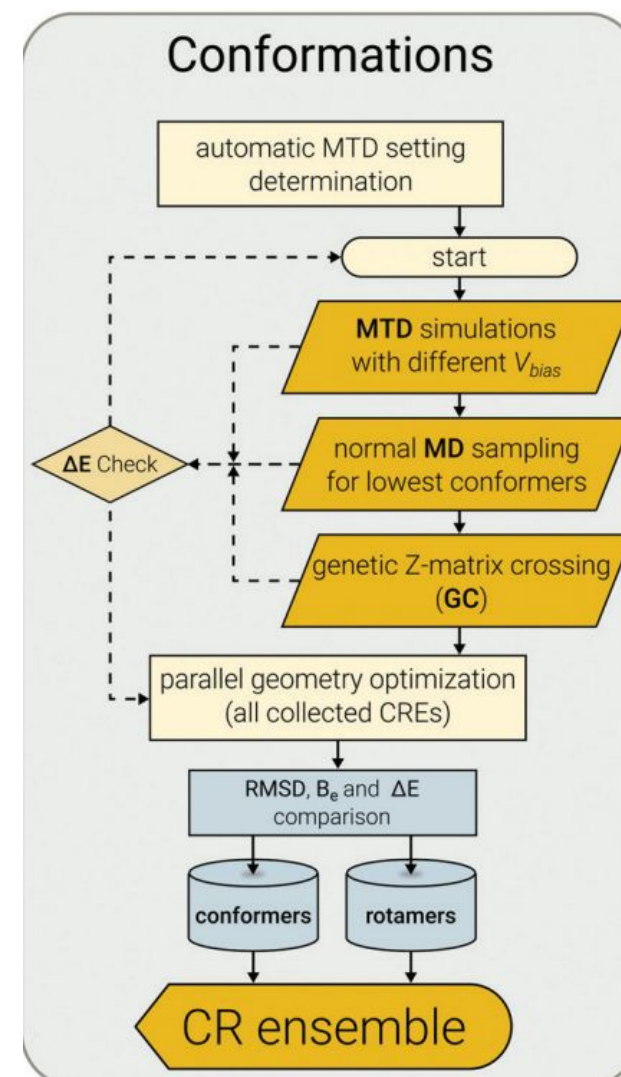
Lõpptulemus: otsitavad konformeerid

CREST (Conformer-Rotamer Ensemble Sampling Tool)

- Autorid: Prof. Dr. Stefan Grimme töögrupp Bonni ülikoolis
- **Effektiivne** konformatsioonide otsing
- **Kiired** energiaarvutusmeetodid: GFN2-xTB, GFN-FF
- **Unersaalne**: kuni elemendini 86, radoon
- **Vabavara**, autorid haldavad aktiivselt

CREST-i töövoo skeem, allikas:

Pracht et al., Automated exploration of the low-energy chemical space with fast quantum chemical methods, Phys. Chem. Chem. Phys., 2020, 22, 7169



DFT arvutuste üldinfo: 17 CPU aastat

DFT funktsionaal	Optimeerimis-samm, CPU tunnid	Koguaeg, CPU aastad	Koguaeg (wall time), päevad	Eksperimentide arv
PBE0/def2-SV(P)	4.3	6.7	40	269
↓ PBE0/cc-pVTZ	78.4	6.8	12	10 ↓
Muud		3.4	21	725
Kokku		16.9	73	1004

~ +3x

~ +20x

~ -20x

- **17 CPU aastat** teostatud DFT arvutuste kogumaht (Tabel 2, lk.32)
- **Arvutused skaleeruvad $O(n^3)$** suurendades keemilist täpsust
- **GFN-FF: 10^{-5} CPU tundi** (<0.04s)
- **GFN2-xTB: 10^{-3} CPU tundi** (<4s)

Suhteliste energiatega korrelatsioonid

1. Suhteliste energiatega korrelatsioon PBE0/cc-pVTZ-ga, võrdlus samal optimeerimissammul (Tabel 5, lk.44).

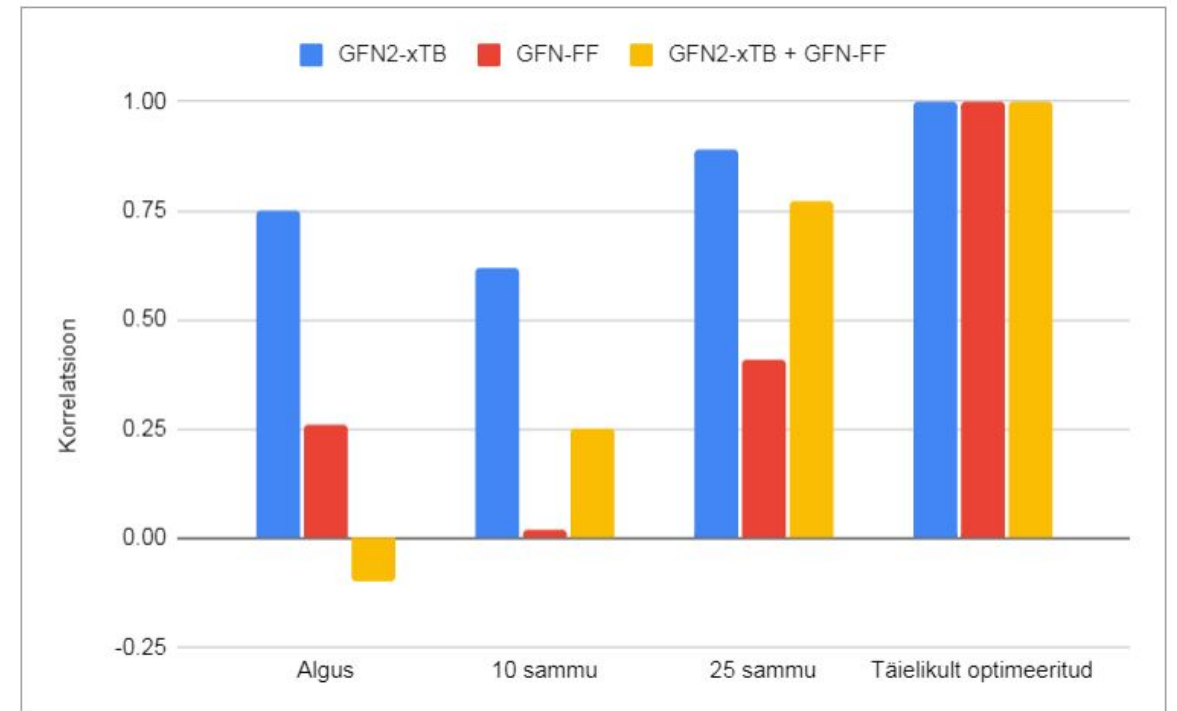
DFT Funktsionaal	Algus	10 sammu	25 sammu	Täielikult optimeeritud
PBE0/def2-SV(P)	0.999	0.994	0.994	0.992
Valimi suurus	8	8	8	8

2. Suhteliste energiatega korrelatsioon osaliselt ja täielikult optimeeritud geomeetria vahel, PBE0/def2-SV(P) (Tabel 8, lk.45).

Sisendite allikas	Algus	10 sammu	25 sammu	Täielikult optimeeritud
GFN2-xTB	0.75	0.62	0.89	1.00
GFN-FF	0.26	0.02	0.41	1.00
GFN2-xTB + GFN-FF	-0.10	0.25	0.77	1.00

1. Kas kiirem mudelkeemia sobib asendama täpsemat mudelkeemiat? JAH (20x ajavõit)

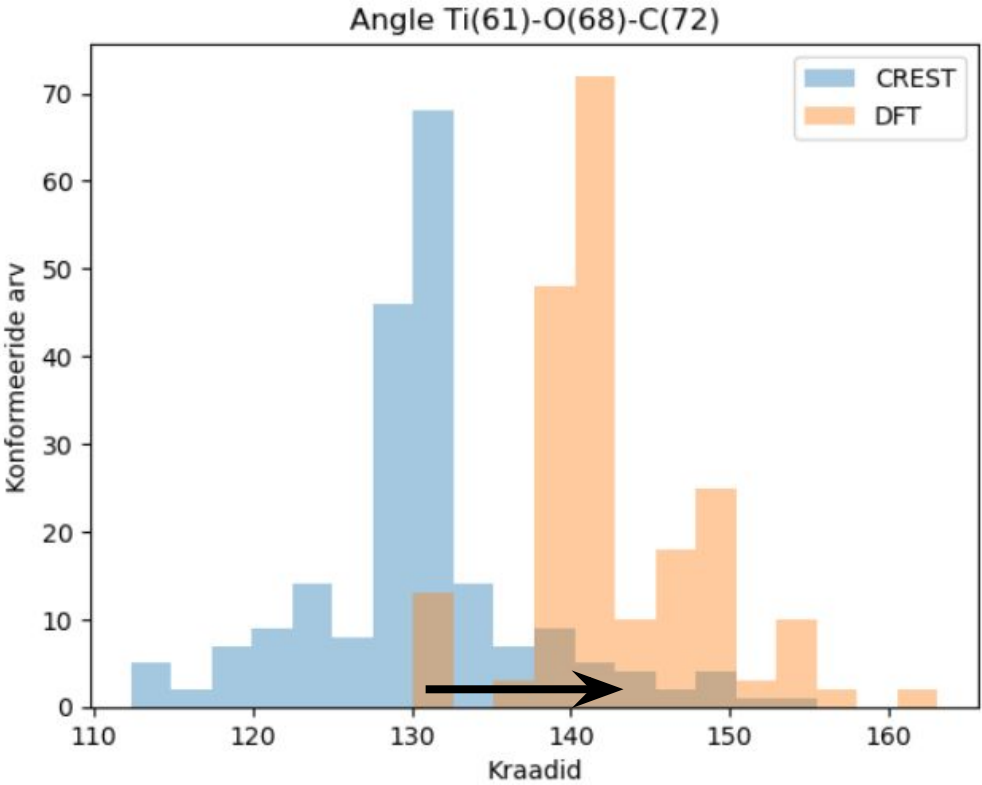
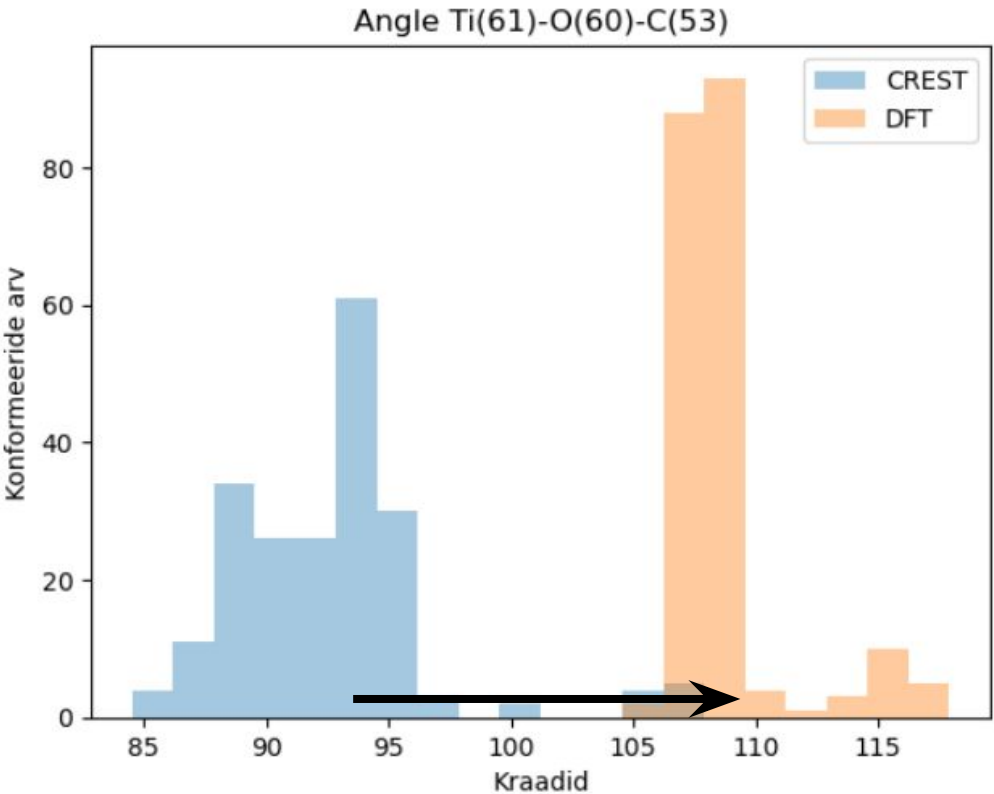
2. Kas saab kasutada osalist optimeerimist? JAH (2-3x ajavõit)

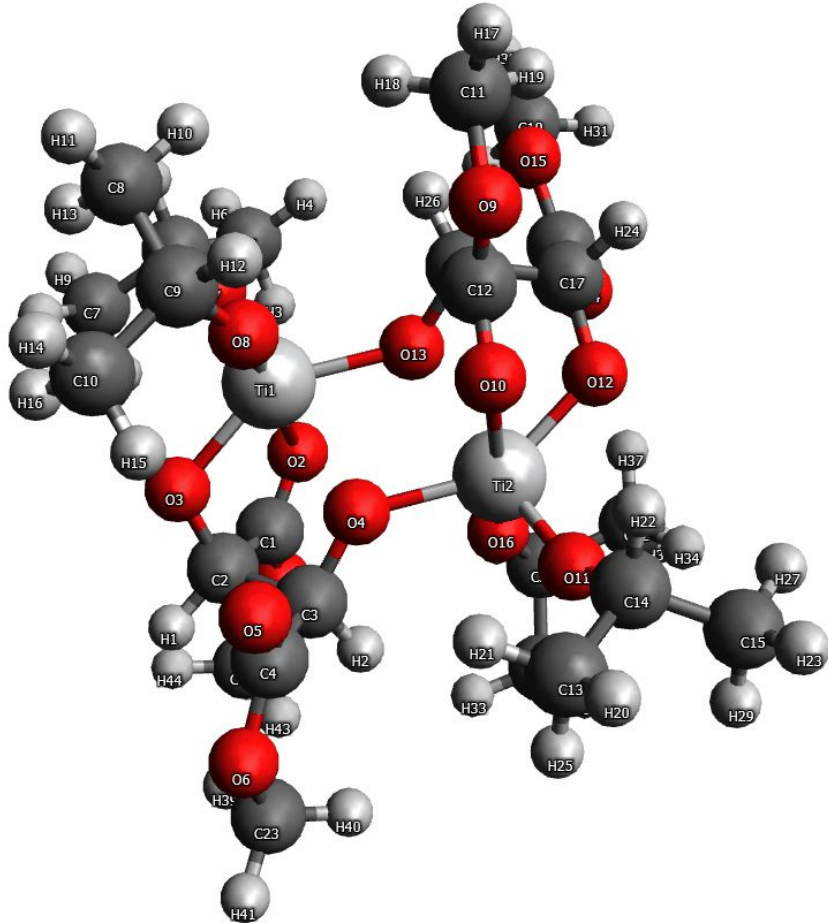


Ti-O-C nurkade analüüs: CREST (GFN) vs DFT

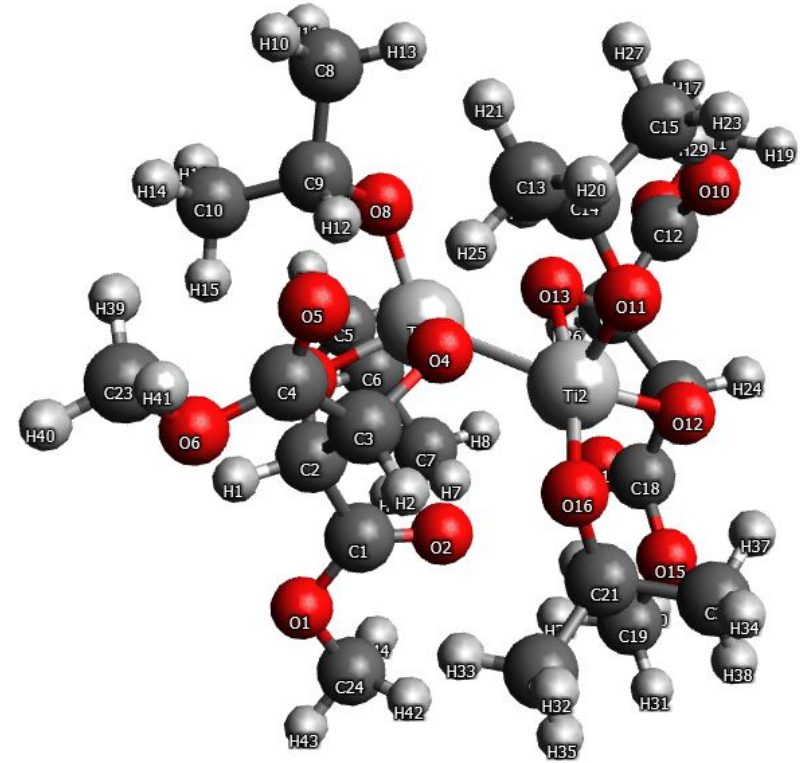
- 1. Ti-O-C nurgad kraadides (Tabel 13, lk.53).
- 2. Histogrammid kahest valitud nurgast. CREST väljund ja DFT optimeerimistulemus (Tabelid A3.4 ja A3.5, lk.76-77).

Nurk	Alguses ehk CREST		Lõpus ehk DFT		Erinevus keskmises
	keskmine	st.hälve	keskmine	st.hälve	
Ti(7)-O(20)-C(18)	139.7	11.5	154.4	5.3	14.7
Ti(7)-O(31)-C(29)	127.5	5.5	146.4	4.3	18.9
Ti(61)-O(56)-C(48)	134.4	8.9	150.4	5.1	16.0
Ti(61)-O(60)-C(53)	92.6	4.1	108.7	2.3	16.1
Ti(61)-O(68)-C(72)	130.6	6.9	143.0	5.7	12.4





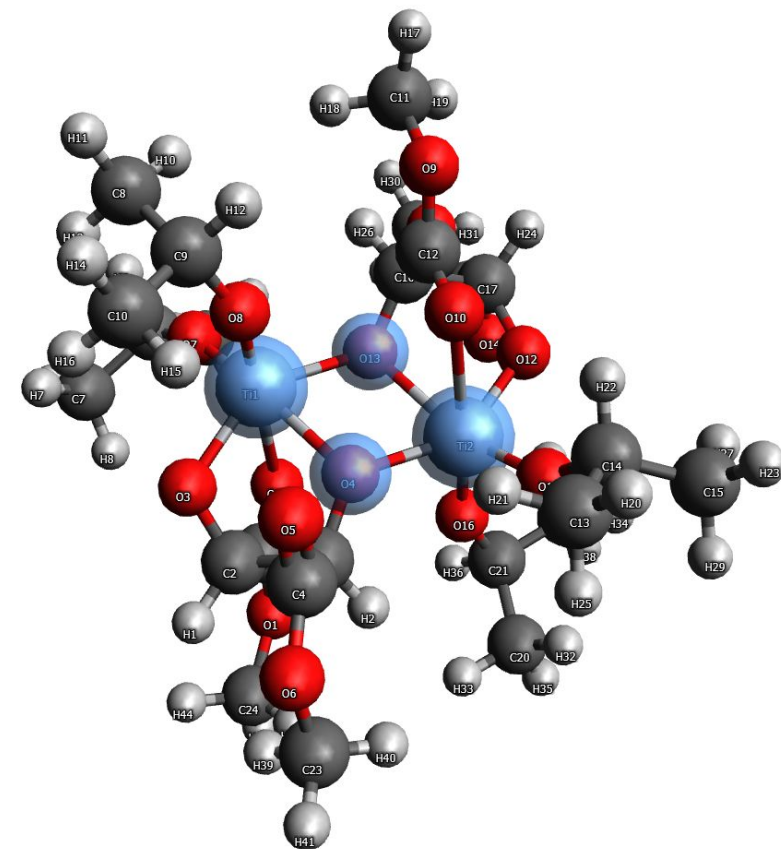
**CREST, GFN-FF optimeerib TargetMol
alggeomeetriat [BP86/def2-SV(P)]**



**DFT (PBE0/cc-PVTZ) optimeerib
GFN-FF geomeetriat**

Kokkuvõte

- **Leidsime konformeeride hulga** otsitavale titaantartraadi molekulile, kasutades vabavara CREST
- Tulemused näitavad, et **CREST on väga võimekas töövahend** meie kontekstis
- GFN meetodid genereerivad näiliselt deformeerunud geomeetriaid, edasine **optimeerimine DFT-ga parandab**
- **GFN-FF** meetod on väga hea kandidaat
- **Lühiteed: eel-optimeerimine kiire DFT-ga** teatud ulatuses aitab kiirendada valikuprotsessi, kui geomeetria valim on suur



TargetMol 3D-struktuur

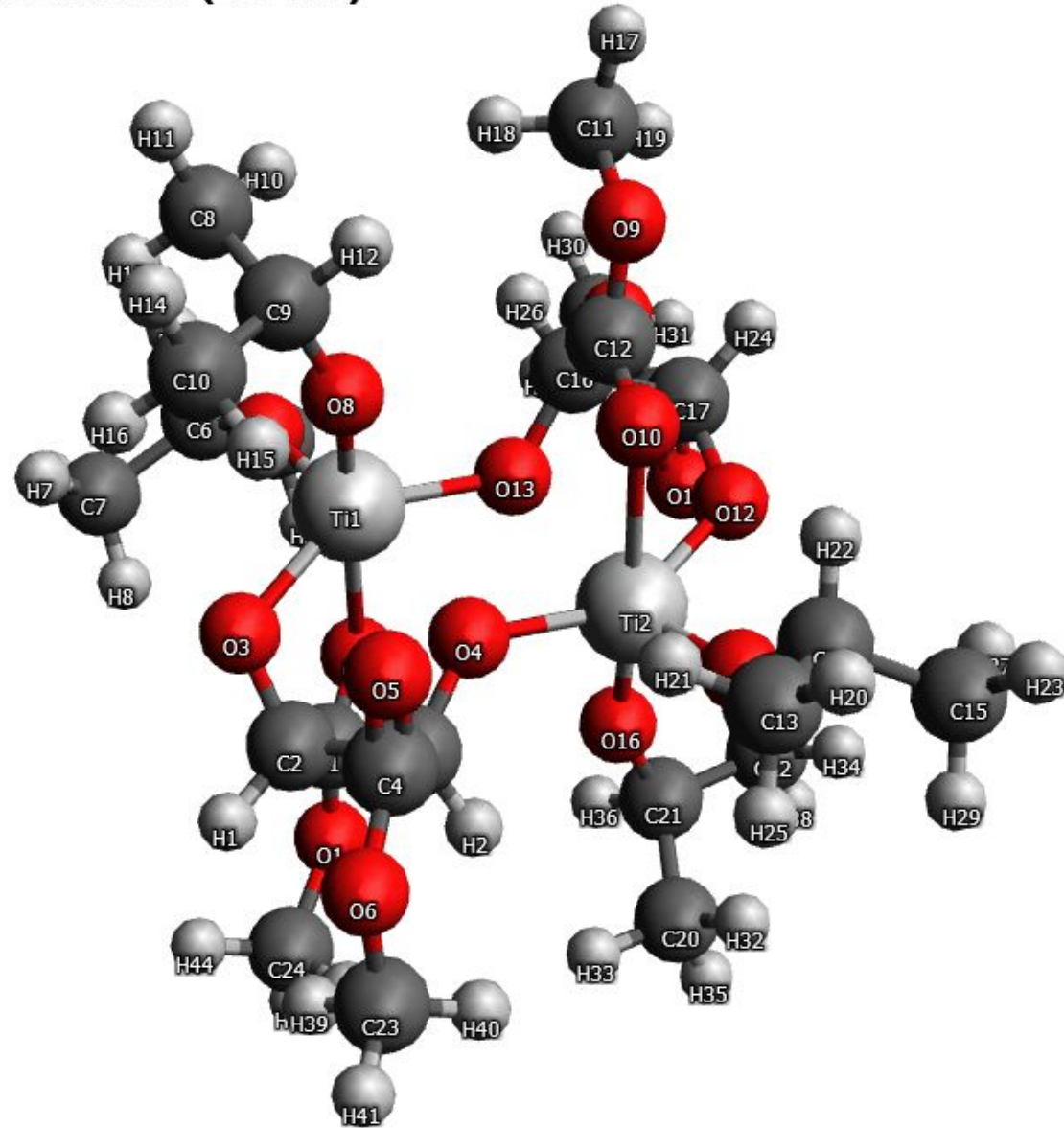
Aitäh!

Alggeomeetria (2011a.)

Parima konformeeri elutsükkel.

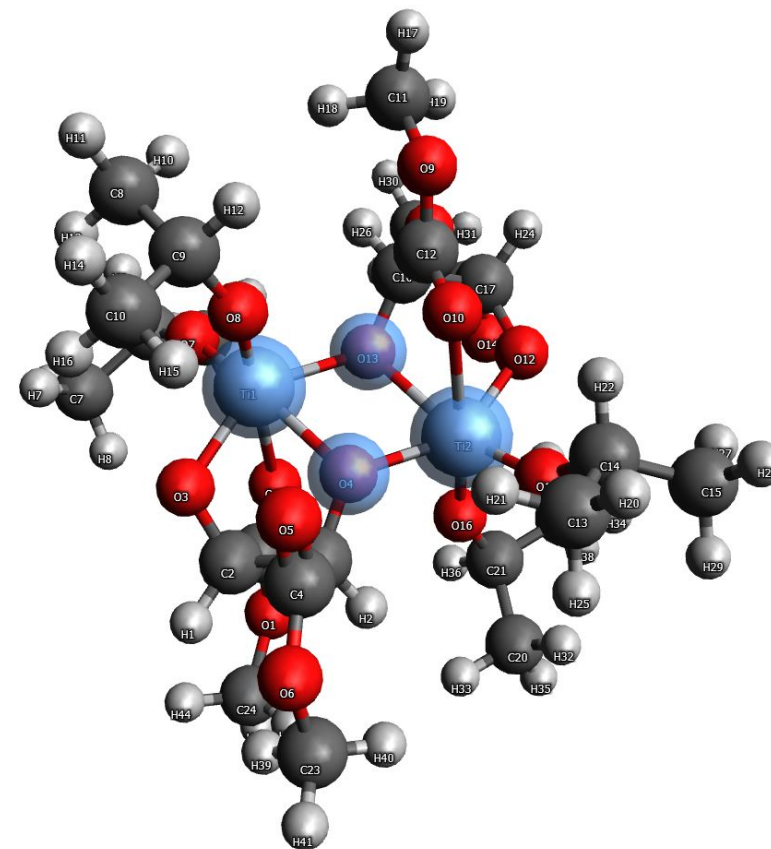
Konformeer ex21_c12

1. Alggeomeetria (2011 a.)
2. CREST GFN2-xTB, 5-s parim
3. CREST GFN-FF, 12-s parim
4. DFT optimeerimine, PBE0/def2-SV(P)
5. Energia 108 kJ/mol madalam



Magistritöö (numbrilised) tulemused

- Leidsime konformeeride hulga otsitavale titaantartraadi molekulile, kasutades vabavara CREST.
 - 8 konformeeeri mudelkeemiaga PBE0/cc-pVTZ, energiavahemikus 27 kJ/mol.
 - 205 konformeeeri mudelkeemiaga PBE0/def2-SV(P), energiavahemikus 67 kJ/mol.
- GFN meetodid genereerivad näiliselt deformeerunud geomeetriaid, edasine optimeerimine DFT-ga parandab.
- GFN-FF meetod on väga hea kandidaat
 - väga kiire: 30-70 korda kiirem kui GFN2-xTB, ehk u 100-1000+ korda kiirem kui DFT.
 - GFN-FF andis 100% top10, 75% top20 konformeeridest (21% sisenditest genereeritud GFN-FF meetodiga).
- GFN meetodid DFT ennustajatena: vastakad tulemused.
- Lühiteed: eel-optimeerimine kiire DFT-ga teatud ulatuses. 2 näidet, kus 10 sammu eeloptimeerimist vähendasid arvutusmahtu 2 korda.



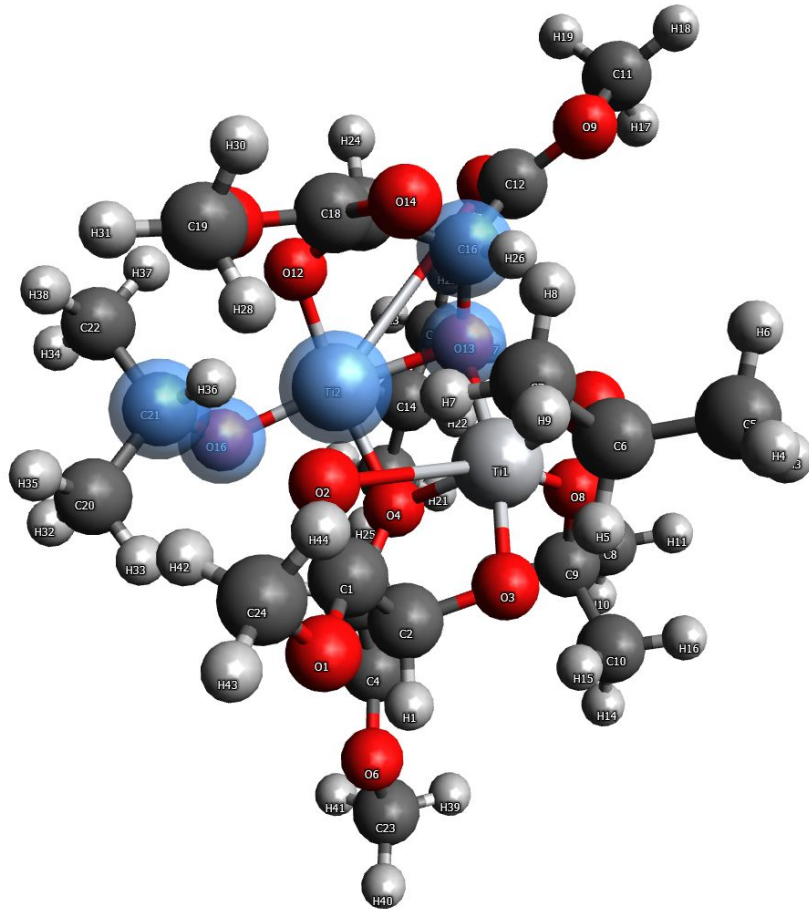
TargetMol 3D-struktuur

Teoreetiline taust

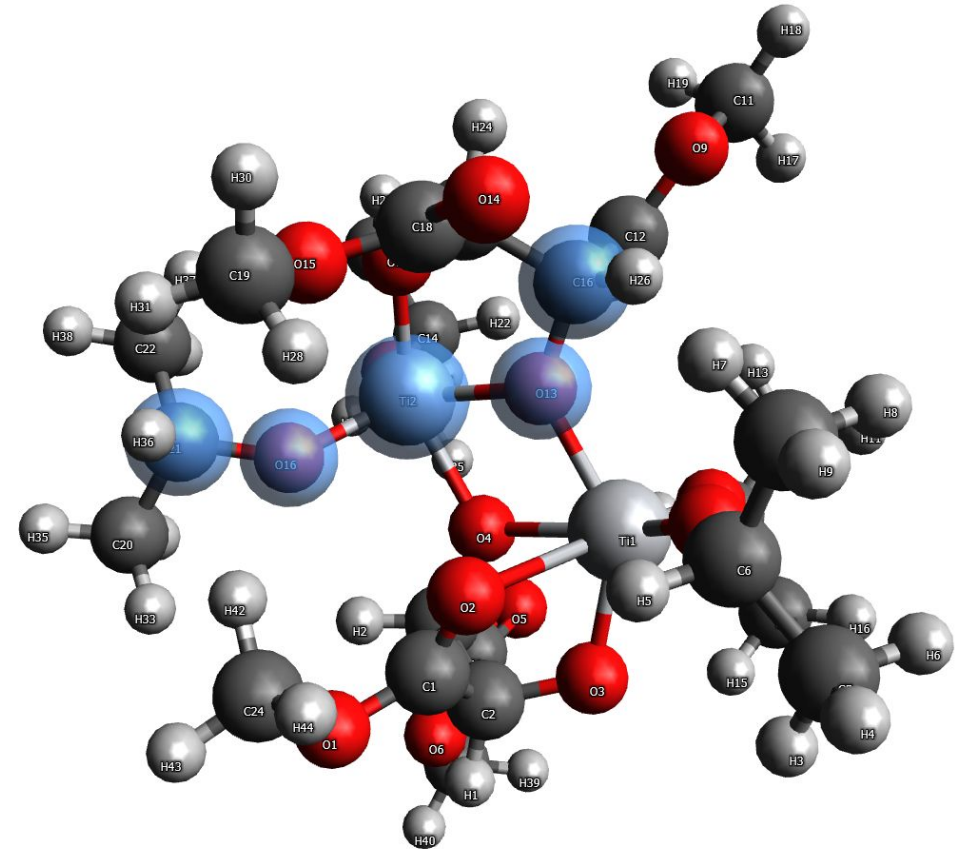
- Keemia: molekulid, aatomid, elektronid.
- Perioodilisustabel, üleminekumetallid: Ti, Cr, Fe, Co, Ni, Cu jne.
- Arvutuskeemia = keemia + arvutused
- Kvantmehaanika põhiteoreemidest: Schrödingeri võrrand: $H\Psi = E\Psi$
süsteemi omadused määratud lainefunktsiooniga, võimatu lahendada.
- Tihedusfunktsionaalide teooria: lähendus Schrödingeri võrrandi lahenduseks, tuletab süsteemi omadused elektronide tihedusest.
- Molekulaargeomeetria → energia = potentsiaalse energia hüperpind
- Konformatsioonianalüüs: otsime energeetiliselt sobivamaid geomeetriaid.

Valitud Ti-O-C nurgad

Urmas Pitsi 2023



CREST



DFT

Retsensent dr. Sven Nõmm, kriitika 1

1. The title of the thesis suggests a stronger connection to data science. The reviewer was unable to find the problem statement in terms used in data science. Also, the workflow itself was not properly described.

Vastus:

1. "Data Science **Inspired**", "Data-driven": populaarsed märksõnad, seose tugevus töö sisuga on objektiivselt määratlemata. Erinevad koolkonnad pealkirjastamisel. "Data Science Inspired" võib pealkirjast ära jätta.
2. Töö reaalne ulatus ja arusaamine valdkonnast muutusid. Leida sobiv kompromiss andmeteaduse (masinõppe) ja arvutuskeemia vahel.
3. Andmeteaduse seisukohalt oleme Titaantartraadi uurimisel pigem EDA (Exploratory Data Analysis) staadiumis, kuna me ei oska **küsida "õigeid" küsimusi**.

ChemRxiv[®]

[How To Submit](#) [Browse](#) [About](#) [News](#) [↗](#)

data science



15654 results for **data science**

14198 results for **data driven**

Theoretical and Computational Chemistry

Working Paper

Data-science driven autonomous process optimization

Melodie Christensen, Lars Yunker, Folarin Adedeji, Florian Häse, Loic Roch, Tobias Gens...

Categories

Retsensent dr. Sven Nõmm, kriitika 2

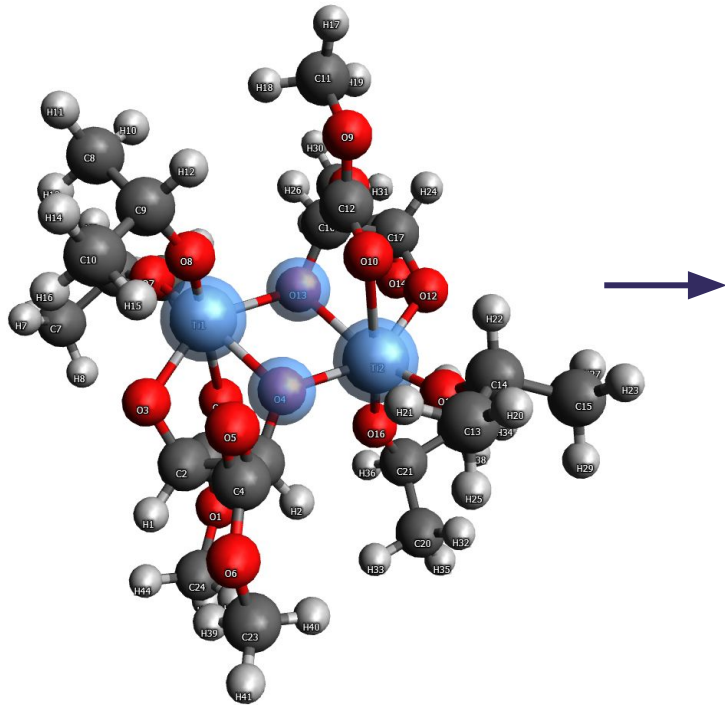
- Background information is given from the chemistry or computational chemistry perspective, whereas the computational part is basically missing. When presenting the concepts of computational chemistry, the author did not give any examples, allowing the reader to understand, try, and follow the concept. It is recommended to present such an example during the defence, preferably illustrated by a proper diagram. Also, immediately in the beginning, the author refers to “level of PBE0/ccpVTZ” without explicitly defining it. There are a number of similar references/statements throughout the text.

Vastus:

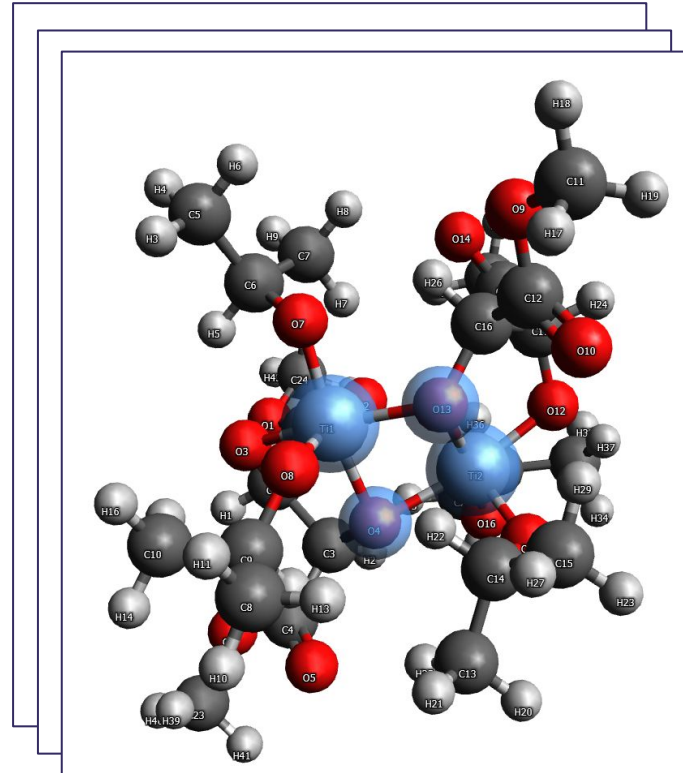
1. Vt slaidid 8, 9.

Geomeetria üldistatud töövoog

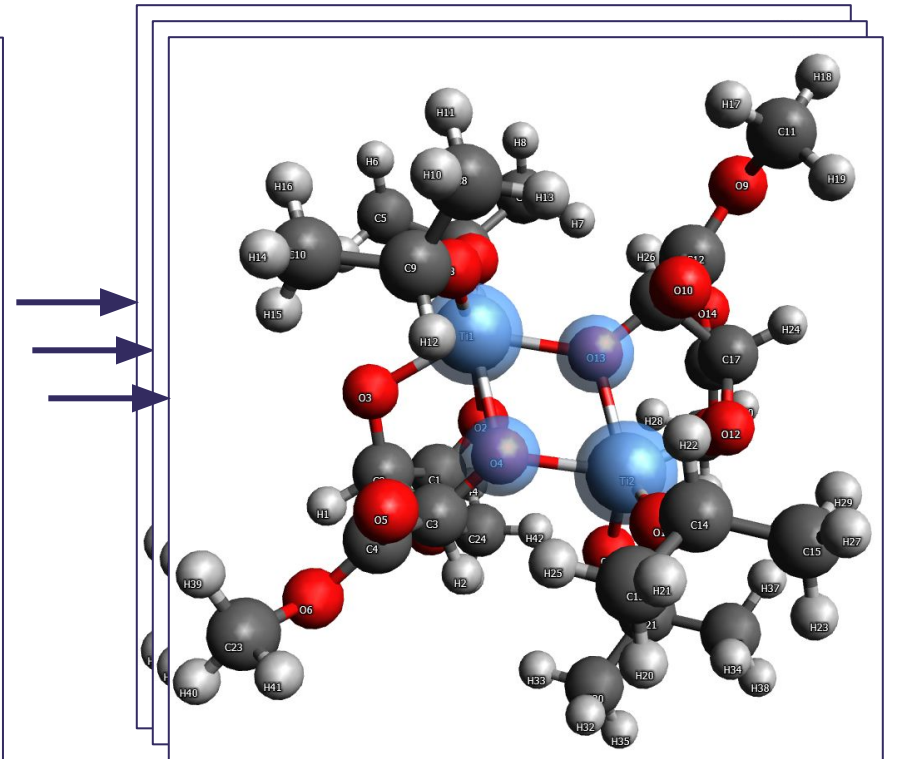
Lähtegeomeetria



CREST-i väljund



DFT optimeeritud



Retsensent dr. Sven Nõmm, kriitika 3

- Despite the large references section, the motivation to study this particular problem is not explained with references and is not positioned with respect to the literature. The absence of positioning makes it difficult to evaluate the novelty.

Vastus (vt ka slaid lk.5 Uurimisobjekt):

1. Osa suuremast projektist, mille detaile ei saa avaldada. Käeliste ühendite sünteesimiseks on vajalik kirjeldada reaktsiooni toimemehanismi, selleks on vaja konformeere. Mille käesolev töö leidis. Projekt ootel 10+ aastat.
2. Autor lähtus kuni dets. 2022 keskpaigani teadmisest, et magistritöö tuleb agnostiline, st ei sisalda detaile vaadeldavast molekulist: "TargetMol on 50-100 aatomist koosnev üleminekumetalli ühend".

Retsensent dr. Sven Nõmm, kriitika 4.1.

- In the first part of the thesis, the author did not present any validation framework and numeric thresholds to compare the results with. This makes it difficult to follow Subsections 5.6 and 5.8.

Vastus :

1. Peatükk 5.6 (lk.49-50) defineerib kriteeriumid üheselt, vastavus diskreetne JAH/EI:

5.6 Validation of structural correctness of molecular geometries

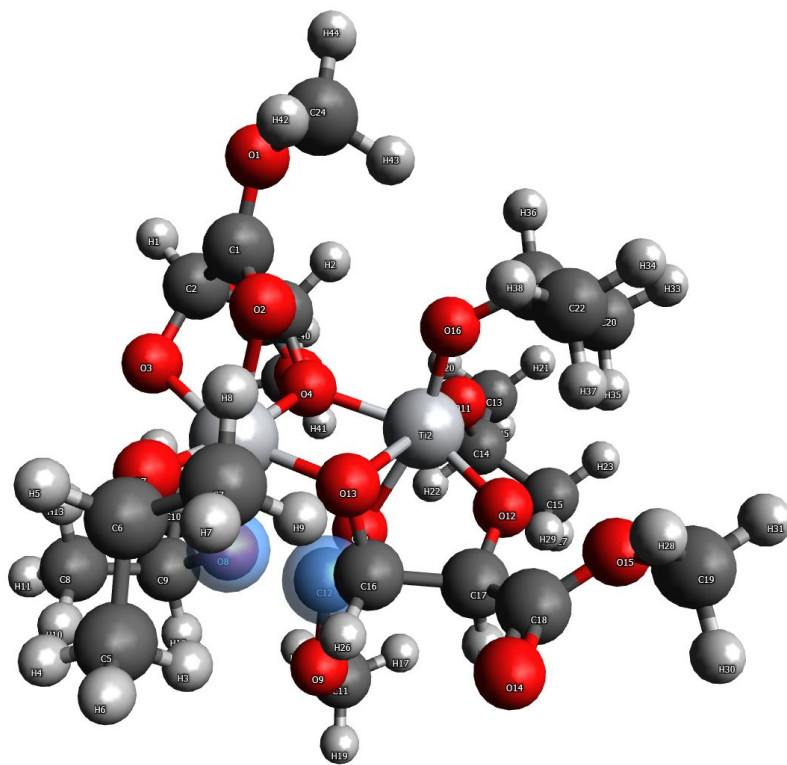
Based on the TargetMol we defined a set of criteria that had to be satisfied to be validated as a correct structure, e.g., certain interatomic bonds that had to be retained.

The general validation criteria were that all Ti-O, C-O, C-C and C-H bonds should stay intact, and neither removal nor creation of any such bonds should occur. The atomic neighbour lists and corresponding interatomic bonds were determined using covalent radii as implemented in ASE software [35], which are based on [36].

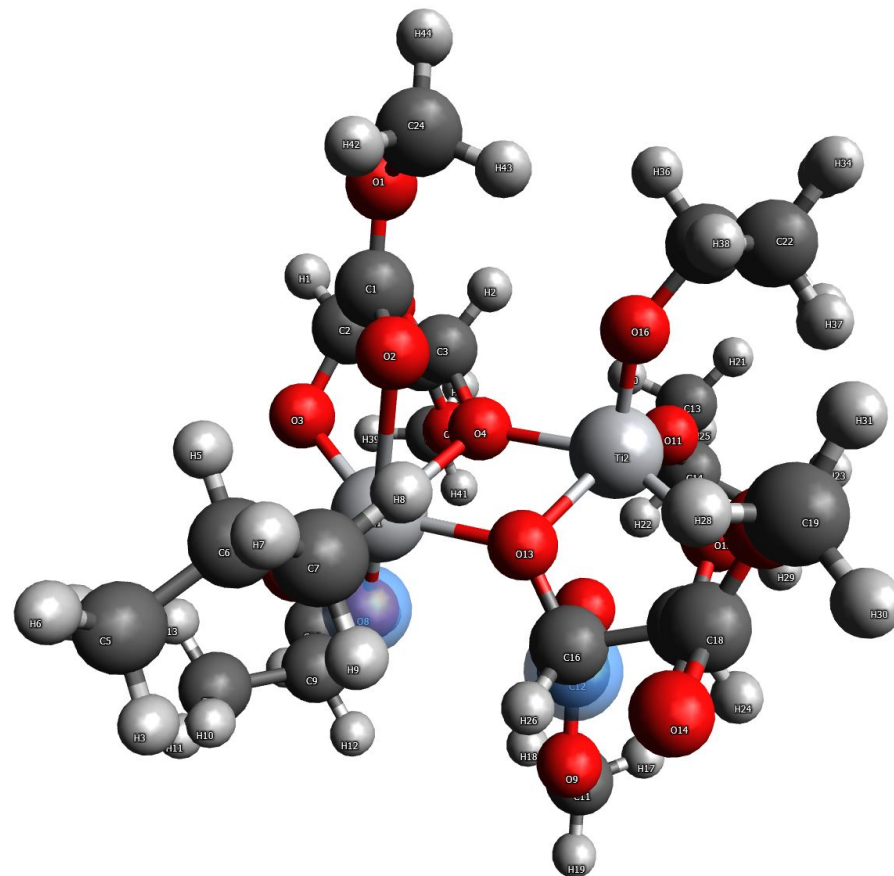
DFT parandab geomeetria

- CREST (GFN-xTB) liigutab aatomid O8 ja C12 teineteisele liiga lähedale.
- DFT "parandab": liigutab ligandid üksteisest kaugemale.
- Tsentraalne Ti-O-Ti-O romb läheb DFT-ga sümmeetrilisemaks.

CREST-i väljund



DFT optimeeritud



Retsensent dr. Sven Nõmm, kriitika 4.2.

- In the first part of the thesis, the author did not present any validation framework and numeric thresholds to compare the results with. This makes it difficult to follow Subsections 5.6 and 5.8.

Vastus (vt. slaid lk.13, Ti-O-C nurkade analüüs):

1. Mõte on hea ja oli meil ka arutlusel. Probleem selles, et meil pole objektiivset tõde, milleks oleks eksperimentaalne katse. Meil on parimaks võrdluseks kvantkeemilised arvutused valitud täpsusastmel. Nagu ka tekstis on mainitud, siis see on teada probleem, et poolempiiriline meetod GFN2-xTB alahindab Metall-Hapnik-Süsinik nurki. Paragrahv 5.8. on kirjeldav statistika, kus me näitame sama fenomeni ilmnemist, nii tabeli kui ka histogrammi kujul.

Retsensent dr. Sven Nõmm, kriitika 5

- The methods used by the author relay on the density functional theory; again, no examples of the functionals have been provided. Overall, the presentation is very poor with respect to technical details.

Vastus:

1. Tihedusfunktsionaalid on antud töö kontekstis nn “musta kasti” meetodid, me viitame neile nimedega ja vastava viitega referentside all. See on tavapraktika arvutuskeemia teadusartiklites. Tihedusfunktsionaale ei kirjeldata täpsemalt, kui nimega ja viitega vastavale publikatsioonile.
2. Põhjus on ilmselt selles, et kui vaadata vastavaid originaalpublikatsioone, siis pole eriti triviaalne sealt kõik asjassepuutuv info üle kopeerida. Soovitan vaadata tihedusfunktsionaali PBE0 originaalpublikatsiooni, ref.[23].

Retsensent dr. Sven Nõmm, kriitika 5: tihedusfunktsionaali näide

- The methods used by the author relay on the density functional theory; again, no examples of the functionals have been provided. Overall, the presentation is very poor with respect to technical details.

Vastus:

- Näide tihedusfunktsionaalist GFN meetodite loomisel (Bannwarth et. al, Extended tight-binding quantum chemistry methods, DOI: 10.1002/wcms.1493)

$$E_{\text{tot}} = E_{nn} + \sum_i^{N_{\text{MO}}} n_i \int \psi_i^*(\mathbf{r}) \left[\hat{T}(\mathbf{r}) + V_n(\mathbf{r}) + \varepsilon_{\text{XC}}^{\text{LDA}}[\rho(\mathbf{r})] + \frac{1}{2} \int \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} + \Phi_{\text{C}}^{\text{NL}}(\mathbf{r}, \mathbf{r}') \right) \rho(\mathbf{r}') d\mathbf{r}' \right] \psi_i(\mathbf{r}) d\mathbf{r}.$$

$$\text{kus} \quad \rho(\mathbf{r}) = \sum_i^{N_{\text{MO}}} n_i \int \psi_i^*(\mathbf{r}) \psi_i(\mathbf{r}) d\mathbf{r}.$$

Retsensent dr. Sven Nõmm, kriitika 6

- Throughout the thesis, the author cites many existing software solutions. Without proper flow chart style diagrams, it is difficult to follow their organization and proportion in the software developed by the author. The reviewer suggests including such a diagram in the defence presentation slides. In addition, the diagram explaining the general workflow of the experiments would be a welcomed addition.

Vastus:

1. Molli - praegusel kujul on see pigem abifunktsioonide kogum, mida autor kasutas magistritöö teostamiseks. Programmeeritud nii, et oleks ka kasutatav väljaspool antud magistritööd. Kuigi autor viitab sellele tekstis "software", siis praeguses teostuses ei vasta see tüüpilisele "valmis" tarkvarale.

Molli näide 1: Gaussiani programmi log failide parsimine


```
from gaussian_utils import process_many_log_files

gaussian_log_files = [
    Path("C:/tmp/molli_examples/gaussian_logs_1/ex0a_c7_gfn2_pbe1pbe_def2svpvp_svpfit.log"),
    Path("C:/tmp/molli_examples/gaussian_logs_1/ex0a_c24_gfn2_pbe1pbe_def2svpvp_svpfit.log"),
    Path("C:/tmp/molli_examples/gaussian_logs_1/ex0ff_c1_gfnff_pbe1pbe_def2svpvp_svpfit.log"),
]

output_dir = Path("C:/tmp/molli_examples/gaussian_logs_1/molli_output")

aggregate_xyz_file = output_dir.joinpath("conformers.xyz")

process_many_log_files(
    input_paths=gaussian_log_files,
    output_dir=output_dir,
    aggregate_log_file_name="aggregate_log.txt",
    write_last_opt_steps_file_path=aggregate_xyz_file,
)
```



- aggregate_log.txt
- conformers.xyz
- ex0a_c7_gfn2_pbe1pbe_def2svpvp_svpfit.xyz
- ex0a_c7_gfn2_pbe1pbe_def2svpvp_svpfit_last_step.xyz
- ex0a_c7_gfn2_pbe1pbe_def2svpvp_svpfit_opt_steps.xyz
- ex0a_c7_gfn2_pbe1pbe_def2svpvp_svpfit_scf_summary.txt
- ex0a_c24_gfn2_pbe1pbe_def2svpvp_svpfit.xyz
- ex0a_c24_gfn2_pbe1pbe_def2svpvp_svpfit_last_step.xyz
- ex0a_c24_gfn2_pbe1pbe_def2svpvp_svpfit_opt_steps.xyz
- ex0a_c24_gfn2_pbe1pbe_def2svpvp_svpfit_scf_summary.txt
- ex0ff_c1_gfnff_pbe1pbe_def2svpvp_svpfit.xyz
- ex0ff_c1_gfnff_pbe1pbe_def2svpvp_svpfit_last_step.xyz
- ex0ff_c1_gfnff_pbe1pbe_def2svpvp_svpfit_opt_steps.xyz
- ex0ff_c1_gfnff_pbe1pbe_def2svpvp_svpfit_scf_summary.txt

Molli näide 1 jätk: aggregate_log.txt

```
aggregate_log.txt
1 {
2   "summary": {
3     "num_experiments_total": 3,
4     "num_experiments_successful": 3,
5     "num_experiments_failed": 0,
6     "energy_diff_best_worst": "-0.003855 a.u., -2.42 kcal/mol, -10.12 kJ/mol",
7     "ranking": [
8       "1: diff best: 0.0 a.u., 0.0 kcal/mol, 0.0 kJ/mol, source: ex0a_c7_gfn2_pbelpbe_def2svpvp_svpfit.log",
9       "2: diff best: 0.00054 a.u., 0.34 kcal/mol, 1.42 kJ/mol, source: ex0a_c24_gfn2_pbelpbe_def2svpvp_svpfit.log",
10      "3: diff best: 0.003855 a.u., 2.42 kcal/mol, 10.12 kJ/mol, source: ex0ff_c1_gfnff_pbelpbe_def2svpvp_svpfit.log"
11    ],
12    "last_opt_steps_file": "C:\\tmp\\molli_examples\\gaussian_logs_1\\molli_output\\conformers.xyz"
13  },
14  "experiments": [
15    {
16      "input_path": "C:\\tmp\\molli_examples\\gaussian_logs_1\\ex0a_c7_gfn2_pbelpbe_def2svpvp_svpfit.log",
17      "output_dir": "C:\\tmp\\molli_examples\\gaussian_logs_1\\molli_output",
18      "results": {
19        "final_xyz": "C:\\tmp\\molli_examples\\gaussian_logs_1\\molli_output\\ex0a_c7_gfn2_pbelpbe_def2svpvp_svpfit.xyz",
20        "opt_steps": "C:\\tmp\\molli_examples\\gaussian_logs_1\\molli_output\\ex0a_c7_gfn2_pbelpbe_def2svpvp_svpfit_opt_steps.xyz",
21        "last_opt_step": "C:\\tmp\\molli_examples\\gaussian_logs_1\\molli_output\\ex0a_c7_gfn2_pbelpbe_def2svpvp_svpfit_last_step.xyz",
22        "scf_summary": {
23          "scf_summary_file": "ex0a_c7_gfn2_pbelpbe_def2svpvp_svpfit_scf_summary.txt",
24          "gaussian_version": "Gaussian 16, Revision C.02",
25          "gaussian_command": "#T PBE1PBE/Def2SVPP/SVPfit opt=(calcfc) formcheck",
26          "dft_functional": "PBE1PBE/Def2SVPP/SVPfit",
27          "dft_info": "696 basis functions, 1318 primitive gaussians, 740 cartesian basis functions, 180 alpha electrons 180 beta electrons",
28          "chemical_formula": "C24H44O16Ti2",
29          "optimization_converged": true,
30          "elapsed_time_str": "0d 3h 27m 14.6s",
31          "elapsed_time_minutes": 207.2,
32          "num_steps": 57,
33          "minutes_per_step": 3.6,
34          "energy_start": -3839.28103693,
35          "energy_end": -3839.37826483,
36          "energy_delta": -0.0972278999977977,
37          "energy_delta_text": "-0.0972279 a.u., -61.01 kcal/mol, -255.27 kJ/mol",
38          "job_cpu_time": "8d 18h 34m 53.3s",
39          "job_cpu_hours": 210.6,
40          "job_cpu_hours_per_step": 3.69,
41          "job_completion_datetime": "11-Dec-2022 00:11:29",
42          "energy_diff_to_best": "0.0 a.u., 0.0 kcal/mol, 0.0 kJ/mol"
43        }
44      }
45    ],
46  }
```


Molli näide 2: features analysis

Load conformers generated by CREST GFN2-xTB from xyz-files.

```
mols_ex0_gfn2 = au.create_ase_atoms_list_from_xyz_files(
    input_paths=[
        Path("C:/tmp/gaussian/workflow/data/conformers/crest_gfn2/ex0a_gfn2_crestconfs.xyz"),
        Path("C:/tmp/gaussian/workflow/data/conformers/crest_gfn2/ex0b_gfn2_crestconfs.xyz"),
    ]
)

len(mols_ex0_gfn2)
```

68

Create features list that we are investigating.

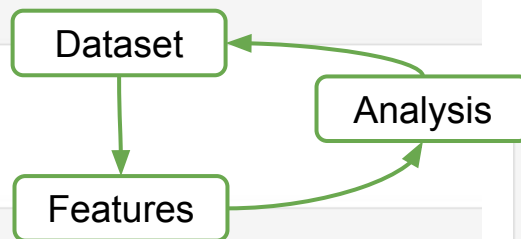
```
angles_TiOC = ft.create_angles(
    atom_idxs=tm.angles_TiOC,
    mol=mol_original
)

angles_TiOTi = ft.create_angles(
    atom_idxs=tm.angles_TiOTi,
    mol=mol_original
)

dihedrals_OCCO = ft.create_dihedrals(
    atom_idxs=tm.dihedrals_OCCO,
    mol=mol_original
)

all_features = angles_TiOC + angles_TiOTi + dihedrals_OCCO
len(all_features)
```

atom_idxs:
aatomite järjekorra numbrite kolmikud:
[(7,6,4), (7,20,18), ...]
mol (optional): referents molekul, et
informatiivsem kuvamine



1. Laeme geomeetriad xyz-faigest
2. Defineerime otsitavad omadused
3. Arvutame otsitavad väärtused

```
df_ex0_gfn2 = A.analyze_by_features_to_dataframe(
    molecules=mols_ex0_gfn2,
    features_list=all_features
)

df_ex0_gfn2[:3]
```

	mol_names	Angle Ti(7)-O(6)- C(4)	Angle Ti(7)- O(20)- C(18)	Angle Ti(7)- O(31)- C(29)	Angle Ti(7)- O(60)- C(53)
0	ex0a_gfn2_crestconfs_1	108.408354	131.202147	129.067155	119.084795
1	ex0a_gfn2_crestconfs_2	108.200779	131.049456	130.122016	119.181737
2	ex0a_gfn2_crestconfs_3	108.472254	130.337212	128.600648	119.146988

19

Retsensent dr. Mario Öeren, kriitika 2

- Konformeeride nimetused (e.g. ex19_c23) on lugejale keerulised jälgida ja süstemaaline nimi ei oma põhitekstis lisaväärtust.

Vastus:

1. Konformeeride nimedel on semantika, võimaldab üheselt jälgida konformeeeri elutsüklit: kuidas/kust konformeer tekkis ja kuhu välja jõudis. Lisab olulist informatsiooni konformeeride nimekirjale.

Retsensent dr. Mario Öeren, kriitika 4

- Sagedusarvutuse oleks võinud kõige madalamale konformeerile siiski ära teha.

Vastus:

1. Jah, võibolla. Tehniliselt on lihtne teostada, pole veendunud, et see uut olulist informatsiooni annaks. Pigem prof. Tamm-e otsustada mida ja kuidas edasi analüüsida.

Tegemist on väga spetsiifilise arvutuskeemia küsimusega. Kui teha sagedusarvutust, siis kõigile (vähemalt valimile), saaksime statistilist informatsiooni konformeeride kohta.

Siin oleks väga kasulik aru saada küsimuse (küsiija) motivatsioonist, MIKS sellist arvutust teha.

Retsensent dr. Mario Öeren, küsimus 1

- Kas te kaalusite ka titaan-tartraat kompleksi spetsiifilise jõuvälja treenimist (jõuväli, mis on mõeldud ainult ühele molekulile)? Sellisel juhul oleks konformeeride süstemaalne läbi arvutamine põhimõtteliselt võimalik. Kas te oskate öelda, mis selle ideega valesti on?

Vastus:

1. Jah, igasuguseid ideid oli
2. Treenimine eeldab treening ja test valimite olemasolu. Magistritöö tulemusena tekitatud konformeerid võiksid olla treening ja test valimiks, kui keegi sooviks uut spetsiifilist mudelit treenida.
3. Küsimus on väga mitmetahuline ja minu jaoks moodustab jää-mäe tipu, kuna küsimust võib tõlgendada ja edasi arendada väga paljudes erinevates suundades. Tegelikuses jääb selgusetuks, mida küsija TÄPSELT silmas peab.

Retsensent dr. Mario Öeren, küsimus 2

- Kas te oskate oletada, kui palju aega võiks käesoleva töövooga kokku hoida (näitena võib kasutada Sharpless'i katalüsaatorit)?

Vastus:

1. Magistritöös näitasime kahte eksperimenti, millega oli ajavõit 2 korda. Mudelkeemia PBE0/cc-PVTZ kiirendamine mudelkeemiaga PBE0/def2-SV(P), kasutates 10 sammu eeloptimeerimist.
2. Samade mudelkeemiate korral (hinnang sooritatud arvutuste baasil):
 - u. 4 korda ajavõitu, kui 20 sammu eeloptimeerida
 - u. 8 korda ajavõitu, kui 40 sammu eeloptimeerida
 - u. 16 korda ajavõitu, kui 80 sammu eeloptimeerida.
3. TargetMol puhul oleks minu ettepanek järgmine:
 - Täpse mudelkeemiaga optimeerida PBE0/def2-SV(P) täisoptimeeritud geomeetriaid (hinnang < 5..10 sammu)
 - Täpse mudelkeemiaga arvutada energia PBE0/def2-SV(P) täisoptimeeritud geomeetriaatest (tavapraktika)

Retsensent dr. Mario Öeren, küsimus 3

- Millisel hetkel võib teie töövoogu kasutav teadlane olla enesekindel, et rohkem konformeere läbi ei pea sõeluma ja leitud madalaim konformeer on tõenäoliselt „see õige“?

Vastus:

1. Teoreetiliselt ei saa kunagi olla kindel, et oleme leidnud madalaima konformeeri, sest praktikas me saa kõiki konformeere läbi sõeluda.
2. Tööprotsessis võib teha loomingulise pausi, kui CREST väljastab juba nähtud geomeetriaatega identseid (või väga sarnased) geomeetriaid.

Retsensent dr. Mario Öeren, küsimus 4

- Kaalusite alguses Psi4 kasutamist, kuid tehnilistel põhjustel jätkasite Gaussian'iga; miks te esialgu eelistasite Psi4?

Vastus:

1. Küsimus polnud Psi4 vs Gaussian. Me alustasime mitmete kvantkeemia vabavarade läbi katsetamisega. Enne Psi4 katsetasime ka BigDFT-d, millega ei jõudnud nii kaugele kui Psi4-ga ja seetõttu ei maininud ka töös.

Praktiline põhjus: magistritöö oli vaja valmis saada, see eeldas, et saaks arvutusi teha.

2. Konkreetselt miks Gaussian:

HPC-s installeeritud

GPU-tugi, lootsime katsetada ka GPU paralleelsust. Paraku andsid GPU, vaid u 10-20% ajalist võitu.