TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Keemia instituut

Karl Marti Toots

GAAS-IOONNE VEDELIK JAOTUSKOEFITSIENDI MODELLEERIMINE

Magistritöö (30 EAP)

Juhendajad:
Uko Maran, PhD
Sulev Sild, PhD
Jaan Leis, PhD

Tartu 2020

INFOLEHT

Gaas-ioonne vedelik jaotuskoefitsiendi QSPR modelleerimine

Magistritöös koostati kvantitatiivsed struktuur-omadus sõltuvused (QSPR) ioonsete vedelike $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$ gaas-ioonne vedelik jaotuskoefitsiendi prognoosimiseks. Mudelite loomisel kasutati multilineaarset regressiooni ja juhumetsa (*Random Forest*) regressiooni. Nimetatud ioonsete vedelike eksperimentaalsed jaotuskoefitsiendid erinevate orgaaniliste ühendite suhtes jaotati ristvalideeritud hinnangute arvutamiseks viieks hulgaks. Nende orgaaniliste ühendite jaoks arvutati molekulaartunnused ja valiti tunnused ortogonaalse sobitusalgoritmi (OMP) ning ettesuunatud valiku (*Forward Selection*) abil. Juhumetsa mudelite koostamise algoritmi mõjutavaid parameetreid optimeeriti. Parimate ristvalideeritud mudelite võrdlus näitas, et mittelineaarsed juhumetsa mudelid suudavad gaas-ioonne vedelik jaotuskoefitsiente prognoosida täpsemini, kui multilineaarse regressiooni mudelid. Samas ei olnud multilineaarse regressiooni lihtsamad mudelid oluliselt väiksema täpsusega. Optimaalsetesse mudelitesse valitud molekulaartunnuste analüüsi põhjal ilmnes, et gaas-ioonne vedelik jaotuskoefitsient on antud andmeseeria piires olulisel määral mõjutatud struktuursetest omadustest, nagu aatomite arv, aromaatse tuuma olemasolu, kindla funktsionaalrühma leidumine, eriti OH-rühm, polaarsus ja süsinikust raskemate aatomite esinemine molekulis. Välja töötatud QSPR mudelite testiks kasutati ka välist valideerimist, mille tulemused tõendavad nende mudelite ennustusvõimekust.

Märksõnad: QSPR, gaas-ioonne vedelik jaotuskoefitsient, jaotuskoefitsient, juhumets, multilineaarne regressioon, molekulaartunnuste valik, masinõpe

CERCS kood ja nimetus: P410 Teoreetiline ja kvantkeemia, T150 Materjalitehnoloogia, P400 Füüsikaline keemia

QSPR modelling of gas-to-ionic liquid partition coefficients

In the present master thesis, quantitative structure-property relationships (QSPR) were developed to predict the gas-to-ionic liquid partition coefficient for ionic liquids $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ *and* $[MeoeMPyrr]^+[FAP]^-$. *Multilinear regression and random forest regression were applied to create the models. The experimental partition coefficients of the ionic liquids to the various organic compounds were divided into five sets for cross-validated estimates. Molecular descriptors were calculated for these organic compounds and feature selection algorithms based on the orthogonal matching pursuit (OMP) and Forward Selection were implemented. The hyperparameters of the random forest models were tuned. The comparison of the best models according to the cross-validated coefficient of determination showed that the nonlinear random forest model predicts gas-to-ionic liquid partition coefficients more accurately than the multilinear regression model. However, the simpler models obtained with multilinear regression were not significantly inferior. Based on the selected descriptors in the optimal models, it was found that the gas-to-ionic liquid partition coefficient is significantly influenced by the properties of a molecule encoded in the molecular descriptors: number of atoms, presence of an aromatic ring, presence of a specific functional group, especially the hydroxyl group, polarity and presence of atoms heavier than carbon. The predictive ability of the developed QSPR models was confirmed by external validation.*

Keywords: *QSPR, gas-to-ionic liquid partition coefficient, partition coefficient, random forest, multiple linear regression, molecular descriptor selection, machine learning*

CERCS codes and names: *P410 Theoretical and quantum chemistry, T150 Material technology, P400 Physical chemistry*

SISUKORD

KASUTATUD LÜHENDID

QSPR	kvantitatiivne struktuur-omadus sõltuvus (ingl. k. *quantitative structure-property relationship*)
DNA	desoksüribonukleiinhape
RNA	ribonukleiinhape
$\log K_{giv}$, K_{giv}	gaas-ioonne vedelik jaotuskoefitsient
GLC	gaas-vedelik kromatograafia (ingl. k. *gas-liquid chromatography*)
LSER	lineaarne lahustuvusenergia seos (ingl. k. *linear solvation energy relationship*)
CV	ristvalideerimine (ingl. k. *cross-validation*)
CCC	korrelatsiooni ühilduvuskordaja (ingl. k. *Concordance Correlation Coefficient*)
OMP	ortogonaalne sobitusalgoritm (ingl. k. *orthogonal matching pursuit*)
E-olek	elektrotopoloogiline olek (ingl. k. *electrotopological state* ehk *E-state*)

SISSEJUHATUS

Ioonsete vedelike unikaalsed omadused nagu ülimadal aururõhk, kõrge polaarsus ja termiline stabiilsus on taganud nende laiaulatusliku uurimise lahustitena sünteesis [1, 2], katalüüsis [1, 3], elektrokeemilistes rakendustes [2, 4-6], ainete eraldamiseks [7-9] ja veel paljudes rakendustes [5, 10-13]. Ioonse vedeliku oluline omadus iseloomustamaks keemilise aine jaotumist ioonse vedeliku ja ümbritseva keskkonna vahel on gaas-ioonne vedelik jaotuskoefitsient. [14] Pidevalt sünteesitakse uusi ioonseid vedelikke püüdes saavutada vastavaks rakenduseks optimaalse jaotuskoefitsiendiga keskkonda. [15-17] Seega jaotuskoefitsiendi modelleerimine ja arvutusmudeli loomine on vajalik, et valida kiirelt, kulutult ja võimalikult täpse omadusega ioonne vedelik.

Gaas-ioonne vedelik jaotuskoefitsiendi määramisel on olnud kasutusel eksperimentaalsed meetodid, mis enamasti pole sobilikud suure hulga ühendite läbi sõelumiseks. [18] Seejuures võimalikke katiooni-aniooni kombinatsioone ioonse vedeliku valmistamiseks on määratu suur arv. Tänapäeva arvutite järjest kasvav arvutuslik jõudlus on tekitanud võimaluse selle rakendamiseks keemiliste ja füüsikaliste suuruste modelleerimisel. Gaas-ioonne vedelik jaotuskoefitsiendi arvutite abil modelleerimisel on leidnud laialdast kasutust kvantitatiivsed omaduse ja struktuuri vahelised seosed (ingl k QSPR ehk *quantitative structure-property relationship*) ehk matemaatilised mudelid, mille abil seostatakse ühendi struktuurist tulenevaid omadusi selle ühendi eksperimendist mõõdetava omadusega. [19, 20] Varasemates jaotuskoefitsiendi QSPR uurimustes on rakendatud mudelite koostamiseks näiteks multilineaarset regressiooni, tugivektori regressiooni ja närvivõrke. [18, 21] Juhumetsa regressioon on modelleerimisalgoritm mittelineaarsete mudelite loomiseks ning eelnevad tööd on näidanud algoritmi võimet konstrueerida suurepärase täpsusega mudeleid. [22-24] Jaotuskoefitsiendi QSPR mudelite koostamisel juhumetsa regressiooni ja multilineaarse regressiooni abil võimaldab saadud mudeleid ka analüüsida ja võrrelda. Analüüsil võib leida seaduspärasusi keemilise ja füüsikalise sisu osas, kuidas koefitsiendi väärtus kujuneb olenevalt keemilisest ühendist.

Seetõttu keskendutakse magistritöös orgaaniliste ühendite gaas-ioonne vedelik jaotuskoefitsiendi modelleerimisele QSPR meetodiga, kasutades juhumetsa ja multilineaarse regressiooni lähenemisi. Mudeleid vaadeldakse kolme ioonse vedeliku korral: $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$.

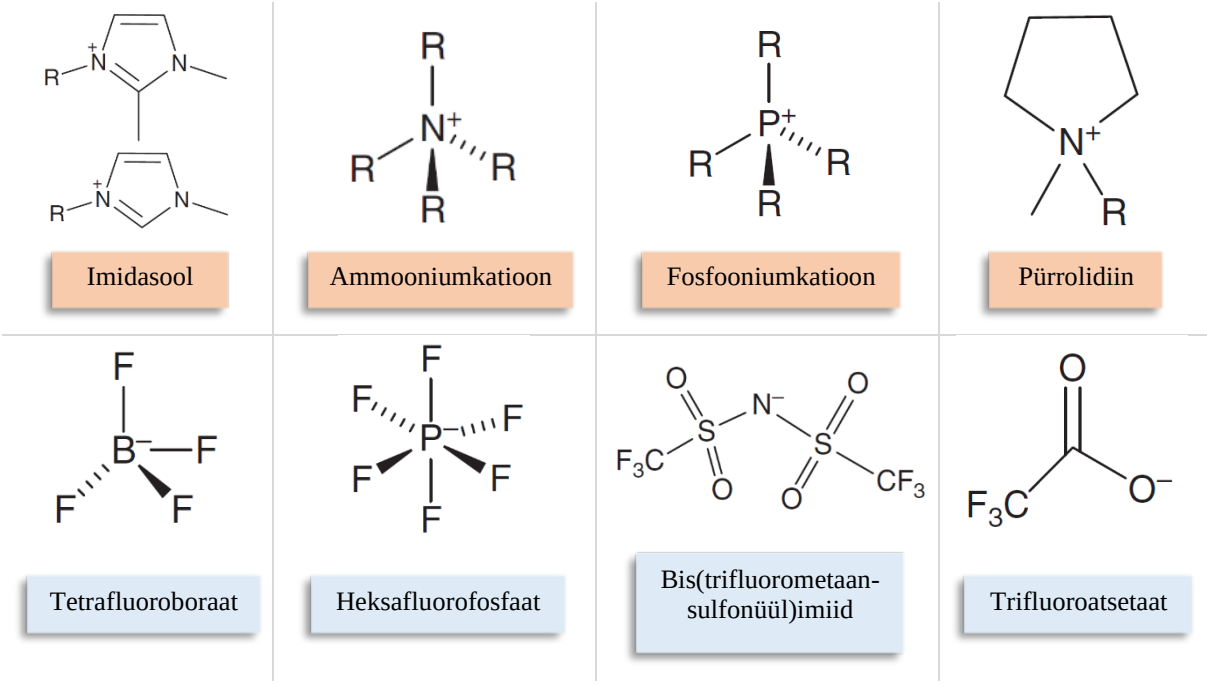
1. KIRJANDUSE ÜLEVAADE

1.1 Ioonne vedelik

Ühe levinud definitsiooni järgi on ioonne vedelik täielikult ioonidest koosnev vedelik. [25] Kuigi selle definitsiooni alla lähevad ka sulatatud soolad – kristalsed soolad vedelas faasis, näiteks sula NaCl, siis enamasti kasutatakse mõistet ioonne vedelik madala sulamistemperatuuriga orgaanilise katiooniga soolade korral. [25] Eristamaks ioonseid vedelikke lisab osa kirjandust definitsioonile piirangu, et aine peab olema vedelas faasis alla 100^{o}C juures. [25-27] Teisalt leidub hulgaliselt rakendusi üle 100^{o}C-ses keskkonnas täielikult ioonidest koosnevatele vedelikele, nagu näiteks mangaanoksiidkile elektrokeemiline sadestus ioonsest vedelikust (etüülammooniumnitraadist) vee oksüdeerimisreaktsiooni katalüüsiks. [28] Lisaks leidub ka selliseid aineid, mille puhul sulamistemperatuuri leidmine on liialt keeruline, mistõttu nende liigitamine sulamistemperatuuri järgi pole võimalik. [25] Ioonsete vedelike allajahutamisel läheb vedel sool üle kristalsesse faasi, kuid mõni allajahutatud vedelik jääb piisavalt viskoosseks, mistõttu kristalne faas ei teki praktilises ajaskaalas ja moodustub klaasjas tahkis. Enamikes ioonsetes vedelikes leidub ka vähesel määral neutraalseid komponente ja muid sünteesi protsessist pärinevaid lisandeid. Olenevalt rakendusest on kasutusel erineva puhtusastmega ioonsed vedelikud (näiteks 95% või 99% puhtusastmega). Käesolevas töös mõistetakse ioonse vedeliku all praktiliselt täielikult ioonidest koosnevat vedelikku.

Ioonsete vedelike liigitamine, süntees ja puhtus

Ioonsed vedelikud koosnevad enamasti kogukatest orgaanilistest katioonidest ja orgaanilistest või anorgaanilistest anioonidest (vt **Joonis 1**). [27, 29] Neid võib liigitada protoonseteks, kiraalseteks, magnetilisteks, polümeerseteks, kelaatseteks, fluoritud, oksüdeerivateks ja diprotoonseteks. [27, 30] Ioonse vedeliku sünteesistrateegiast sõltub olulisel määral nende puhtus. Ioonsete vedelike sünteesi puhul on oluline ka väike mõju keskkonnale, nagu näiteks mitmeastmeliste mikrolaine- ja ultraheli-ergastusel põhinevate sünteesimeetodite korral. Teaduslikes uurimustes usaldusväärsete tulemuste saamiseks vajatakse kõrge puhtusastmega ioonseid vedelikke. Olenevalt ioonsest vedelikust võib nende puhastamine olla keerukas. Ioonsetele vedelikele omane ülimadal aururõhk takistab nende ümberdestilleerimist, kuid enamus ioonsete vedelike lisandeid on siiski kõrvaldatavad destillatsiooni abil. Puhtusastet on võimalik suurendada ka ioonse vedeliku aeglasel kristallisatsioonil. [27]



Joonis 1. Sagedamini kasutatavates ioonsetes vedelikes leiduvate katioonide (üleval) ja anioonide (all) struktuurid ja nimetused.

Ioonsete vedelike olulisus ja kasutusala

Üha laialdasemalt kasutatakse ioonseid vedelikke alternatiivina lenduvatele orgaanilistele lahustitele, mineraalhapetele, alustele, tahketele hapetele ja veel paljudele ühenditele. [27] Peamiseks põhjuseks on siin ioonsete vedelike kaduvväike aururõhk. Potentsiaalne keskkonnasäästlik kasu on siinjuures oluline edasiviiv jõud ioonsete vedelike uurimisel ja rakendamisel. [25, 26, 31] Lisaks on ioonsete vedelike tähelepanuväärne omadus nende kohandatavus. Ioonse vedeliku valmistamisel tuleb kõrge sulamistemperatuuri vähendamiseks kasutada suuri hajutatud laenguga ioone. [25] Seega kasutada on erakordselt suur valik ioone. Teades, kuidas ioonid ioonsete vedelike omadusi mõjutavad, saab ioonseid vedelikke peenhäälestada vastavalt rakendusele. [25, 27] Soovides luua ioonset vedelikku vastavalt rakendusele, osutuvad kaasaja arvutuslikud modelleerimise tehnikad tõenäoliselt väga kasulikuks. [25]

Ioonsete vedelike kasutamisel on põhirõhk suunatud füüsikalise keemia, keemiatehnika, materjaliteaduse ja ka mitmeid keemia alamdistsipliine ühendavatele valdkondadele. [27] Ioonsed vedelikud on head lahustid nii orgaanilistele kui ka anorgaanilistele ainetele, mistõttu saab neid kasutada näiteks erinevate reagentide viimiseks samasse faasi. [32, 33] Nende häid lahustamisomadusi kasutatakse ära ka erinevates sünteesides, materjalide töötlemisel, biomassi töötlemisel, ekstraheerimisel ja gaaside eraldamisel. Võime lahustada pea lahustamatuid aineid ja vähene keskkonnamõju on

peamiseks põhjuseks ioonsete vedelike kasutamisel orgaanilises, anorgaanilises ja bioloogilises sünteesis. [25] Ioonsete vedelike keskkonnamõju on pälvinud tähelepanu ka seetõttu, et väga vähesed neist on mürgised või mitte biolagunevad. [29, 31] Elektrokeemiline stabiilsus lisaks unikaalsetele lahusti omadustele loob hea aluse ioonsete vedelike kasutamisele haruldaste muldmetallide soolade eraldamiseks ja puhta muldmetalli elektrokeemiliseks sadestamiseks. Madal sulamistemperatuur võimaldab ioonseid vedelikke kasutada ravimites, kus vedeliku katioon või anioon on ravivaks toimeaineks. Vedelal kujul ravim võib hõlbustada ka selle manustamist. Ioonsete vedelike võime lahustada ja stabiliseerida ensüüme, valke, DNA-d ja RNA-d on erakordselt väärtuslik omadus biotehnoloogilistes rakendustes. [25]

1.2 Jaotuskoefitsient

Keemilise ühendi lahustumisel toimub molekulide jagunemine kahes omavahel segunematus lahustis. [14, 34] Seda nähtust iseloomustab kvantitatiivselt jaotuskoefitsient ehk ühendi kontsentratsioonide suhe segunematutes lahustites tasakaaluolekus (1). [14, 34] See suhe on ühendi lahustuvuse erinevuse mõõt nende kahe lahusti vahel. Näiteks keskkonnakeemias tavaliselt on üks lahustitest vesi ja teine hüdrofoobne aine, milleks tihti valitakse 1-oktanool. [34] Sellises segus annab keemilise ühendi jaotuskoefitsient mõõdu aine hüdrofiilsusele või hüdrofoobsusele. [34, 35] Lahustiteks võivad olla nii orgaanilised kui anorgaanilised ühendid erinevates olekutes, näiteks õhk, vesi, pinnas, setted ja aerosoolid. Usaldusväärsete jaotuskoefitsientide kättesaadavus on teaduslikel ja regulatiivsetel eesmärkidel hädavajalik. Üldine eesmärk on mõista ja ennustada keemiliste ainete jaotust paljudest erinevatest ainetest koosnevas keskkonnas. Rahvusvaheline Puhta ja Rakenduskeemia Liit (IUPAC) on soovitanud kasutada jaotuskoefitsiendi asemel terminit jaotuskonstant (ingl. k. *partition constant*) või jaotussuhe (ingl. k. *partition ratio*) ning on defineerinud jaotuskonstandi järgmiselt: [34]

$$\mathrm{K}_A = \frac{c_{A,lahusti_1}}{c_{A,lahusti_2}} \quad (1)$$

Paljud autorid eelistavad siiski terminit jaotuskoefitsient (ingl. k. *partition coefficient* või *distribution coefficient*). Jaotuskoefitsient on kasulik mis tahes süsteemi kemikaalide jaotuse hindamisel. Looduses jaotub kõrge oktanool-vesi jaotuskoefitsiendiga hüdrofoobne keemiline aine peamiselt ökosüsteemi hüdrofoobsetele aladele. Seevastu hüdrofiilset kemikaali, mille oktanool-vesi jaotuskoefitsient on madal, leidub peamiselt märgades piirkondades, näiteks sood, tiigid, jõed ja järved. Jaotuskoefitsienti kasutatakse ka ravimite väljatöötamisel lahustunud aine hüdrofoobsuse mõõtmiseks ja selle alusel ravimi membraanist läbilaskvuse

määramiseks. Kui aga üks lahustitest on gaas ja teine vedelik, iseloomustab lahustuva aine jagunemist nende faaside vahel gaas-vedelik jaotuskoefitsient. [34]

Gaas-ioonne vedelik jaotuskoefitsient

Keemilise ühendi jagunemist gaasi ja ioonse vedeliku vahel kirjeldab gaas-ioonne vedelik jaotuskoefitsient, K_{giv}: [14, 18]

$$K_{giv} = \frac{c_{iv}}{c_g}, \tag{2}$$

kus c_g on ühendi kontsentratsioon gaasis ja c_{iv} on ühendi kontsentratsioon ioonses vedelikus tihti esitatakse seda ka logaritmilisel kujul $\log K_{giv}$. Koefitsienti K_{giv} saab leida arvutuste teel isotermilise gaas - vedelik kromatograafia (ingl. k. *gas-liquid chromatography* ehk *GLC*) mõõtmise tulemustest. [36, 37] Jaotuskoefitsient leitakse lahustunud aine elueerimiseks vajaliku kandegaasi ruumala ja statsionaarse vedelikfaasi ruumala suhtena. [36] K_{giv} määramisega seotud eksperimentaalsed meetodid on töömahukad, kallid ja aeganõudvad ning vajavad piisavas koguses puhtaid ühendeid. Need meetodid enamasti ei ole sobivad suure hulga ühendite läbi sõelumiseks. Selle probleemi lahendamiseks on välja töötatud teoreetiline ja arvutuslik metoodika jaotuskoefitsientide hindamiseks. [18]

1.3 Omaduse ja struktuuri vaheline kvantitatiivne seos

Kvantitatiivsed omaduse ja struktuuri vahelised seosed on matemaatilised mudelid, mille abil seostatakse ühendi struktuurist tulenevaid omadusi selle ühendi bioloogilise või füüsikalis-keemilise aktiivsuse või muu omadusega. QSPR meetodid on *in silico* meetodite alamhulk ning neil on kirjeldav ja ennustav võime eeldusel, et struktuurilt sarnastel ühenditel on sarnased omadused. QSPR meetodit saab kasutada näiteks ühendite bioloogilise mõju ennustamiseks enne tegelikku bioloogilist testimist või huvipakkuvaid omadusi mõjutavate struktuuri iseärasuste analüüsimisel. [19, 20]

1.3.1 QSPR meetodi kirjeldus

QSPR mudelite arendamine koosneb mitmest järjestikusest etapist: andmete kogumine, molekulaartunnuste (ingl. k. *molecular descriptors*) arvutamine, andmete treening- ja testhulgaks jaotamine, treeninghulga põhjal mudeli treenimine ning testhulgal mudeli valideerimine. Mudelite väljatöötamine algab huvipakkuva omaduse kohta andmete kogumisega, näiteks kindlale molekulile või materjalile vastava keemilise või füüsikalise

omaduse mõõtmisega. Hoolikas andmete töötlus ja andmevalimi koostamine mõjutab suurel määral saadud mudeli kvaliteeti, mille tõttu esmalt välistatakse mudelit halvendavad madala kvaliteediga andmed. Seejärel transformeeritakse iga molekuli või materjali kohta kogutud info ümber tema struktuuri eripärasid kirjeldavateks tunnusteks, mida nimetatakse molekulaartunnusteks. Molekulaartunnuseid võib olla tohutult palju, kuid kõik neist pole konkreetse probleemi jaoks kasulikud. Seega tuleks enne modelleerimist eemaldada mitteinformatiivsed või ülearused tunnused. Seejärel jagatakse QSPR uuringus kasutatav täielik andmestik enne mudeli koostamist treeninghulgaks ja testhulgaks. QSPR mudeli koostamiseks kasutatakse erinevaid modelleerimismeetodeid, nagu näiteks lineaarset regressiooni, logistilist regressiooni või masinõppe meetodeid, et luua matemaatiline sõltuvus, mis kirjeldaks huvipakkuva omaduse ja struktuuri vahelist seost. Optimaalne mudel saadakse, valides samaaegselt parimad mudeli algoritmi kontrollivad parameetrid (ingl. k. *hyperparameters*) ja parim tunnuste komplekt mudelis. [20]

Mudeli hindamine

Treenitud mudelile antakse hinnang kasutades andmete testhulka, mis kinnitab mudeli ennustusvõimet ennenägemata andmetel. [20] Mudelite ennustusvõimet võrreldakse erinevate hindamisparameetrite kaudu, näiteks korrelatsioonikordaja R, määramiskoefitsient r^2 (ingl. k. *coefficient of determination*) ja ruutkeskmine viga MSE (vt **Joonis 2**). [20, 38, 39] Tihti kasutatakse mudeli hindamiseks spetsiaalset metoodikat nimega ristvalideerimine (ingl. k. *cross-validation* ehk *CV*). Ristvalideerimisel koostatakse *n* treening- ja testhulka, mudel treenitakse ja hinnatakse kõigil *n* hulgal ning mudeli hinnang arvutatakse kõigi testhulkade hinnangute keskmisena. Ristvalideerimise eesmärk on võtta parima mudeli valikul arvesse kogu andmehulk, kindlustada mudeli suutlikkus ennenägemata andmete põhjal prognoosida ning vältida olukorda, kus mudel juhuslikult vaid kindlal testhulgal häid tulemusi annab.

QSPR uurimustes on mudeli ennustusvõimekuse kinnitamisel saanud tavaks kasutada ka mudeli koostamise välist hindamisparameetrit, milleks on mitmesuguseid meetodeid. [40 - 42] Üks võimalik selline hindamisparameeter on korrelatsiooni ühilduvuskordaja (ingl. k. *Concordance Correlation Coefficient*) ehk CCC (vt **Joonis 2**). [42] Korrelatsiooni ühilduvuskordaja peamine eelis on selle lihtsus, sest mitmed teised parameetrid kasutavad nii treening- kui testhulki ning vajavad mõnel juhul isegi viie erineva tingimuse kontrolli. [38, 39, 42] Teiseks on CCC võimeline hindama nii ennustatud punktide kaugusi eksperimentaalsest regressioonijoonest kui ka mudeli ennustustest koostatud regressioonijoone tõusu hälbimist

eksperimentaalse regressioonijoone tõusust. Mudeli, mille CCC on >0,85 mudel loetakse ennustusvõimeliseks. [42]

Määramiskoefitsient

$$r^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (3)$$

Korrelatsioonikordaja

$$\mathrm{R} = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}} \quad (4)$$

Ruutkeskmine viga

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

Korrelatsiooni ühilduvuskordaja

$$CCC = \frac{2\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{\hat{y}} - \bar{y})^2} \quad (6)$$

Joonis 2. Hindamisparameetrid ja nende arvutuseeskiri. y_i ja $\hat{y}_i$ on vastavalt i-ndal vaatlusel eksperimentaalne omaduse väärtus ja mudeli ennustatud omaduse väärtus, $\bar{y}$ ja $\bar{\hat{y}}$ on vastavalt eksperimentaalsete omaduse väärtuste keskmine ja mudeli ennustatud omaduse väärtuste keskmine. [40, 42]

QSPR modelleerimismeetodid võib jagada regressiooni ülesanneteks ja klassifitseerimise ülesanneteks. Regressioonanalüüsis püütakse leida sõltuvust molekulaartunnuste ja huvipakkuva omaduse vahel. Klassifitseeriv mudel kategoriseerib huvipakkuvat omadust mõnesse ettemääratud rühma või annab tõenäosuse kuuluda kindlasse rühma. Levinud regressioonanalüüsi meetoditeks on näiteks lineaarne regressioon, juhumets (ingl k *random forest*), tehisnärvivõrk (ingl k *artificial neural network*) ja tugivektormasin (ingl k. *support vector machine* ehk *SVM*). Klassifikatsiooni mudelid on näiteks logistiline regression, lineaarne diskriminant analüüs (ingl k *linear discriminant analysis*), otsustuspuu (ingl k *decision tree*), juhumets, k lähimat naabrit (ingl k *k nearest neighbours*), tõenäosuslik närvivõrk (ingl k *probabilistic neural network*) ja tugivektormasin. [20]

1.3.2 Molekulaartunnused

Molekulaartunnused on keemilise ühendi või materjali struktuuri põhjal arvutatud parameetrid, mille hulgast valida QSPR sisendparameetreid. Tunnuseid võib saada kas eksperimentaalselt või arvutuslikult ning erinevaid tunnuseid leidub tavaliselt tuhandeid. Molekulaartunnused klassifitseeritakse kahe skeemi põhjal, kas nende mõõtmelisuse järgi või nende määramise meetodi valdkonna järgi. [43] Erinevad tunnuste liigitamise valdkonnad on koostisosalised (näiteks aatomite arv), topoloogilised (graafiteooria põhised), geomeetrilised (nurgad, kaugused, pinnad), kvantkeemilised (laengujaotuspõhised) ja termodünaamilised (entroopia,

tekkeentalpia) tunnused. Tunnuste mõõtmete järgi klassifitseerimisel eristatakse 0D tunnuseid (koostisosalised tunnused), 1D tunnuseid (struktuurvalem), 2D tunnuseid (topoloogilised näitajad) ja 3D tunnuseid (pinnad, ruumalad, kvantkeemilised tunnused, struktuur elektrondifraktsiooni alusel). Mudelitesse 3D tunnuste kaasamise ning nende ennustusvõimekuse osas on jätkuvalt erinevaid arvamusi. 3D tunnuste kaasamise pooldajad toovad välja näiteks lisanduva stereokeemilise informatsiooni eeliseid, kuid võrreldavalt head uurimustulemused ainult 0D, 1D ja 2D tunnustega koostatud lihtsamad ja väiksema arvutusliku taagaga mudelid on tõestanud ka ilma 3D tunnusteta mudelite adekvaatsust. [43] Molekulaartunnuste arvutamiseks on töötatud välja mitmeid tarkvarateeke sealhulgas kommertstarkvarasid nagu näiteks DRAGON 6 (4885 tunnust) ja CODESSA^{TM} (>600 tunnust) ning vabavarasid nagu Mold^2 (777 tunnust) ja Mordred (1826 tunnust). [44-46]

Tunnuste valiku meetodid

QSPR modelleerimine sõltub suurel määral molekulaartunnuste valikust mudelisse. Valikut on võimalik teostada intuitiivselt või kasutades kõiki tunnuseid, kuid peamised põhjused valida alamhulk on (i) ülemääraste ja ebaoluliste tunnuste kahjulik mõju mudeli täpsusele, (ii) vähemate tunnustega mudel on lihtsam, interpreteerivam ja potentsiaalselt kiirem kasutada ning (iii) QSPR algoritmide ajaline keerukus kipub olema suurem kui lineaarne, mis takistab suure tunnuste hulgaga andmehulkade analüüsi. Näide ebaolulisest tunnusest on näiteks konstantse väärtusega tunnus. Tihti eemaldatakse QSPR uurimustes madala dispersiooniga ja omavahel korreleeruvad tunnused. Tunnuste valiku meetod kasutab tavaliselt kindlat algoritmi püüdes saavutada mõne klassifikatsiooni või regressiooni hindamisparameetri maksimaalset väärtust. Seega on protseduur lähedalt seotud vastava kasutatava regressioonialgoritmiga, näiteks lineaarsete ja mittelineaarsete tehnikatega saadakse peaaegu alati erinevad tulemused. Tunnuste valiku algoritme ja nende variatsioone on palju, kuid mõned tuntumad on näiteks ükshaaval valimine (ingl. k. *forward selection*), elimineerimine (ingl. k. *backward elimination*), tunnuste valik koos modelleerimisega, tehisintellektipõhised meetodid (geneetilised algoritmid, tehisnärvivõrk) ja k lähimat naabrit. [41]

Ortogonaalne sobitusalgoritm

Ortogonaalne sobitusalgoritm (ingl. k. *Orthogonal Matching Pursuit* ehk OMP) on arvutuslik meetod, mis valib mudelisse omavahel vähekorreleeruvaid tunnuseid. [47] OMP algoritmi põhimõte on iteratiivselt leida omaduse suhtes kõrgeima korrelatsiooniga tunnus, lisada valitud tunnus igal iteratsioonil lineaarse regressiooni mudelisse ning mudeli ennustatud väärtus

omaduse väärtusest järgmiseks iteratsiooniks maha lahutada. [48] Algoritmi tööpõhimõtte detailid on kirjeldatud algoritmi pseudokoodis (vt **Joonis 3**). Korrelatsioonide arvutamiseks tarvilik eeldus on, et sisendina antud tunnused on standardiseeritud keskväärtusele 0,0 ja standardhälbele 1,0.

ORTOGONAALNE SOBITUSALGORITM		
1: $\mathrm{I} := \emptyset,\ \boldsymbol{r} := \boldsymbol{y},\ \widehat{\boldsymbol{\beta}} := \mathbf{0}$		Seatakse algtingimused; valitud on tühihulk tunnuseid I
2: WHILE ($\lvert \mathrm{I} \rvert < \mathrm{K}$)		Kontrollitakse, kas K tunnust on juba valitud
3:	$\hat{k} := arg\,\max_k \lvert \mathbf{X}_k{}^{\mathrm{T}} \boldsymbol{r} \rvert$	Arvutatakse $\boldsymbol{r}$-i korrelatsioonikoefitsient iga tunnuse suhtes, valitakse suurima koefitsiendi saavutanud tunnuse indeks $\hat{k}$
4:	$\mathrm{I} := \mathrm{I} \cup \{\hat{k}\}$	Valitud indeks lisatakse hulka I
5:	$\widehat{\boldsymbol{\beta}}_{\mathrm{I}} := (\mathbf{X}_{\mathrm{I}}{}^{\mathrm{T}}\mathbf{X}_{\mathrm{I}})^{-1}\mathbf{X}_{\mathrm{I}}{}^{\mathrm{T}}\,\boldsymbol{y}$	Arvutatakse seni valitud tunnuste I väärtustele lineaarse regressiooni koefitsiendid $\widehat{\boldsymbol{\beta}}_{\mathrm{I}}$
6:	$\boldsymbol{r} := \boldsymbol{y} - \mathbf{X}_{\mathrm{I}}\widehat{\boldsymbol{\beta}}_{\mathrm{I}}$	Järgmiseks iteratsiooniks arvutatakse uus $\boldsymbol{r}$ väärtus
7: END WHILE		Naasetakse punkti **2.**

Joonis 3. Ortogonaalse sobitusalgoritmi võimalik teostus (vasakul) ja iga rea ülesande selgitus (paremal). Algoritmis esinevad muutujad on I – valitud tunnusemaatriksi indeksite hulk, $\boldsymbol{r}$ – ajutine muutuja, $\boldsymbol{y}$ – omaduse väärtuste vektor, K – valitud tunnuste kogus, et algoritm lõpetaks, $\hat{k}$ – iteratsioonis valitud tunnusemaatriksi indeks, $\widehat{\boldsymbol{\beta}}$ – regressioonikordajate vektor, $\widehat{\boldsymbol{\beta}}_{\mathrm{I}}$ – regressioonikordajad indeksitel hulgas I, $\mathbf{X}_k$ – tunnus tunnusemaatriksis **X** indeksiga k ja $\mathbf{X}_{\mathrm{I}}$ – tunnused tunnusemaatriksis **X** indeksitega hulgast I. [48]

1.3.3 Multilineaarne regressioon

Multilineaarne regressioonanalüüs on sagedasemalt rakendatud statistiline meetod kahe või enama muutuja omavahel seostamiseks. Lineaarse regressiooni mudelis arvutatakse soovitud omaduse väärtus $\hat{y}$ molekulaartunnuste lineaarse kombinatsioonina valemi kaudu: [20, 49, 50]

$$\hat{y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_{\mathrm{k}} \mathrm{X}_{\mathrm{k}}, \tag{7}$$

kus $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_k$ on molekulaartunnused ja $\beta_0, \beta_1, \ldots, \beta_k$ on tunnuste kordajad ehk regressioonikoefitsiendid. Kordajate leidmisel kasutatakse n vaatlust omaduse väärtuseid y_i ja vastavaid muutujaid $\mathrm{X}_{i1} .. X_{in}$. Üksikul vaatlusel leitud omaduse väärtus avaldub tunnuste kaudu: [49]

$$y_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \beta_2 \mathrm{X}_{i2} + \cdots + \beta_k \mathrm{X}_{ik} + \varepsilon_i, \qquad i = 1,2,\ldots,n, \tag{8}$$

kus ε_i on i-nda vaatluse hindamisviga ennustatud ja eksperimentaalse väärtuse vahel. Antud n vaatlust võib panna kirja maatrikskujul: [49]

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & \mathrm{X}_{11} & \mathrm{X}_{12} & \cdots & \mathrm{X}_{1\mathrm{k}} \\ 1 & \mathrm{X}_{21} & \mathrm{X}_{22} & \cdots & \mathrm{X}_{2\mathrm{k}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{X}_{\mathrm{n1}} & \mathrm{X}_{\mathrm{n2}} & \ldots & \mathrm{X}_{\mathrm{nk}} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{\mathrm{k}} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \text{ ehk } \boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (9)$$

Mudeli parameetrite $\boldsymbol{\beta}$ saamiseks on laialt kasutust leidnud vähimruutude meetod. Leitakse kordajate väärtused, mille puhul hindamisvigade ruutude summa on vähim:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \mathrm{X}_{i1}\beta_1 - \mathrm{X}_{i2}\beta_2 - \ldots - \mathrm{X}_{ik}\beta_k)^2. \quad (10)$$

Ülesanne lahendatakse hindamisvigade ruutude summa iga kordaja järgi osalise tuletise võrdsustamisel nulliga. Tulemusena saadakse kordajate arvutamiseks valem: [49]

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}. \quad (11)$$

Olenevalt ülesandest kasutatakse ka teisi meetodeid peale vähimruutude meetodi, näiteks hindamisvigade absoluutväärtuste summa minimaliseerimist või kaalutud vähimruutude meetodit. Eeltöötlemata tunnustel leitud regressioonikoefitsiendid sisaldavad tunnuste keskväärtust ja standardhälvet. Koefitsientide väärtuste võrreldavaks muutmiseks võib andmeid standardiseerida ehk viia tunnused ning omaduse keskväärtused väärtusele 0,0 ja standardhälbed väärtusele 1,0: [50]

$$\tilde{\mathrm{X}}_{ij} = \frac{\mathrm{X}_{ij} - \overline{\mathrm{X}}_{\cdot j}}{\mathrm{s_j}}, \qquad i = 1,2,\ldots,n\,, \qquad j = 1,2,\ldots,k\,, \quad (12)$$

$$\tilde{y}_i = \frac{\hat{y}_i - \bar{y}}{\mathrm{s}_y}, \qquad i = 1,2,\ldots,n\,, \quad (13)$$

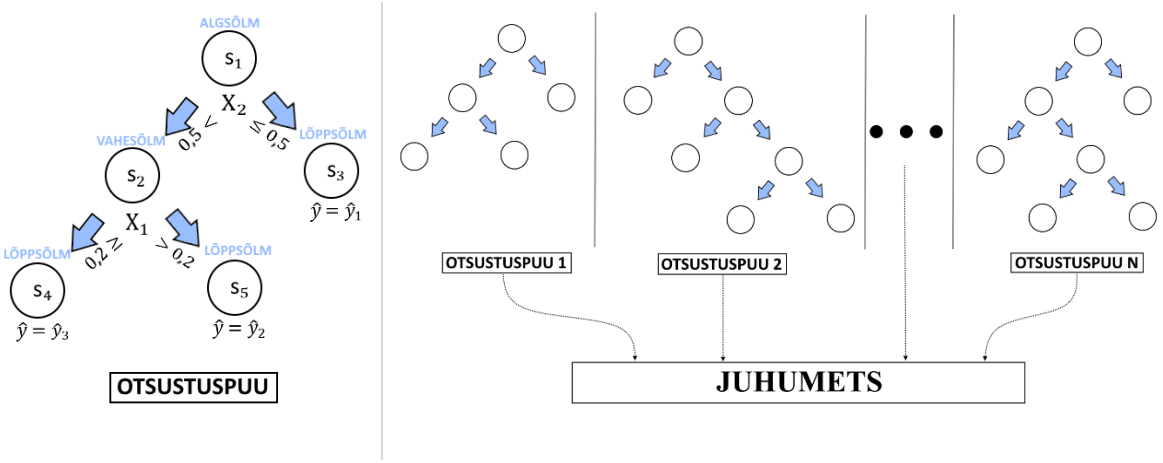
$$\tilde{\beta}_j = \frac{\beta_j \mathrm{s}_j}{\mathrm{s}_y}, \qquad j = 1,2,\ldots,k\,, \quad (14)$$

kus $\overline{\mathrm{X}}_{\cdot j}$ ja s_j on j-nda tunnuse keskväärtus ja standardhälve, $\bar{y}$ ja s_y on omaduse keskväärtus ja standardhälve ning $\tilde{\mathrm{X}}, \tilde{y}, \tilde{\beta}$ on vastavalt standardiseeritud tunnuse, omaduse ja koefitsiendi väärtused. Sel viisil saadud standardiseeritud regressioonikoefitsiendid võimaldavad hinnata nende mõju omaduse väärtuse määramisel võrreldes teiste koefitsientidega. [50]

Eelnevast võime järeldada, et multilineaarne regressioon võimaldab andmestikus leida mõjuväärseid molekulaartunnuseid, nende mõju omavahel võrrelda ja avastada kõrvalekalduvaid andmepunkte. Meetod annab häid tulemusi, kui suhe muutujate ja omaduse vahel on lineaarne. See-eest analüüsitulemus sõltub andmete küllusest ja kvaliteedist ning tihti võivad tunnused olla ka mittelineaarses seoses omadusega, mida multilineaarne regressioon arvesse ei võta.

1.3.4 Juhumetsa regressioon

Juhumetsa regressioon on modelleerimisalgoritm, mis koosneb otsustuspuudest ning juhumetsa ennustatud väärtus on otsustuspuude ennustatud väärtuste aritmeetiline keskmine (vt **Joonis 4**, paremal). [51, 52] Otsustuspuu on hierarhiline otsustussõlmedest $s_1, s_2, \ldots, s_n$ koosnev puud meenutava struktuuriga mudel (vt **Joonis 4**, vasakul). [20] Mudeli ennustatud väärtuse arvutamisel otsustuspuus alustatakse algsõlmest, igas sõlmes valitakse tunnuste väärtuste põhjal järgmine alamsõlm, kuni jõutakse lõppsõlme, millel pole enam alamsõlme. Igale lõppsõlmele on määratud mudeli loomisel arvutatud modelleeritava omaduse väärtus, mis ongi otsustuspuu ennustatud väärtus. [20] Tulenevalt otsustuspuu mudeli loomise protsessist võib juhumetsas iga otsustuspuu struktuur omada erinevat arvu ning erinevalt asetsevaid sõlmesid.



Joonis 4. Otsustuspuu (vasakul) ja juhumetsa (paremal) mudeli struktuur ja tööpõhimõte. Tegemist on hüpoteetiliste mudelite struktuuridega. Näiteks vaatlusel tunnuste väärtustega $X_1 = 0{,}2$ ja $X_2 = 0{,}8$ ennustatakse otsustuspuus omaduse väärtust $\hat{y}_3$. Alustades algsõlmest s_1, valitakse alamsõlm s_2, sest $X_2 = 0{,}8 > 0{,}5$ ja sõlmes s_2 valitakse alamsõlm s_4, sest $X_1 = 0{,}2 \leq 0{,}2$.

Otsustuspuu loomisel valitakse põhisõlmes ja iga kihi alamsõlmedes tunnus ning väärtus, mille põhjal sõlmes olevad andmepunktid jagatakse kahe alamsõlme vahel, kuni alles on vaid lõppsõlmed. [20, 53] Jagamiseks tunnuse ja selle väärtuse valikul on kasutusel mitmeid kriteeriume, millest enamus teevad valiku teatud veafunktsiooni (ingl. k. *impurity function*) hinnangu põhjal. [20, 53] Üks võimalik veafunktsioon regressiooni korral on: [53]

$$i(t) = \frac{1}{N_s} \sum (y - \hat{y}_s)^2, \tag{15}$$

kus N_s on andmepunktide arv sõlmes s, $\hat{y}_s$ on ennustatav väärtus sõlmes s ja y on andmepunkti omaduse väärtus. [53] Võimalikke alamsõlmedeks jaotamise variante võrreldakse selle põhjal, milline jaotamine saavutab kõige suurema veahinnangu langemise: [54]

$$\Delta i(t) = \frac{N_s}{N}[i(s) - \frac{N_{s_P}}{N_s} i(s_P) - \frac{N_{s_V}}{N_s} i(s_V)], \quad (16)$$

kus N, N_{s_P} ja N_{s_V} on vastavalt kõigi andmepunktide arv, andmepunktide arv sõlme s parempoolses alamsõlmes ja andmepunktide arv sõlme s vasakpoolses alamsõlmes ning $i(s)$, $i(s_P)$ ja $i(s_V)$ on vastavalt veahinnang sõlmes s, veahinnang sõlme s parempoolses alamsõlmes ja veahinnang sõlme s vasakpoolses alamsõlmes. Seda, kas sõlm on lõppsõlm otsustatakse mudeli parameetrina valitud kriteeriumi põhjal, milleks on tavaliselt minimaalne treeninghulga andmepunktide arv sõlmes. [54]

Juhumetsa mudelit luues võib valida mitmeid mudeli parameetreid, mis mõjutavad üksikute otsustuspuude ja kogu juhumetsa loomise algoritmi. Parameetreid on väga mitmeid, kuid olulisemate hulka kuuluvad otsustuspuude arv, puu maksimaalne kõrgus ehk maksimaalne otsustuspuu sõlmede kihtide arv, maksimaalne lõppsõlmede arv ja minimaalne andmepunktide arv sõlmes alamsõlmede loomiseks. [54] Üheks parameetriks on valik, kas igale otsustuspuule antakse kõik treeningandmed või juhuvalim treeningandmeid, kus mõned andmepunktid esinevad mitmel korral ning umbes 1/3 andmepunktidest jäävad välja. [51, 52, 54, 55] Teise variandi kasutamist nimetame edaspidi asendustega valimiks (ingl. k. *bootstrap aggregating* või *bagging*). [55] Igas otsustuspuus teatud andmepunktide välja jätmisel on võimalik neid kasutada ka testhulgana (ingl. k. termin *out-of-bag estimate*). [52, 54] Veel üks oluline parameeter on tunnuste arv, mille vahel valitakse otsustuspuus sõlme kaheks alamsõlmeks jagamisel. Levinud valikud on kas kõik tunnused (n) või $\sqrt{n}$ suurune juhuvalim tunnuseid. [54]

Otsustuspuu ei tee eeldusi omaduse väärtuse jaotuse osas, võimaldab kasutada nii reaalväärtusega kui kategoorilise väärtusega andmeid, teostab juba mudelit koostades tunnuste valikut ja on robustne vigaste väärtuste või erandlike väärtuste esinemisel. See võimaldab otsustuspuudega eeldusteta modelleerida keerukaid mittelineaarseid sõltuvusi kasutades andmeid, kus müra esinemisel mudelid on vähem mõjutatud. [53] Otsustuspuu treenimisel saadakse tihti ületreenitud mudel, kus mudel on väga spetsiifiliselt häälestatud antud treeningandmestikule, kuid testandmetel saavutab halvema tulemuse. Juhumetsa kasutamine võimaldab vähendada üksiku otsustuspuu ületreenimise efekti ning suurendada mudeli täpsust testhulgal. [20] Juhumets on võimeline tunnuste olulisust hindama ja on veelgi vähem mõjutatud mürast sisendandmetes, kuid mudeli suutlikkust mõjutab ka väike andmehulk,

mitmekesisuse puudumine andmehulgas, kasutatud otsustuspuude arv ja muud mudeli parameetrid. [20, 51]

1.4 Jaotuskoefitsiendi varasemad arvutusmudelid

Gaas-ioonne vedelik jaotuskoefitsiente on ka varem arvutuslikult modelleeritud ja oluliselt laialdast kasutust on leidnud Abrahami lahustuvusmudel. [56-71] Mudelit on aja möödudes järk-järgult täiustatud üha üldisemale kujule. [58, 61] Abrahami arvutusmudel võimaldab arvutada füüsikalis-keemilisi parameetreid mitme lahusti süsteemides nagu näiteks gaas-ioonne vedelik jaotuskoefitsienti kasutades selleks lahusti ja lahustatava aine tunnusparameetreid. [56] Algse mudeli konstrueeris Michael H. Abraham 1993. aastal gaas-vedelik jaotuskoefitsientide arvutamise tarbeks eksperimentaalsete sisendparameetrite põhjal, nimega üldine lineaarne lahustuvusenergia sõltuvus (ingl. k. *general linear solvation energy relationship* e. *LSER*) või lineaarne vabaenergia sõltuvus (ingl. k. *linear free energy relationship* e. *LFER*). [56] Kirjeldatakse LSER (17) koefitsientide leidmist kahele solvendile ($n_1 = 45$ ja $n_2 = 22$) eksperimentaalselt leitud gaas-vedelik jaotuskoefitsientide jaoks: [56]

$$\log K_{gv} = c + eE + sS + aA + bB + lL, \tag{17}$$

kus E on vedeliku liig molaarmurdumisnäitaja (ingl. k. *excess molar refraction*) ning on arvutatav ühendisegu optilisest murdumisnäitajast, S on vedeliku Kamlet - Taft dipoolsus/polariseeritavus, A ja B on vedeliku summaarne vesiniksideme happelisus ja aluselisus ning L on logaritm gaasifaasis vedeliku - heksadekaani jaotuskoefitsiendist 298 K juures. [57, 58] Uuringus treeninghulgal arvutatud Pearsoni korrelatsioonid ennustatud ja eksperimentaalsete väärtuste vahel olid $R_1 = 0{,}9949$ ning $R_2 = 0{,}9978$. [56] Valemis (17) suurused S, A, B ja L on arvutatavad GLC meetodi tulemustest. [57, 58] Regressioonikordajad c, e, s, a, b ja l on leitavad vähimruutude meetodiga lineaarsel regressioonil. [57] Kordajad väljendavad teatud lahustatava aine ja ioonse vedeliku vahelisi interaktsioone ning iseloomustavad ioonse vedeliku vastavaid keemilisi omadusi. [57] Hilisemad tööd Abrahami mudeliga on leidnud regressioonikoefitsiendid mitmete ioonsete vedelike gaas-ioonne vedelik jaotuskoefitsientide modelleerimiseks LSER abil. [57-71]

Algselt käsitletud Abrahami mudel sisaldas nõrka külge, kus iga uue ioonse vedeliku jaoks $\log K_{giv}$ modelleerimiseks on tarvis eraldi võrrandit ja eksperimentaalseid andmeid. Probleemi adresseerimiseks pakuti 2007. aastal eraldada regressioonikordajad katioonist ja anioonist tulenevateks kordajateks (nn. ioon-spetsiifiline Abrahami mudel): [58]

$$\log K_{giv} = c_{kat} + c_{an} + (e_{kat} + e_{an})E + (s_{kat} + s_{an})S + (a_{kat} + a_{an})A + (b_{kat} + b_{an})B + (l_{kat} + l_{an})L, \quad (18)$$

kus katioonne ja anioonne osa on eraldatud vastavalt alamindeksitega kat ja an. Selline käsitlus võimaldab kasutada juba leitud anioonseid ja katioonseid kordajaid, et ennustada eksperimentaalselt leidmata kombinatsioonide gaas-ioonne vedelik jaotuskoefitsienti. Töös kasutati eelnevas kirjanduses leitud 584 eksperimentaalset $\log K_{giv}$ väärtust katioonsete ja anioonsete regressioonikordajate leidmiseks ning mudel oli kõrge ennustusvõimega treeninghulgal, $R^2 = 0{,}992$. Edasiste regressioonanalüüside tulemuste fikseerimiseks lepiti kokku anioonseteks ja katioonseteks komponentideks jaotamisel nullpunkt, milleks valiti aniooni $[(Tf)_2N]^-$ anioon-spetsiifiline c_{an}. [58] Hilisemad uurimused ioon-spetsiifilise Abrahami mudeliga on leidnud paljude ioonsete vedelike katiooni- ja aniooni-spetsiifilised regressioonikoefitsiendid. [58-71]

Veelgi suurema hulga ioonsete vedelike modelleerimiseks jaotati ühes töös kõik katiooni-spetsiifilised kordajad eraldi alküülahela osade ja funktsionaalrühmade CH_3, CH_2, N, O, $CH_{tsükliline}$, jne kordajateks. Kordajad arvutati summana (nimeks grupi-spetsiifiline Abrahami mudel), näiteks c_{kat}: [61]

$$c_{kat} = \sum_{\text{funkts.rühmad}} n_i c_i, \quad (19)$$

kus c_i on katiooni - spetsiifiline funktsionaalrühma i spetsiifiline kordaja ning n_i on vastava funktsionaalrühma arv katioonis. Nimetatud artiklis kasutati 1450 eksperimentaalset $\log K$ väärtust ning treeninghulgal ennustusvõime oli kõrge, $R^2 = 0{,}997$. [61] Järgnevates artiklites on veelgi täiendatud eksperimentaalandmeid ning ümber arvutatud kindla aniooni- ja katiooni-spetsiifilisi kordajaid ja grupi-spetsiifilisi kordajaid LSER gaas-ioonne vedelik jaotuskoefitsiendi valemites (17), (18) ja (19). [62-71]

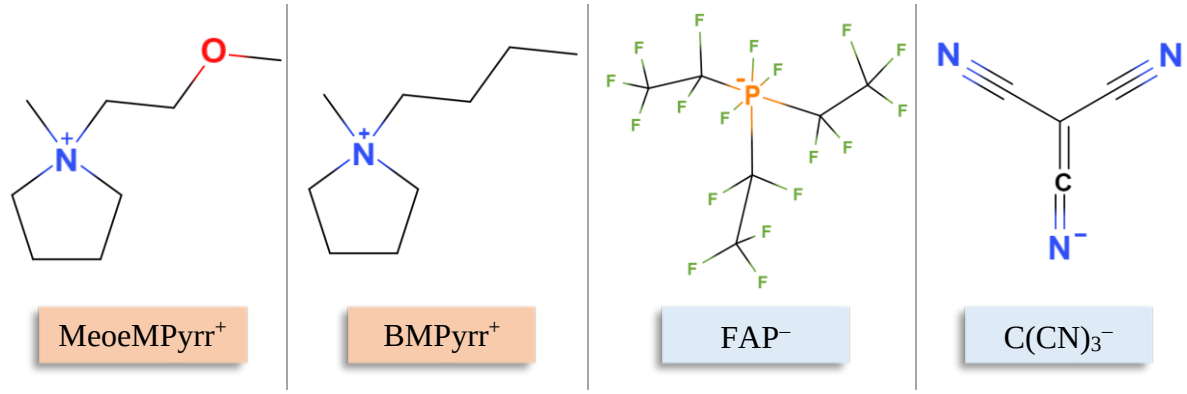
Keemiliste ühendite gaas-ioonne vedelik jaotuskoefitsiendi modelleerimiseks on rakendatud teisigi QSPR meetodeid. Nii näiteks kasutati 90 erineva orgaanilise ühendi $\log K$ väärtust ühendite lahustamisel ioonses vedelikus $[EtOHMIm]^+[FAP]^-$ ehk 1-(2-hüdroksüetüül)-1-metüülimidasool tris(pentafluoroetüül)trifluorofosfaat temperatuuril 323 K. Pärast molekulaartunnuste genereerimist treeniti kaht erinevat QSPR mudelit 60 andmepunkti peal ning testiti 30 andmepunkti peal. Ühe meetodina kasutati tunnuste valimisalgoritmi koos

multilineaarse regressiooniga ning teises meetodis treeniti regressiooniks Gaussian tuumaga SVM koos regulariseerimise ja mudeli parameetrite optimeerimisega. Mudelite kvaliteeti kontrolliti mudeli ennustatud ja eksperimentaalse jaotuskoefitsiendi vahelise korrelatsiooni R arvutamise teel. Valimisalgoritmiga multilineaarne regressioon saavutas korrelatsioonikordaja 0,957 ja SVM korrelatsioonikordaja 0,993 testhulgal. Järeldati, et Gaussian tuumaga SVM on võimeline mittelineaarseid sõltuvusi molekulaartunnuste ja jaotuskoefitsiendi vahel paremini modelleerima, kui alternatiivne multilineaarset regressiooni kasutav meetod. [18]

2. METOODIKA

2.1 Andmekomplekt

Magistritöös kasutatakse varasemas kirjanduses eksperimentaalselt leitud gaas-ioonne vedelik jaotuskoefitsiente ($\log K_{giv}$) temperatuuril 323 K. Edasises tekstis käsitletakse mõisteid omadus, modelleeritud omadus, jaotuskoefitsient, gaas-ioonne vedelik jaotuskoefitsient ja $\log K_{giv}$ samatähenduslikuna. Jaotuskoefitsientide eksperimentaalandmed on üheks andmetabeliks kokku kogutud koostööpartneri William E. Acree Jr. poolt. Pärast andmete ühtlustamist ja kontrollimist valiti andmetabelist kolme erineva ioonse vedeliku, $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$, vastavalt 82, 82 ja 91 jaotuskoefitsienti erinevate orgaaniliste ühendite suhtes. Kolme ioonse vedeliku kohta oli kokku 96 erinevat orgaanilist ühendit (vt ühendite loetelu **Lisas 1**). Modelleerimiseks valitud ioonsete vedelike valiku kriteerium oli, et leiduks vedelike paar, millel erineb vaid katioonne osa ning vedelike paar, millel erineb vaid anioonne osa (vt **Joonis 5**). Kasutatud eksperimentaalsed $\log K_{giv}$ väärtused kuulusid vahemikku 0,016 – 5,477. Andmekomplektist jäeti välja N_2-$[BMPyrr]^+[FAP]^-$ jaotuskoefitsient väärtusega –0,523, mis on ekstreemne võrreldes teistega. Selle jaotuskoefitsiendi kaasamisega mudelite koostamisel kannatas mudelite ennustusvõimekus, mille võimalik põhjus on, et andmestikus ei leidunud N_2-le sarnaste omadustega teisi ühendeid. Lämmastik on ainuke täielikult mittepolaarne ja inertne gaas kõigi orgaaniliste ühendite seast andmestikus.



Joonis 5. Uuritud ioonsete vedelike anioonide ja katioonide molekulstruktuurid.

2.2 Molekulaartunnuste arvutamine

Molekulaartunnuste arvutamiseks kasutati vabavaralist tunnuste arvutamise teeki Mordred. [72] Teek vajas 2D tunnuste arvutuseks sisendina iga molekuli SMILES esitust. SMILES ehk *Simplified Molecular Input Line Entry System* on keemilise struktuuri joontähistuse viis. [73] Näiteks tsükloheksaani molekuli SMILES tähistus on C1CCCCC1. Molekulide SMILES-id koostati molekulide nimetuste põhjal. Saadud ühendite tähistuste sisendiga arvutas Mordred iga ühendi kohta 1613 2D molekulaartunnust ja tulemuseks oli 96 X 1613 ühendite tunnuste andmemaatriks. Nende arvutamine võttis kokku aega ~7,0 sekundit. Lisaks arvutatud tunnustele loodi üks uus molekulaartunnus nimetusega *isPolar*, mille väärtus oli 1 või 0, vastavalt sellele, kas molekul on polaarne või mittepolaarne (vt väärtuseid **Lisas 1**). Uue molekulaartunnuse väärtuste määramisel määrati mittepolaarseks selline molekul, millel kas ei leidunud polaarseid sidemeid (suurema Paulingi elektronegatiivsuse vahega kui 0,5) või leiduvate polaarsete sidemete põhjustatud dipoolmomendi suunavektorid liitusid kokku nullvektoriks.

Molekulaartunnuste eeltöötlus

Molekulaartunnuste arvutamisele järgnes andmete eeltöötlus andmete korrastamiseks ja ülemääraste ning ebaoluliste tunnuste eemaldamiseks. Molekulaar tunnuste andmemaatriksis esinesid arvutusprotsessis tekkinud puuduvad väärtused, mis eeltöötluse etapis asendati nulliga. Kõik konstantse väärtusega molekulaartunnused eemaldati maatriksist. Selle tulemusena oli alles 1239, 1230 ja 1252 tunnust vastavalt ühenditel $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$. Järgnevalt eemaldati tunnused, mis olid samade väärtustega, ning alles jäi vastavalt 1189, 1184 ja 1200 tunnust.

Modelleerimiseks kuluv aeg ja selle tulemus sõltub palju sisendandmetest ning mõlemat annab parandada segava või ebaolulise informatsiooni eemaldamine, mistõttu paljudes QSPR uurimustes eemaldatakse andmestikust omavahel korreleeruvad tunnused. Ka siin töös korreleerusid mitmed tunnused omavahel tugevalt. Seetõttu koostati väiksema kui 0,9 absoluutse omavahelise korrelatsioonikordajaga ($|R|$) vähekorreleeritud tunnustega andmestik. Erinevates QSPR uurimustes on kasutatud nii korrelatsioonikordajat 0,9 kui ka teisi sarnaseid piirväärtuseid. [74, 75] Vähekorreleeritud andmestiku loomiseks leiti iga tunnuse kohta kõik teised tunnused, mille puhul $|R| \geq 0{,}9$ ning alles jäeti üks tunnus. Valikus jäeti alati alles tunnus, mille korrelatsioonikordaja jaotuskoefitsiendiga oli suurim. Lõplikku valikusse jäi seega 430, 439 ja 445 omavahelise korrelatsioonikoefitsiendiga $|R| \geq 0{,}9$

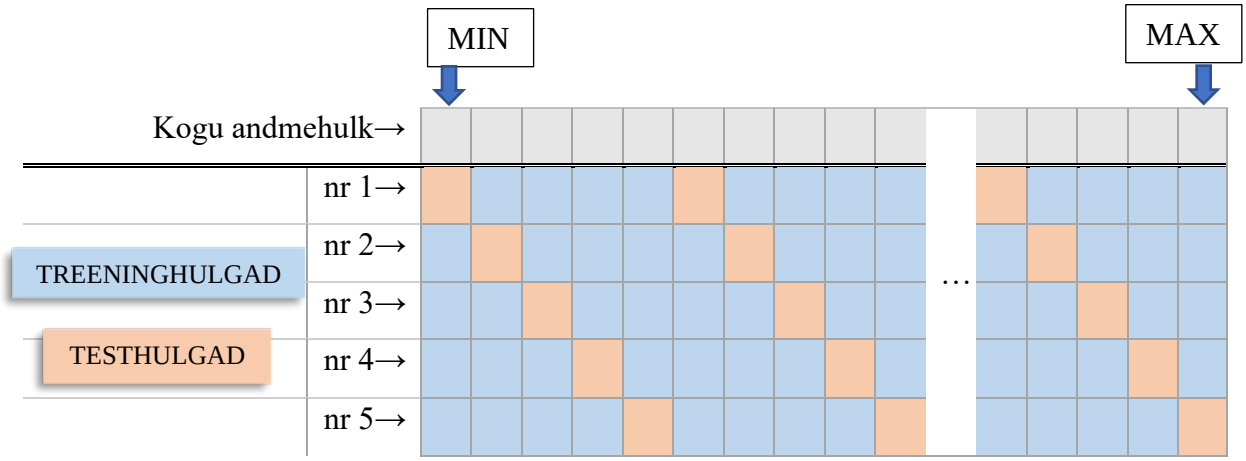
molekulaartunnust vastavalt ühenditele $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$. Siit peale viidatakse seda andmestikku kui **vähe korreleeritud tunnuste hulk** ning tavalist korreleeritud tunnuseid sisaldavat hulka kui **kõigi tunnuste hulk**.

2.3 Mudeli hindamine

Parima mudeli leidmiseks tuleb mudeleid omavahel võrrelda kasutades mudeli ennustusvõimet iseloomustavat hindamisfunktsiooni. Antud töös kasutati parimate mudelite otsingul hindamisfunktsiooni määramiskoefitsient rakendades valemit (3). Lisaks rakendati ristvalideerimise metoodikat (ristvalideeritud määramiskoefitsiendid on tähistatud r_{CV}^2). Mudeli koostamise väliseks hindamiseks arvutati märkimisväärsete tulemustega mudelitele korrelatsiooni ühilduvuskordaja kasutades valemit (6).

Ristvalideerimine

Koostatud mudeliga ennenägemata andmete põhjal omaduse väärtuse prognoosimiseks oli vaja andmestik lahutada treening- ja testhulgaks. Treening- ja testhulga koostamisel on üheks võimaluseks koostada valim eksperimentaalse omaduse jaotus järgides. Ristvalideerimiseks koostatakse mitu valimit, võttes mudeli lõplikuks hinnanguks nende valimite hindamisfunktsiooni väärtuste aritmeetiline keskmine. Jaotuskoefitsiendi väärtuselt sarnaste hulkade koostamiseks sorteeriti orgaanilised ühendid jaotuskoefitsiendi väärtuse järgi ning grupeeriti viieks võimalikult võrdseks hulgaks. Esimesse hulka kuulus kasvavas jaotuskoefitsiendi järjekorras 1., 6., 11., … orgaaniline ühend, teise hulka 2., 7., 12., … orgaaniline ühend ja sarnasel viisil võttes iga 5. ühendi, koostati ka kolmas, neljas ja viies hulk (vt **Joonis 6**). Nii koostatud viis ühendite hulka sisaldasid kogu jaotuskoefitsientide väärtuste vahemikus ühtlaselt jaotunud väärtuseid. Lisaks võimaldas grupeering arvutada mudeli hinnang viiel erineval testhulgal. Neli gruppi võeti treeninghulgaks ja allesjäänud hulka kasutati testhulgana. Mudeli võrreldavaks hinnanguks võeti käesolevas uuringus viiel erineval testhulgal arvutatud ristvalideeritud hinnang.



Joonis 6. Andmete grupeerimine ristivalideerimiseks. Iga rida tähistab eraldi treening- ja testhulgaks jaotamist. Helesinisega on tähistatud treeninghulga elemendid ja oranžiga on tähistatud testhulga elemendid. Näiteks testhulka 1 kuuluvad kasvavas log K järjekorras elemendid 1., 6., 11., ... ja treeninghulka 1 kuuluvad elemendid 2. – 5., 7. – 10., … .

2.4 Multilineaarne regressioon

Multilineaarse regressioonanalüüsi teostamiseks kasutati *sklearn* teegi moodulit *LinearRegression*. Mudelisse valiti tunnused kasutades ortogonaalset sobitusalgoritmi. Sobitusalgoritmi tööks oli vajalik enne tunnuste valiku etappi jõudmist standardiseerida tunnuste väärtused keskväärtusele 0,0 ja standardhälbele 1,0, milleks kasutati valemit (12). Algoritmi teostusena kasutati OMP moodulit *sklearn* teegis. [54] OMP algoritm leiab arvutuslikult tõhusal viisil omavahel vähe korreleeruvad tunnused, mis võimaldab mudelis arvesse võtta võimalikult erineva keemilise sisuga informatsiooni. Tunnuste arvuks, mille juures OMP algoritm lõpetab töö valiti neli tunnust ning *sklearn* moodulis määrati selleks parameeter $n\_nonzero\_coefs = 4$. Viienda tunnuse lisamisel ei paranenud mudeli täpsus märkimisväärsel määral.

Multilineaarse regressiooni tunnuste valiku algoritmi kirjeldab **Joonis 7**. Algoritmi otsinguruumi laiendamiseks eemaldati omadusega suurimat korrelatsiooni väärtust omav tunnus võimalikest valikutest ning kasutati OMP algoritmi uute parimate tunnuste valikuks. Seda eemaldusmeetodit korrati, kuni mudeli kasutada jäid alles tunnused, millel $R < 0{,}4$ jaotuskoefitsiendi suhtes.

LINEAARSE REGRESSIOONI TUNNUSTE VALIKU ALGORITM

Algsisendid: Kõigi tunnuste hulk M, valitud tunnuste hulkade tühihulk T
1. Hulgast M valida tunnused I kasutades OMP algoritmi
2. Salvestada valitud tunnused I hulka T
3. Leida iga hulka I valitud tunnuse ja omaduse vaheline Pearsoni korrelatsioon R, valem (4)
4. Valida **3.**-s leitud korrelatsioonide hulgast kõrgeima absoluutse korrelatsioonikordajaga R_{max} tunnus ning eemaldada see hulgast M
5. Juhul, kui $R_{max} \geq 0{,}4$, siis alustada uuesti **1.**-st
Väljundid: Valitud tunnuste hulkade hulk T

Joonis 7. Multilineaarse regressiooni tunnuste valiku algoritm.

2.5 Juhumetsa regressioon

Juhumetsa regressiooni mudelite koostamiseks kasutati *sklearn* teegi moodulit *RandomForestRegressor*. [54] Mudelite koostamisel viidi esmalt läbi tunnustekomplekti valik kahes etapis ning sellele järgnes parameetrite optimeerimine samuti kahes etapis. Mudeli parameetrid algse tunnustekomplekti valikul võtab kokku **Tabel 1**. Parima mudeli tunnustekomplekti valikul treeniti 100 otsustuspuud. Algsel otsustuspuude arvu valikul tugineti kaalutlustel kasutada piisavalt vähe puid arvutusliku koormuse kokkuhoiuks ning samas võimalikult palju puid, et ennustusvõime ei kannataks. Oshiro et al. uuring on näidanud, et nende eesmärkide täitmiseks on mõistlik valida puude arv vahemikus 64 … 128. [76] Mudeli reprodutseeritavuse eesmärgil määrati ka juhuslikkuse seeme väärtusega 42. Selle parameetri määramisel juhumetsa algoritmis asendustega andmepunktide valimi koostamisel valitakse juhuvalim igal algoritmi käivitusel samal viisil. Seega mudeli tulemuste reprodutseerimiseks saab sel viisil panna algoritmi samasid otsuseid tegema ehk saavutatakse sama tulemus. Ülejäänud mudeli parameetrid jäeti vaikeväärtustele. Juhumetsa tunnuste valiku algoritmi kirjeldab detailselt **Joonis 8**.

Tabel 1. Juhumetsa mudeli parameetrid tunnuste valikul.

Parameetri nimi (*nimi sklearn-is*) →	Parameetri väärtus
Otsustuspuude arv (*n_estimators*) →	100
Alamsõlmedeks jagamisel tunnuse valiku veafunktsioon (*criterion*) →	valem (15)
Suurim otsustuspuu kihtide arv (*max_depth*) →	Piiritu
Vähim arv andmepunkte sõlmes jagamiseks (*min_samples_split*) →	2
Vähim arv admepunkte lõppsõlmes (*min_samples_leaf*) →	1
Jagamisel juhuvalimi suurus kõigi n tunnuse hulgast (*max_features*) →	n
Asendustega andmepunktide valim ehk bagging (*bootstrap*) →	Jah
r^2 väljajäetud andmepunktidel (*oob_score*) →	Ei
Juhuslikkuse seeme (*random_state*) →	42

JUHUMETSA REGRESSIOONI TUNNUSTE VALIKU ALGORITM
Algsisendid: Tunnuste hulk M, valitud tunnuste tühihulk I, leitud tunnuste hulkade hulk T
Etapp 1 (tunnuste valik)
1. Iga tunnusega hulgast M treenida kõigil treeninghulkadel juhumets, kokku 5 mudelit iga tunnuse kohta. Mudeli tunnusteks on üks tunnus hulgast M koos seni valitud tunnuste hulgaga I **2.** Leida ristvalideeritud määramiskoefitsient iga tunnuse kohta üle 5 mudeli kasutades vastavaid testhulki **3.** Määrata valitud tunnuste hulgaks I tunnuste hulk, millega saavutati punktis **2.** kõrgeim ristvalideeritud hinnang **4.** Korrata alapunkte **1.** - **3.** valides igal iteratsioonil hulka I juurde üks tunnus, kuni kokku on neli tunnust
Etapp 2 (tunnuste asendamine)
5. Asendada valitud tunnuste hulgast I üht tunnust kõigi teiste tunnustega hulgast M ning leida iga asenduse puhul ristvalideeritud hinnang. Salvestada suurima ristvalideeritud hinnanguga tunnustekomplekt hulka T **6.** Määrata valitud tunnuste hulgaks I tunnuste hulk, millega saavutati punktis **5.** suurim ristvalideeritud hinnang **7.** Korrata alapunkte **5.** - **6.**, proovides iga kord asendada erinevat tunnust, kuni ühelgi neljast tunnusest asendamisel enam paremat ristvalideeritud hinnangut ei leita
Väljundid: Valitud tunnuste hulk I, leitud tunnuste hulkade hulk T

Joonis 8. Juhumetsa regressiooni tunnuste valiku algoritm.

Tunnuste valiku algoritmis prooviti algse nelja tunnuse valikut nii vähekorreleeritud tunnuste hulgast kui ka kõigi tunnuste hulgast. Hilisemas tunnuste asendamise etapis kasutati alati kõikide tunnuste hulka. *Sklearn* teegi juhumetsa mooduli sisseehitatud tunnuste valiku meetodit mudeli väljundparameetri *feature_importances_* kujul ei kasutatud, sest see ei andnud usaldusväärseid ja korratavaid tulemusi. Juhumetsa korral on seda probleemi täheldatud ka varasemates töödes. [77, 78] Usaldusväärsuse probleem on seotud omavahel kõrgelt korreleeruvate tunnustega, mida arvutatud molekulaartunnuste hulgas leidus hulganisti.

Tabel 2. Juhumetsa mudeli parameetrite optimeerimisel proovitud väärtused esimeses etapis.

Parameetri nimi (*nimi sklearn-is*) →	{Võimalikud väärtused}
Otsustuspuude arv (*n_estimators*) →	$\{16, 32, 64, 128, 256, 512, 1024, 2048\}$
Jagamisel juhuvalimi suurus kõigi n tunnuse hulgast (*max_features*) →	$\{n, \sqrt{n}\}$
Suurim otsustuspuu kihtide arv (*max_depth*) →	$\{1,4,9,16,25,36,49,64,81, 'Piiritu'\}$
Vähim arv andmepunkte sõlmes jagamiseks (*min_samples_split*) →	$\{2, 3, 4\}$
Vähim arv admepunkte lõppsõlmes (*min_samples_leaf*) →	$\{1, 2, 3\}$
Asendustega andmepunktide valim (*bootstrap*) →	$\{'Jah', 'Ei'\}$

Mudeli parameetrite optimeerimine

Tunnuste valikule järgnes juhumetsa mudeli parameetrite optimeerimine. Optimeerimine toimus kahes etapis. Esmalt prooviti laiemas vahemikus võimalikult erinevaid parameetrite väärtuseid käies läbi vaid alamhulga kõikidest võimalikest parameetrite väärtuste kombinatsioonidest (vt **Tabel 2**). Selle tulemusena koostati kõigi 2880 võimaliku kombinatsiooni hulgast 300 erineva kombinatsiooniga mudelit. Kuna ristvalideeritud hinnang

arvutati viiel testhulgal, siis reaalsuses treeniti 300 kombinatsiooni puhul 1500 mudelit. Reprodutseeritavuse eesmärgil kasutati juhuslikkuse seemet väärtusega 42. Esmase juhuotsingu etapi järgselt kasutati mudelites esimese etapi tulemuste põhjal valitud kitsamaid väärtuste vahemikke ning läbi prooviti kõik kombinatsioonid. Lisaks tunnuse valiku etapis leitud parimatele tunnustele prooviti parameetreid optimeerida ka kasutades mudeleid tunnustekomplektidega hulgast T tunnuste valiku algoritmi väljundis. Teatud mudeli parameetrite kombinatsioon võib mõnda parimatest mudelitest tõsta kõrgema hinnangu peale, kui seda on algsete mudeli parameetritega leitud kõrgeima ristvalideeritud hinnanguga tunnustekomplekt. Hulka T sattus kümneid erinevaid tunnustekomplekte ning paljud neist saavutasid parima tunnustekomplektiga võrreldes märkimisväärselt madalama tulemuse (kuni 0,1 r_{CV}^2 võrra), mistõttu parameetreid optimeeriti vaid parimale tulemusele lähedaste tulemustega tunnustekomplektidega. Lähedaseks tulemuseks loeti kuni 0,01 võrra kehvemat r_{CV}^2 väärtust.

Tulemuse juhuslikkuse kontroll

Parameetrite optimeerimisel tuli ette olukordi, kus suure hulga parameetrite kombinatsioonide läbi proovimisel leiti selline kombinatsioon, mis juhuste kokkulangemisel saavutas teistest parema tulemuse. Tunnuste valiku algoritm leidis seega mudelisse tunnuseid, mis saavutasid juhuslikkuse seemne 42 juures parema hinnangu kui mõni teine tunnus, mis oli keskmiselt parema tulemusega. Üheks selliseks näiteks on $[BMPyrr]^+[C(CN)_3]^-$ jaotuskoefitsientide mudelisse leitud tunnustekomplekt AATS1m, ATS0s, ATSC2se ja Xch-7dv, mis seemnega 42 saavutas r_{CV}^2 väärtuse 0,8861, kuid seemnete 43 – 47 peale keskmise tulemuse 0,8753. Seetõttu tehti juhumetsa tunnuste valikut (**Joonis 8**) ka algoritmi variatsiooniga, mis kontrollis parema hinnanguga tunnustekomplekti leidmisel punktides **3.** ja **6.**, kas seemnete 43 – 47 peale keskmine tulemus oli parem eelmisest parimast keskmisest tulemusest. Juhuslikkuse seemnete 43 – 47 peale keskmine tulemus on lisatud juhumetsa tulemustele tähisega $\overline{r_{CV}^2}$.

3. TULEMUSED JA ARUTELU

Tulemuste esitamisel vaadeldakse kõigepealt optimaalseid multilineaarse regressiooni mudeleid koos mudelitesse valitud molekulaartunnuste analüüsiga. Mudelitesse valitud tunnuste väärtuste tagamaid uurides on võimalik kirjeldada mudelis loodud seoseid molekulide keemilise ja füüsikalise sisu ning vastava ioonse vedeliku jaotuskoefitsiendi vahel. Lineaarse regressiooni mudelite analüüsile järgneb juhumetsa mudelite tunnuste valiku ja parameetrite optimeerimise tulemused, millega kaasneb samuti optimaalsete juhumetsa mudelite tunnuste analüüs. Kasutades mudelite analüüsi tulemusi võrreldakse konstrueeritud mudeleid lähtudes modelleerimismeetodist ning edasi võrreldakse ka ühist iooni omavate ioonsete vedelike mudeleid.

3.1 Multilineaarse regressiooni mudelid

Multilineaarse regressioonanalüüsi tulemusel saadi kolm lineaarset mudelit (võrrandid 20 – 22). Standardiseeritud andmetel arvutatud regressioonikordajatega mudelid on:

$$\log K_{g[\mathrm{BMPyrr}]+[\mathrm{FAP}]-} = 3{,}051 + 0{,}6247 \cdot VR1\_A + 0{,}4299 \cdot SsOH - 0{,}3283 \cdot AATS0i - 0{,}6943 \cdot GATS1are, \quad (20)$$

$$\log K_{g[\mathrm{BMPyrr}]+[\mathrm{C(CN)3}]-} = 3{,}044 + 0{,}6248 \cdot nHBDon + 0{,}5046 \cdot VSA\_EState9 + 0{,}3264 \cdot MDEC - 22 - 0{,}7149 \cdot GATS1m, \quad (21)$$

$$\log K_{g[\mathrm{MeoeMPyrr}]+[\mathrm{FAP}]-} = 3{,}270 + 0{,}6204 \cdot ATS1m + 0{,}5711 \cdot isPolar - 0{,}4114 \cdot MAXssO - 0{,}7021 \cdot AXp - 1dv. \quad (22)$$

Need mudelid on edaspidi tähistatud vastavalt LR1, LR2 ja LR3. Kõikidel mudelitel arvutatud statistilised hindamisparameetrid treening- ja testhulkadel on toodud **Tabelis 3**. Kõrged määramiskoefitsiendid ja madal ruutkeskmine viga testhulkadel näitavad, et mudelid on hea ennustusvõimega. Kõikidel mudelitel arvutatud ristvalideeritud määramiskoefitsient oli suurem kui 0,87 ning on suurima väärtusega LR1 korral, 0,925. Korrelatsiooni ühilduvuskordajad kõikidel test- ja treeninghulkadel ületasid kokkuleppelist ennustusvõimeliseks lugemise väärtust 0,85. Arvutatud ristvalideeritud korrelatsiooni ühilduvuskordajad olid kõik suuremad kui 0,93, mis kinnitab, et mudelid on suurepärase ennustusvõimega. Iga ioonse vedeliku parima tunnustehulgaga multilineaarse regressiooni mudeli LR1, LR2 ja LR3 ennustusvead ehk eksperimentaalse ja ennustatud väärtuse vahed on töö **Lisades 2, 3 ja 4** kõrvutatuna juhumetsa parimate mudelite ennustusvigadega.

Tabel 3. Multilineaarse regressiooni optimaalsete mudelite määramiskoefitsiendid, ruutkeskmised vead (MSE) ja korrelatsiooni ühilduvuskordajad (CCC) treening- ja testhulkadel 1 – 5. Arvude tausta värviskaalas punasest roheliseni tähistatakse paremaid tulemusi järjest tumedama rohelisega.

	r^2			MSE			CCC		
	LR1	LR2	LR3	LR1	LR2	LR3	LR1	LR2	LR3
Treening1	0,9286	0,8928	0,9054	0,0907	0,1365	0,1032	0,9630	0,9433	0,9504
Test1	0,9424	0,8619	0,9015	0,0988	0,2090	0,1225	0,9739	0,9293	0,9499
Treening2	0,9281	0,8912	0,8969	0,0988	0,1458	0,1118	0,9627	0,9425	0,9457
Test2	0,9586	0,8547	0,9364	0,0539	0,1818	0,0808	0,9784	0,9098	0,9681
Treening3	0,9369	0,8888	0,9231	0,0863	0,1494	0,0882	0,9674	0,9412	0,9600
Test3	0,9209	0,8726	0,8088	0,1062	0,1577	0,1960	0,9569	0,9386	0,8872
Treening4	0,9329	0,8815	0,9087	0,0940	0,1581	0,1049	0,9653	0,9370	0,9522
Test4	0,9386	0,9021	0,8925	0,0742	0,1252	0,1095	0,9676	0,9457	0,9456
Treening5	0,9475	0,8814	0,9071	0,0735	0,1570	0,1063	0,9731	0,9370	0,9513
Test5	0,8633	0,9025	0,8915	0,1634	0,1280	0,1124	0,9216	0,9512	0,9428
Treening_{CV}	0,9348	0,8871	0,9082	0,0887	0,1494	0,1029	0,9663	0,9402	0,9519
Test_{CV}	0,9248	0,8788	0,8862	0,0993	0,1603	0,1242	0,9597	0,9349	0,9387

Molekulaartunnuste analüüs

Põhinedes standardiseeritud multilineaarse regressiooni koefitsientidel valemites (20) on võimalik hinnata erinevate tunnuste olulisust ning seda, kuidas need mudeliga ennustatud jaotuskoefitsiendi väärtust mõjutavad. Multilineaarse regressiooni mudelite molekulaartunnuste analüüsil võeti aluseks nende väärtused (**Lisad 5 – 15**) ja need omavad selget keemilist ja füüsikalist sisu.

LR1 tunnused

Molekulaartunnuse *VR1_A* väärtuste võrdlemisel oli näha, et molekulis sideme vahetamisel kaksik- või kolmiksideme vastu väärtus ei erinenud ning süsinikaatomi väljavahetamisel mõne teise aatomi vastu samuti mitte (vt väärtuseid **Lisa 5**). See-eest mõjutas kõige rohkem tunnuse väärtust vesinikust raskemate aatomite arv molekulis, sest suurema aatomite arvuga molekulil arvutatud *VR1_A* väärtus oli alati suurem. Lisaks ilmnes, et sama aatomite arvuga, kuid vähem hargnenud süsinikuahelaga molekulil arvutatakse kõrgem *VR1_A* väärtus ning aromaatset tuuma sisaldaval molekulil veelgi kõrgem *VR1_A* väärtus. Seega vähematest aatomitest (vesinikke arvestamata) koosneval molekulil määratakse LR1 mudelis madalama *VR1_A* väärtuse tõttu madalam log $K_{g[BMPyrr]+[FAP]-}$ väärtus.

Tunnuse *SsOH* väärtus oli null, kui molekulis ei leidu hüdroksüülrühma (vt väärtuseid **Lisa 6**). Seega *SsOH* esinemisest mudelis positiivse koefitsiendiga saame järeldada, et

hüdroksüülrühma esinemisel gaasis g on ennustatud log $K_{g[BMPyrr]+[FAP]-}$ väärtus suurem. Lisaks *SsOH* väärtuste uurimisel oli ilmne, et pikema süsinikuahelaga molekulidel oli *SsOH* väärtus suurem ning aromaatset rühma sisaldades veelgi suurem. Seega põhjustavad nimetatud omadused ka suuremat ennustatud log $K_{g[BMPyrr]+[FAP]-}$ väärtust.

AATS0i tunnuse lähemal uurimisel ilmnes, et tunnuse väärtused aromaatsetel molekulidel olid erandlikult madalad ning lühema ahelaga hapnikku või lämmastikku sisaldavate mittearomaatsete molekulide puhul olid tunnuse väärtused kõrged (vt väärtuseid **Lisa 7**). Seega aromaatse tuuma esinemisel gaasis g on *AATS0i* väärtus madalam ning LR1 mudeli ennustatud log $K_{g[BMPyrr]+[FAP]-}$ väärtus suurem.

Molekulaartunnuse *GATS1are* väärtused olid väga sarnased sama funktsionaalrühma sisaldavatel ühenditel ning üldiselt mida suurem laengute eraldatus esines molekulis, seda madalam oli tunnuse väärtus (vt väärtuseid **Lisa 8**). Näiteks nitro- või nitriilrühma sisaldavate molekulide *GATS1are* väärtused olid madalaimate seas. Lisaks kordsete sidemeteta alkaanide ahelatel olid suurimad *GATS1are* väärtused, millele järgnesid veidi madalama väärtusega kordsete sidemetega või tsüklit sisaldavad alkaanid. Arvutati ka *GATS1are* korrelatsioonikordaja tunnusega isPolar ja selleks oli –0,7304. Seega väljendasid need tunnused teatud määral sarnast informatsiooni ja olid üksteise suhtes vastandliku mõjuga jaotuskoefitsiendile. Suurema laengute eraldatusega molekulidel määrati LR1 mudelis järelikult tunnuse *GATS1are* panuse tõttu kõrgem log $K_{g[BMPyrr]+[FAP]-}$ väärtus ning kordsete sidemeteta alkaaniahelatel madalam väärtus.

LR2 tunnused

Tunnus *nHBDon* väljendab vesiniksidet andvate rühmade arvu molekulis ning selles andmestikus oli see tunnus nullist erinev hüdroksüülrühma sisaldavatel ühenditel ja pürroolil (vt väärtuseid **Lisa 9**). Seega ennustas mudel kõrgemat jaotuskoefitsiendi log $K_{g[BMPyrr]+[C(CN)3]-}$ väärtust ühenditel, mis olid võimelised andma vesiniksidet.

Seaduspärasused tunnuse *VSA_Estate9* väärtustes pole nii ilmsed, kuid võttes arvesse tunnuse arvutuseeskirja Mordredi lähtekoodis saab näha, et tunnuse arvutamisel on kokku liidetud molekuli aatomite elektrotopoloogilise oleku indeksid, lühidalt E-olekud (ingl. k. *electrotopological state* ehk *E-state*). [79, 80] E-olek väljendab aatomi elektronegatiivsust või elektronide ruumilist ligipääsetavust ning seda võib interpreteerida kui tõenäosust teise molekuliga interakteeruda. [81] Näiteks oktaan ei sisalda kõrge elektronegatiivsusega funktsionaalrühmasid ning selle *VSA_Estate9* väärtuseks arvutati 13,0, samas kui oktanaal ja

1-oktanool sisaldavad vastavalt okso- ja hüdroksügruppi ning nende *VSA_Estate9* väärtused olid palju kõrgemad, vastavalt 20,0 ja 18,5 (vt väärtuseid **Lisa 10**). Järelikult ühenditel, milles on rohkem kõrgema elektronegatiivsusega funktsionaalrühmasid, prognoosis LR2 mudel ka suuremat log $K_{g[BMPyrr]+[C(CN)3]-}$ väärtust.

Sarnasel viisil interpreteeriti ka tunnust *MDEC-22*, mille arvutamisel lähtekoodis on loetud kokku sekundaarsete süsinikupaaride arv jagatuna nende vahelise kaugusega. [80] Seega tunnuse *MDEC-22* panuse läbi ennustati suurema arvu sekundaarsete süsinikega ühendil g kõrgemat log $K_{g[BMPyrr]+[C(CN)3]-}$ väärtust (vt väärtuseid **Lisa 11**).

Tunnus *GATS1m* omas suurimaid väärtuseid lühikeste alkaanide puhul, millele järgnevad pikemad kordsete sidemetega ahelatega molekulid ning väärtused olid sarnased ühenditel, milles esines sama funktsionaalrühm (vt väärtuseid **Lisa 12**). *GATS1m* väärtuste vähenemise järjekorras mõningate eranditega sisaldasid molekulid järgmiseid funktsionaalrühmi: hüdroksü-, eeter-, okso-, karboksürühm, lämmastikku sisaldavad rühmad ja halogeenrühm. Üldiselt, mida rohkem süsinikust raskemaid aatomeid ja kordseid sidemeid esines, seda madalam oli molekulil arvutatud *GATS1m*. Järelikult negatiivse regressioonikoefitsiendi tõttu valemis ennustati rohkete süsinikust raskemate aatomitega ja kordsete sidemete arvuga ühendite puhul suuremat log $K_{g[BMPyrr]+[C(CN)3]-}$ väärtust.

LR3 tunnused

Mudelis LR3 tunnuse *ATS1m* väärtused on järjest suuremad rohkemate aatomitega molekulides (vt väärtuseid **Lisa 13**). Lisaks kõige suuremad väärtused tulid molekulide puhul, milles esines seejuures ka mõni süsinikust raskem aatom, näiteks halogeenirühma sisaldavatel ühenditel olid väärtused suurimad. Seega kõrgem log $K_{g[BMPyrr]+[C(CN)3]-}$ ennustati *ATS1m* panuse läbi suurema arvu aatomitega ja raskemaid aatomeid sisaldavatel molekulidel.

Tunnuse *isPolar* tõttu mudelis hinnati polaarsete molekulide puhul suuremat jaotuskoefitsiendi väärtust. Seda kinnitab koefitsiendi positiivne väärtus.

Molekulaartunnuse *MAXssO* väärtused olid nullist erinevad juhul kui esines hapnik, mis annab sideme kahele vesinikust erinevale aatomile (vt väärtuseid **Lisa 14**). Seejuures estrite puhul arvutati madalamad väärtused, kui eetrite puhul. Negatiivse korrelatsioonikoefitsiendi tõttu prognoosis LR3 mudel madalamat jaotuskoefitsiendi väärtust eetritel ja estritel.

Tunnuse *AXp-1dv* puhul võis täheldada madalaimaid väärtuseid aromaatset tuuma sisaldavatel ühenditel (vt väärtuseid **Lisa 15**). Lisaks, mida rohkem hapniku või lämmastiku

aatomeid ühend sisaldas, seda madalam oli *AXp-1dv* väärtus. Suurimad väärtused olid see-eest halogeenrühma sisaldavatel molekulidel. Seega *AXp-1dv* mõjul prognoositi aromaatse tuumaga ja rohkemate hapnike ja lämmastikega ühenditel kõrgemat jaotuskoefitsiendi väärtust ning halogeenrühmaga ühenditel madalamat väärtust.

3.2 Juhumetsa regressiooni mudelid

Parimate juhumetsa mudelite (**Tabel 4**) kõrged määramiskoefitsiendid ja madal ruutkeskmine viga testhulkadel näitavad, et ka juhumetsa mudelid on hea ennustusvõimega. Kõikide juhumetsa mudelite r_{CV}^2 väärtus oli 0,90 läheduses, mis näitab, et mudelitega on võimalik hea täpsusega prognoosida vastavaid gaas-ioonne vedelik jaotuskoefitsiente. Samasugust tulemust näitavad ka keskmised ristvalideeritud määramiskoefitsiendid, $\overline{r_{CV}^2}$, juhuslikkuse kontrollist. Mudelite head ennustusvõimet kinnitasid ka testvalimi ristvalideeritud korrelatsiooni ühilduvuskordajad, mis on kõik suuremad kui 0,94 (**Tabel 5**). Iga ioonse vedeliku optimaalse juhumetsa regressiooni mudeli RF1, RF2 ja RF3 ennustusvead on töö **Lisades 2**, **3** ja **4** kõrvutatuna juhumetsa parimate mudelite ennustusvigadega.

Tabel 4. Juhumetsa regressiooni optimaalsed mudelid ja vastavad ristvalideeritud määramiskordajate r_{CV}^2 ja keskmiste ristvalideeritud määramiskordajate $\overline{r_{CV}^2}$ väärtused. Need mudelid on edaspidi tähistatud vastavalt RF1, RF2 ja RF3.

Tähis	Ioonne vedelik	Molekulaartunnused				r_{CV}^2	$\overline{r_{CV}^2}$
RF1	$[BMPyrr]^+[FAP]^-$	GATS1s	IC0	MATS1p	TMPC10	0,9261	0,9250
RF2	$[BMPyrr]^+[C(CN)_3]^-$	MID_h	MATS1i	PEOE_VSA1	ETA_beta	0,8943	0,8985
RF3	$[MeoeMPyrr]^+[FAP]^-$	AATS1m	ATSC0are	piPC6	ATSC0c	0,8911	0,8912

Mudeli parameetrid	RF1	RF2	RF3
Otsustuspuude arv	2048	128	256
Jagamisel juhuvalimi suurus kõigi n tunnuse hulgast	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$
Suurim otsustuspuu kihtide arv	Piiritu	Piiritu	8
Vähim arv andmepunkte sõlmes jagamiseks	2	2	2
Vähim arv admepunkte lõppsõlmes	1	1	1
Asendustega andmepunktide valim	Ei	Ei	Ei

Tabel 5. Juhumetsa optimaalsete mudelite määramiskoefitsiendid, ruutkeskmised vead (MSE) ja korrelatsiooni ühilduvuskordajad (CCC) treening- ja testhulkadel 1 – 5. Arvude tausta värviskaalas punasest roheliseni tähistatakse paremaid tulemusi järjest tumedama rohelisega.

	r^2			MSE			CCC		
	RF1	RF2	RF3	RF1	RF2	RF3	RF1	RF2	RF3
Treening1	1,0000	0,9975	0,9996	0,0000	0,0027	0,0005	1,0000	0,9988	0,9998
Test1	0,9195	0,8159	0,8645	0,1382	0,2290	0,2051	0,9531	0,9005	0,9203
Treening2	1,0000	0,9967	0,9991	0,0000	0,0036	0,0012	1,0000	0,9983	0,9995
Test2	0,9398	0,8882	0,8446	0,0784	0,1421	0,1945	0,9677	0,9361	0,9159
Treening3	1,0000	0,9943	0,9991	0,0000	0,0065	0,0012	1,0000	0,9971	0,9996
Test3	0,9509	0,9526	0,9533	0,0659	0,0486	0,0578	0,9740	0,9747	0,9758
Treening4	1,0000	0,9955	0,9993	0,0000	0,0052	0,0010	1,0000	0,9977	0,9996
Test4	0,8927	0,9105	0,8908	0,1296	0,0912	0,1397	0,9347	0,9542	0,9391
Treening5	1,0000	0,9973	0,9992	0,0000	0,0031	0,0011	1,0000	0,9987	0,9996
Test5	0,9274	0,9045	0,9024	0,0868	0,0989	0,1281	0,9621	0,9509	0,9490
Treening_{CV}	1,0000	0,9963	0,9992	0,0000	0,0042	0,0010	1,0000	0,9981	0,9996
Test_{CV}	0,9261	0,8943	0,8911	0,0998	0,1220	0,1450	0,9583	0,9433	0,9400

Tunnuste valik

Tunnuste valiku algoritm leidis selle esimeses etapis tunnusekomplektid, mille r^2_{CV} väärtused olid 0,8656^{VK}, 0,8801^{VK} ja 0,8729 vastavalt ioonsetele vedelikele $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$. Ülaindeksiga VK tähistatud väärtusel kasutati esmases tunnuse valikus vähekorreleeritud tunnuste hulka ning ilma ülaindeksita tulemusel kõigi tunnuste hulka. Edasises parameetrite optimeerimise etapis uurimise alla võetud parima $\overline{r^2_{CV}}$ väärtuse suhtes 0,01 suuruse hälbe sisse jäävad tulemused esitab **Tabel 6**.

Tabel 6. Juhumetsa regressiooni parimate mudelite tunnused koos r^2_{CV} väärtustega juhuslikkuse seemnel 42 ja $\overline{r^2_{CV}}$ väärtustega juhuslikkuse seemnetel 43 – 47.

Ioonne vedelik	Molekulaartunnused				r^2_{CV}	$\overline{r^2_{CV}}$
$[BMPyrr]^+[FAP]^-$	GATS1s	IC0	MATS1p	TMPC10	0,9112^{VK}	0,9070^{VK}
	IC0	MATS1p	VR2_Dzp	GATS1s	0,9023^{VK}	0,9000^{VK}
$[BMPyrr]^+[C(CN)_3]^-$	MID_h	MATS1i	PEOE_VSA1	ETA_beta	0,8880^{VK}	0,8927^{VK}
$[MeoeMPyrr]^+[FAP]^-$	AATS1m	piPC6	ATSC0c	NssO	0,8866	0,8858
	IC0	RPCG	SMR_VSA3	VE1_Dzse	0,8857^{VK}	0,8822^{VK}
	IC0	RPCG	SMR_VSA3	VE1_DzZ	0,8857^{VK}	0,8820^{VK}
	IC0	RPCG	SMR_VSA3	VE1_Dzpe	0,8860^{VK}	0,8818^{VK}
	fragCpx	IC0	RPCG	SMR_VSA3	0,8801^{VK}	0,8805^{VK}
	AATS1m	piPC6	ATSC0c	nN	0,8788	0,8802
	AATS1m	ATSC0are	piPC6	ATSC0c	0,8799	0,8773

Mudeli parameetrite optimeerimine ja mudelite hindamine

Juhumetsa mudeli parameetreid optimeeriti parimal ning parimale lähedase tulemusega tunnusekomplektide jaoks ehk $[BMPyrr]^+[FAP]^-$ puhul kahel tunnustekomplektil, $[BMPyrr]^+[C(CN)_3]^-$ puhul ühel ning $[MeoeMPyrr]^+[FAP]^-$ puhul seitsmel tunnusekomplektil. Kõikide optimeeritud mudelite tunnusekomplektid olid juba mainitud **Tabelis 6**. Esimese etapi parameetrite juhuotsingu tulemusena leiti algsete parameetritega mudeliga võrreldes 94, 15 ja 19 parema tulemusega parameetritekomplekti vastavalt ioonsetel vedelikel $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$. Optimeerimise esimeses etapis parima tulemuse ja selle lähedaste tulemuste puhul arvutati juhuslikkuse kontrolliks ka $\overline{r_{CV}^2}$. Lähedaseks tulemuseks loeti ka siin parima r_{CV}^2 väärtusega võrreldes kuni 0,01 hälbe sisse jäänud r_{CV}^2 väärtust.

Parimad parameetrid esimeses etapis

Esmases etapis leitud parema $\overline{r_{CV}^2}$ väärtusega mudelitel olid alati parameetrid 'jagamisel juhuvalimi suurus kõigi n tunnuse hulgast' väärtusega $\sqrt{n}$, 'asendustega andmepunktide valim' väärtusega EI ehk ei kasutatud asendustega andmepunktide valimit, 'vähim arv andmepunkte lõppsõlmes' vaikeväärtusega 1 ning 'vähim arv andmepunkte sõlmes jagamiseks' vaikeväärtusega 2. Seetõttu neid parameetreid optimeerimise teises etapis ei muudetud ja jäeti neile väärtustele. Parameetrid, mille väärtused kõigi ioonsete vedelike parimate mudelite puhul varieerusid olid 'otsustuspuude arv' ning 'suurim otsustuspuu kihtide arv', mistõttu nende mõju teises etapis kitsamates väärtusevahemikes uuriti. Parameetri 'otsustuspuude arv' puhul saadi alati parem tulemus 128 või suurema arvu otsustuspuude rakendamisel mudelis. Näiteks nii $[BMPyrr]^+[FAP]^-$ kui ka $[MeoeMPyrr]^+[FAP]^-$ puhul saavutas parima $\overline{r_{CV}^2}$ väärtuse 2048 otsustuspuuga mudel. Seega teises etapis prooviti väärtusevahemikke alates 128-st. Parimate $\overline{r_{CV}^2}$ väärtustega mudelitel oli 'suurim otsustuspuu kihtide arv' alati kas 9, 16, 81 või 'Piiritu', kusjuures tihti 81 ja 'Piiritu' puhul saadi täpselt sama tulemus. Nendes mudelites ilmselt otsustuspuude kihtide arv ei jõudnud üle 81 kasvada juba üksnes andmepunktide arvu põhjustatud piirangu tõttu. Leitud 'suurim otsustuspuu kihtide arv' väärtuste põhjal valiti teiseks etapiks lähedased väärtused (vt **Tabel 7**).

Tabel 7. Juhumetsa mudeli parameetrite optimeerimisel proovitud väärtused teises etapis.

Parameetri nimi (*nimi sklearn-is*) →	{Võimalikud väärtused}
Otsustuspuude arv (*n_estimators*) →	$\{128, 256, 512, 1024, 2048\}$
Suurim otsustuspuu kihtide arv (*max_depth*) →	$\{8, 12, 16, 20, 'Piiritu'\}$

Lisaks saadi esimeses etapis $[MeoeMPyrr]^+[FAP]^-$ jaoks tunnuste hulga *AATS1m*, *ATSC0are*, *piPC6* ja *ATSC0c* puhul valdava enamusega kõrgemad $\overline{r_{CV}^2}$ väärtused, mistõttu kasutati seda teises etapis. Kahel ülejäänud vedelikul jäi parim tunnustekomplekt samaks nagu see oli tunnuste valiku järgselt. Parimad $\overline{r_{CV}^2}$ väärtused esimeses etapis olid 0,9250, 0,8985 ja 0,8899 vastavalt vedelikel $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$.

Parimad parameetrid teises etapis

Teises optimeerimise etapis leiti, et parameetri 'otsustuspuude arv' väärtusel 128 ja suurematel väärtustel ei saavutanud märkimisväärselt erinevaid tulemusi, sest kõikide ioonsete vedelike puhul $\overline{r_{CV}^2}$ väärtused erinesid üksteisest maksimaalselt ~0,005 võrra. Siiski, $[BMPyrr]^+[FAP]^-$ puhul oli suurema otsustuspuude arvu korral ka kõrgem $\overline{r_{CV}^2}$, kuid $[BMPyrr]^+[C(CN)_3]^-$ puhul oli näha hoopis vastupidist seost ning $[MeoeMPyrr]^+[FAP]^-$ puhul parimad kaks mudelit olid 256 ja 128 otsustuspuuga. Sellest tulenevalt valiti $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$ mudelite otsustuspuude arvuks vastavalt 2048, 256 ning 128.

Parameetri 'suurim otsustuspuu kihtide arv' erinevate väärtuste puhul erinesid $\overline{r_{CV}^2}$ väärtused samuti üksteisest maksimaalselt ~0,005 võrra. $[MeoeMPyrr]^+[FAP]^-$ mudelites parameetri väärtusel 8 saavutati kõrgeim $\overline{r_{CV}^2}$ väärtus, kuid $[BMPyrr]^+[C(CN)_3]^-$ puhul samal parameetri väärtusel saavutati alati madalaim $\overline{r_{CV}^2}$ väärtus. $[BMPyrr]^+[FAP]^-$ mudelites ei leidunud üht selgelt paremat parameetri 'suurim otsustuspuu kihtide arv' väärtust. Seetõttu valiti $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$ mudelite suurimaks otsustuspuu kihtide arvuks vastavalt 'Piiritu', 'Piiritu' ja 8.

Molekulaartunnuste analüüs

Juhumetsa regressiooni valitud tunnuste sisu tõlgendamine võimaldab leida seaduspärasusi, mis antud tunnuse olulisust mudelis võivad selgitada. Molekulaartunnuste analüüsil võeti aluseks nende väärtused (**Lisad 16 – 27**). Otsustuspuu tööpõhimõtte tõttu leiab juhumetsa mudelis aset tunnuse väärtuste grupeerimine väärtusevahemikeks. Kindlatesse väärtusevahemikesse kuuluvate tunnustega molekulil ennustatakse alati sama jaotuskoefitsient erinevalt multilineaarsest regressioonist, kus tunnuse väärtus korrutatakse regressioonikoefitsiendiga.

RF1 tunnused

Tunnuse *GATS1s* (**Tabel 4**) lähemal vaatlusel saab näha, et tunnuse väärtused on enamjaolt sarnased ühenditel, mis sisaldavad samasid struktuure: kordne side, sama funktsionaalrühm, aromaatse ning mittearomaatse tsükli esinemine (väärtused **Lisa 16**).

IC0 väärtused on grupeeritud järgnevalt: alkaanid, mittearomaatse tsükliga ja/või kordse sidemega funktsionaalrühmadeta süsinikuahelad, benseenituuma sisaldavad funktsionaalrühmadeta süsinikuahelad, hapnikku või halogeenrühma sisaldavad ühendid, benseenituuma ja hapnikku sisaldavad ühendid ning lämmastikku või mitut hapnikku sisaldavad ühendid (väärtused **Lisa 17**). *IC0* väärtused tundusid mõnel määral korreleeruvat molekuli polaarsusega ning seda kinnitas tunnuse korrelatsioonikordaja *isPolar* tunnusega, 0,7869.

MATS1p väärtuste puhul võis näha järgmisi eristatavaid gruppe: aromaatseid ja mittearomaatseid tsükleid sisaldavad ühendid, halogeenrühmaga ühendid, sama pika hargnemata süsinikuahelaga ühendid ja hapnikku sisaldavad ühendid (väärtused **Lisa 18**).

TMPC10 väärtustes olid üldiselt sama suure aatomite arvuga molekulidel sarnased väärtused ning benseenituuma sisaldavad ühendid olid enamasti sarnaste väärtustega mõne erandiga (väärtused **Lisa 19**). Eelnevalt kirjeldatud tunnuste iseärasustest tulenevalt võis mudel hüpoteetiliselt jaotuskoefitsiendi prognoosimisel erinevalt arvesse võtta molekuli järgnevaid struktuuri omadusi: benseenituuma olemasolu, mittearomaatse tsükli leidumine, halogeenrühm, aatomite arv, hapniku või lämmastiku olemasolu, süsinikuahela hargnevus ja kordse sideme leidumine.

RF2 tunnused

Tunnus *MID_h* RF2 mudelis (**Tabel 4**) omas järgmiseid väärtuste grupeeringuid: funktsionaalrühmadeta ühendid, estrid, eetrid, alkoholid, sama pikka süsinikuahelat sisaldavad ühendid ja halogeenrühma sisaldavad ühendid (väärtused **Lisa 20**).

MATS1i puhul olid eristatavad grupid: tsüklit sisaldavad ühendid, sama pika süsinikuahelaga ühendid, täpselt samadest aatomitest koosnevad sama funktsionaalrühma sisaldavad ühendid, sealhulgas estrid, eetrid ja alkoholid (väärtused **Lisa 21**).

Molekulaartunnuse *PEOE_VSA1* väärtuseid võib grupeerida vastavalt: eetrid ja estrid, aldehüüdid ning hüdroksüülrühmaga ühendid (väärtused **Lisa 22**).

ETA_beta puhul olid sarnased väärtused lähedase aatomite arvuga ühenditel ja aromaatsetel ühenditel (väärtused **Lisa 23**). Nende tunnuste väärtuste kirjeldustest tulenevalt võis mudel hüpoteetiliselt jaotuskoefitsiendi ennustamisel erinevalt arvesse võtta molekuli järgnevaid struktuuri omadusi: funktsionaalrühma puudumine, esterrühm, eeterrühm, OH-rühm, aldehüüdrühm, halogeenrühm, tsükli olemasolu, aromaatsus, süsinikuahela pikkus ning aatomite arv.

RF3 tunnused

RF3 mudelis (**Tabel 4**) *AATS1m* tunnuse väärtused olid enamjaolt sarnased lähedase aatomite arvuga ühenditel. Lisaks oli võimalik väärtusevahemikke määrata järgmisteks gruppideks: alkoholid, eetrid, oksoühendid, funktsionaliseerimata aromaatsed ühendid, funktsionaliseeritud aromaatsed ühendid ja nitroühendid (väärtused **Lisa 24**).

ATSC0are väärtuste põhjal on võimalik tuvastada järgmisi gruppe: funktsionaalrühmata ühendid, nitrorühma sisaldavad ühendid, lämmastikuga ühendid v.a. nitrorühm, estrid ja karboksüülhapped ning üldiselt hapnikku sisaldavad ühendid (väärtused **Lisa 25**). *ATSC0are* väärtused jaotusid teatud määral molekulide polaarsuse järgi ning korrelatsioonikordaja tunnusega isPolar, 0,7078, näitas, et selle tunnuse läbi võis RF3 mudel arvestada ka molekuli polaarsust.

Tunnuse *piPC6* väärtused võib jaotada järgmistesse kategooriatesse: aromaatsed ühendid ja lähedase aatomite arvuga vähemalt 7-st aatomist koosnevad ühendid (väärtused **Lisa 26**).

ATSC0c väärtused olid enamasti sarnased sama funktsionaalrühma sisaldavatel ühenditel. Seejuures võis eristada gruppe: estrid, lämmastikku ja benseenituuma sisaldavad ühendid, alkoholid ja eetrid (väärtused **Lisa 27**).

Neid grupeerimisi täheldades võis mudel hüpoteetiliselt jaotuskoefitsiendi ennustamisel erinevalt arvesse võtta molekuli järgnevaid struktuuri omadusi: aatomite arv, polaarsus, OH-rühm, eeterrühm, oksorühm, funktsionaalrühm ja aromaatne tuum, funktsionaalrühma puudumine ja aromaatne tuum, nitrorühm, funktsionaalrühma puudumine, esterrühm, karboksürühm, hapniku olemasolu ja üldiselt sama funktsionaalrühma olemasolu.

3.3 Mudelite võrdlus

Järgnevalt võrdleme erinevas regressiooni meetodis valitud molekulaartunnuste sisu sama ioonse vedeliku raames. Mudelites ilmnes sarnasusi kuid ka erinevusi katiooni või aniooni vahetamisel ioonses vedelikus.

3.3.1 Lineaarsed vs juhumetsa mudelid

Multilineaarse ja juhumetsa regressiooni optimaalsete mudelite testhulkadel arvutatud ristvalideeritud määramiskoefitsientide põhjal võib märkida, et mudelite ennustusvõimekus ei erinenud palju, kuid juhumetsa mudelid saavutasid jaotuskoefitsientide prognoosimisel vähesel määral parema täpsuse. Seega üldiselt saavutati paremaid tulemusi mudeliga, mis võtab arvesse lisaks lineaarsetele seostele tunnuste ja omaduse vahel ka mittelineaarseid seoseid.

Tabel 8. Optimaalsete mudelite tunnuste omavahelised korrelatsioonikordajad.

Multilineaarse regressiooni mudelite tunnuste vaheline korrelatsioon

LR1	VR1_A	SsOH	AATS0i	GATS1are
VR1_A	1,000	-0,182	-0,499	-0,007
SsOH	-0,182	1,000	0,144	-0,017
AATS0i	-0,499	0,144	1,000	0,056
GATS1are	-0,007	-0,017	0,056	1,000

LR2	ATS1m	isPolar	MAXssO	AXp-1dv
ATS1m	1,000	-0,025	0,011	0,131
isPolar	-0,025	1,000	0,334	-0,180
MAXssO	0,011	0,334	1,000	-0,087
AXp-1dv	0,131	-0,180	-0,087	1,000

LR3	nHBDon	VSA_EState9	MDEC-22	GATS1m
nHBDon	1,000	-0,014	-0,276	0,137
VSA_EState9	-0,014	1,000	0,128	0,025
MDEC-22	-0,276	0,128	1,000	0,052
GATS1m	0,137	0,025	0,052	1,000

Juhumetsa regressiooni mudelite tunnuste vaheline korrelatsioon

RF1	GATS1s	IC0	MATS1p	TMPC10
GATS1s	1,000	-0,712	0,065	-0,082
IC0	-0,712	1,000	-0,132	-0,100
MATS1p	0,065	-0,132	1,000	0,572
TMPC10	-0,082	-0,100	0,572	1,000

RF2	AATS1m	ATSC0are	piPC6	ATSC0c
AATS1m	1,000	0,086	0,391	0,191
ATSC0are	0,086	1,000	-0,096	0,878
piPC6	0,391	-0,096	1,000	0,013
ATSC0c	0,191	0,878	0,013	1,000

RF3	MID_h	MATS1i	PEOE_VSA1	ETA_beta
MID_h	1,000	-0,489	0,605	-0,049
MATS1i	-0,489	1,000	-0,554	0,366
PEOE_VSA1	0,605	-0,554	1,000	-0,281
ETA_beta	-0,049	0,366	-0,281	1,000

Juhumetsa regressiooni mudelitesse valitud tunnused korreleerusid omavahel rohkem kui tunnused multilineaarse regressiooni mudelites (vt **Tabel 8**). Tunnuste valiku algoritm multilineaarses regressioonis üritas saavutada võimalikult madalat korrelatsiooni tunnuste

vahel samas kui juhumetsa tunnuste valiku algoritmis polnud madal tunnuste omavaheline korrelatsioon valiku tingimuseks. Seega tunnuste vaheline madal korrelatsioonikordaja polnud siin vajalik, et juhumetsa mudel saavutaks samaväärset või täpsemat mudelit kui multilineaarne regressioon. See-eest võimaldab madal tunnuste vaheline korrelatsioon selgemalt eristada molekuli füüsikalisi või keemilisi iseärasusi, mis määravad omaduse väärtust. Samuti seda, kui mõjurikas on üks tunnus võrreldes teisega.

log $K_{g[BMPyrr]+[FAP]-}$ modelleerimine

LR1 ja RF1 mõlema tunnuste väärtustest oli näha mustreid, mis võisid arvesse võtta: aatomite arvu molekulis, molekuli polaarsust, benseenituuma olemasolu, süsinikuahela hargnevust, teatud funktsionaalrühma esinemist, lämmastiku või hapniku olemasolu või kordse sideme leidumist. LR1 mudeli tunnuste *VR1_A* ja *GATS1are* suuremad regressioonikoefitsiendid annavad alust arvata, et teatud funktsionaalrühma esinemine, molekuli polaarsus, aromaatsus ja aatomite arv molekulis olid olulisimad molekuli struktuuri omadused jaotuskoefitsiendi väärtuse kujunemisel. RF1 tunnuste vaheline korrelatsioon oli keskmiselt kõrgem kui LR1 tunnuste vaheline korrelatsioon ning tunnuste sisust oli võimalik näha, et RF1 tunnustes olev informatsioon kattus rohkem, näiteks kõikide RF1 tunnuste korral aromaatset tuuma sisaldavad ühendid olid sarnaste tunnuste väärtustega. LR1 mudeli tunnustes polnud halogeenrühma sisaldavatel ühenditel tunnuse väärtused sarnased ning on võimalik, et RF1 mudel arvestas selle rühma olemasolu erinevalt. Sellegipoolest võis näha LR1 ja RF1 tunnustes rohkelt sarnasusi ning mõlemal mudelil arvutati ka sarnaselt kõrge määramiskoefitsiendi väärtus. Multilineaarse regressiooni mudeli tunnuste väärtuste põhjal prognoosib mudel, et keemilise ühendi jagunemisel gaasi ja $[BMPyrr]^+[FAP]^-$ vahel võib ioonses vedelikus leida suurema osa seda ühendit juhul, kui ühend on polaarne, rohkemate aatomitega molekulidega, sisaldab aromaatset tuuma ja OH-rühma. Jaotuskoefitsiendi väärtuste kasvav järjekord **Lisas 28** kinnitab ka selle paikapidavust.

log $K_{g[BMPyrr]+[C(CN)3]-}$ modelleerimine

Log $K_{g[MeoeMPyrr]+[C(CN)3]-}$ modelleerimisel oli juhumetsa ja multilineaarse mudelite vahel rohkem erinevusi, kui teiste ioonsete vedelike puhul. Sarnasused LR2 ja RF2 mudelite tunnuste sisu vahel on järgnevate molekulide omaduste võimalik arvesse võtmine: OH-, eeter-, okso- ja halogeenrühma esinemine, funktsionaalrühmade puudumine, süsinikuahela pikkus ning aatomite arv. Erinevalt LR2 tunnustest oli võimalik, et RF2 tunnustes arvestati ka aromaatse tuuma ning esterrühma olemasolu. See-eest oli näha, et LR2 mudelis tunnuse *VSA_EState9*

panusega hinnati molekulide aatomite elektronide ruumilist ligipääsetavust samas kui RF2 mudeli tunnuste hulka ei valitud ühtki E-olekul põhinevat tunnust. Lisaks leiti LR2 mudeli optimaalse komponendina arvestada tunnuse *MDEC-22* panuse läbi sekundaarsete süsinike arvu ning tunnuse *GATS1m* kaasamisel kordsete sidemete esinemist ühendis. LR2 tunnuste väärtuste põhjal ennustab mudel, et keemilise ühendi jagunemisel gaasi ja $[BMPyrr]^+[C(CN)_3]^-$ vahel võib ioonses vedelikus leida suurema osa ühendit, mis annab vesiniksidet, on ruumiliselt ligipääsetavamate elektronidega aatomitega, rohkemate süsinikust raskemate aatomitega ja sisaldab kordseid sidemeid. See-eest ennustab mudel vastupidist mõju lühemate alkaanide puhul. Jaotuskoefitsiendi väärtuste kasvava järjekorra põhjal **Lisas 29** on tõesti näha, et lühemate alkaanide puhul on jaotuskoefitsient väiksem ning rohkemate süsinikust raskemate aatomitega ja vesiniksidet andvate molekulide puhul on jaotuskoefitsiendi väärtus üldiselt suurem.

log $K_{g[MeoeMPyrr]+[FAP]-}$ modelleerimine

Kolmanda ioonse vedeliku jaotuskoefitsiendi modelleerimisel leiti järgmised sarnasused LR3 ja RF3 vahel võimalike arvestatavate molekuli omaduste osas: aatomite arv, aromaatse tuuma olemasolu, hapniku olemasolu, esterrühm, eeterrühm, halogeenrühm ja molekuli polaarsus. LR3 mudelis ennustati halogeenrühma sisaldavatel ühenditel *ATS1m* tunnuse tõttu kõrgemat väärtust, kuid samas *AXp-1dv* panuse läbi madalamat väärtust. On võimalik, et RF3-s eristati erinevate funktsionaalrühmade mõju jaotuskoefitsiendile täpsemalt, mis selgitaks mõnevõrra kõrgemat määramiskoefitsiendi väärtust. Multilineaarse regressiooni tunnuste sisu analüüsist võib järeldada, et mudel ennustab keemilise ühendi jagunemisel gaasi ja $[MeoeMPyrr]^+[FAP]^-$ vahel, et ioonses vedelikus võib leida suurema osa ühendit, mis on polaarne, sisaldab esterrühma või eeterrühma, on aromaatse tuumaga, on rohkemate aatomitega, sisaldab süsinikust raskemaid aatomeid ning sisaldab hapnikku või lämmastikku. Jaotuskoefitsiendi väärtuste kasvav järjekord **Lisas 30** kinnitab ka nendel omadustel põhinevate seoste üldist paikapidavust.

3.3.2 Ühist iooni omavate ioonsete vedelike võrdlus

Optimaalsete mudelite tulemuste võrdlusel $[BMPyrr]^+[FAP]^-$ modelleerimisel saavutati palju kõrgem määramiskoefitsient. Võimalik põhjus võis olla, et tunnuse valiku algoritmi ei leidnud teiste ioonsete vedelike puhul niivõrd optimaalseid tunnuseid. Lisaks võis määramiskoefitsienti mõjutada ka valitud ristvalideerimise metoodika, sest erinevate ioonsete vedelike puhul olid

treening- ja testhulgad alati erinevad. Võis juhtuda, et kasutatud metoodika jaotas molekule $[BMPyrr]^+[FAP]^-$ andmestiku puhul mitmekesisemalt treening- ja testhulkade vahel. Tõenäolisem on, et ristvalideerimise metoodika mõjutas siin tulemusi kõige rohkem, sest nii multilineaarse kui juhumetsa regressiooni puhul olid tulemused sarnased.

Katiooniga $[BMPyrr]^+$ ioonsete vedelike modelleerimine

Eelnevast katiooni $[BMPyrr]^+$ sisaldavate ioonsete vedelike mudelite analüüsist lähtudes saab märkida järgnevaid sarnasusi mudelites mõjuvamate orgaaniliste ühendite struktuuriomaduste arvestamisel: aatomite arv molekulis, aromaatse tuuma olemasolu, kordse sideme leidumine, OH- või halogeenrühma esinemine molekulis ning üldisemalt süsinikust raskemate aatomite esinemine molekulis. Eriti selgelt kajastus mõlema mudeli molekulaartunnustes OH-rühma ehk vesiniksidet andva rühma olemasolul suurema jaotuskoefitsiendi väärtuse prognoosimine. Jaotuskoefitsiendi väärtuste uurimisel võib tõepoolest näha, et kui molekulis erineb vaid OH-rühma olemasolu, siis väärtus on alati palju kõrgem, näiteks heksaan ja pentaan võrreldes 1-butanooli ja 1-pentanooliga. Samal viisil on aatomite arv selgelt oluline. Eriti hästi on seda näha pentaani, heksaani, heptaani, oktaani, nonaani ja dekaani puhul jaotuskoefitsiendi väärtuste kasvamise näitel. Erineva aniooniga ioonsete vedelike mudelid erinesid selle poolest, et $[BMPyrr]^+[FAP]^-$ puhul arvestati mudelites selgelt ka molekuli polaarsust, samas kui $[BMPyrr]^+[C(CN)_3]^-$ jaotuskoefitsiendi modelleerimisel selle asemel leidis mudelis kasutust E-oleku tunnus, mis kirjeldab molekuli aatomite elektronide ruumilist ligipääsetavust.

Aniooniga $[FAP]^-$ ioonsete vedelike modelleerimine

Aniooni $[FAP]^-$ sisaldavate ioonsete vedelike mudelite analüüsil põhinedes võib täheldada järgnevaid sarnasusi mudelites mõjuvamate orgaaniliste ühendite struktuuriomaduste arvestamisel: aatomite arv molekulis, polaarsus, aromaatse tuuma olemasolu, OH- või halogeenrühma esinemine molekulis, lämmastiku või hapniku olemasolu molekulis ning üldisemalt süsinikust raskemate aatomite esinemine molekulis. Sama aniooniga ioonsete vedelike modelleerimisel ei leitud palju erinevusi oluliste molekulide omaduste arvestamisel mudelites. Võimalik põhjus võis olla, et nende katioonid $[BMPyrr]^+$ ja $[MeoeMPyrr]^+$ erinesid vaid ühe eeterrühma poolest ning mõlemas ioonis esines pürrolidiin. Ka jaotuskoefitsiendi väärtuste vaatlusel võib märkida, et nende kasvav järjestus on märkimisväärselt sarnane.

KOKKUVÕTE

Magistritöös koostati juhumetsa ja multilineaarse regressiooni mudelid prognoosimaks ioonsete vedelike $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$ gaas-ioonne vedelik jaotuskoefitsienti. Orgaaniliste ühendite gaas-ioonne vedelik jaotuskoefitsiendid jagati ristvalideeritud hinnangute tarbeks viieks hulgaks. Nende orgaaniliste ühendite jaoks arvutati molekulaartunnused ning rakendati tunnuse valikul kahte lähenemist: ortogonaalne sobitusalgoritm (OMP) ja ükshaaval valik (*Forward Selection*) baasil. Juhumetsa mudelite parameetreid optimeeriti. Parimate ristvalideeritud mudelite võrdlus näitas, et mittelineaarne juhumetsa mudel suudab gaas-ioonne vedelik jaotuskoefitsiente prognoosida täpsemini, kui multilineaarse regressiooni mudel. Multilineaarse regressiooniga saavutatud lihtsamad mudelid ei olnud aga märkimisväärselt kehvemad. Suurim ristvalideeritud määramiskoefitsiendi väärtus nii multilineaarse regressiooni kui juhumetsa regressiooni puhul saavutati $\log K_{g[BMPyrr]+[FAP]-}$ mudelitel, mis olid vastavalt 0,925 ja 0,926. Konstrueeritud QSPR mudelite korrelatsiooni ühilduvuskordaja (CCC) on >0,85, mis tõendab nende mudelite ennustusvõimekust. Eriti näitab kõrget ennustusvõimet fakt, et multilineaarse regressiooni mudelite CCC > 0,93 ja juhumetsa regressiooni mudelite CCC > 0,94. Optimaalsetesse mudelitesse valitud tunnuste analüüsil leiti, et gaas-ioonne vedelik jaotuskoefitsient saab olla olulisel määral mõjutatud järgmistest orgaaniliste ühendite struktuuriomadustest: aatomite arv, aromaatse tuuma olemasolu, kindla funktsionaalrühma leidumine, eriti OH-rühm, polaarsus ja süsinikust raskemate aatomite esinemine molekulis. Katiooniga $[BMPyrr]^+$ ioonsete vedelike mudelid erinesid selle poolest, et $\log K_{g[BMPyrr]+[FAP]-}$ mudelites arvestati selgelt ka molekuli polaarsust, samas kui $\log K_{g[BMPyrr]+[C(CN)3]-}$ puhul leidis selle asemel mudelis kasutust E-oleku tunnus. Aniooniga $[FAP]^-$ ioonsete vedelike mudelid omasid näiliselt sarnaseid keemilisi ja füüsikalisi omadusi arvesse võtvaid molekulaartunnuseid, mis võis olla tingitud sellest, et nende katioonid erinesid vaid ühe eeterrühma poolest. Antud tulemused on heaks toeks ja aluseks edasistele uurimustele gaas-ioonne vedelik jaotuskoefitsiendi modelleerimisel ning järgmistes töödes võiks proovida koostada katiooni-spetsiifilisi, aniooni-spetsiifilisi ning veel üldisemate mudelite konstrueerimist. Samuti võiks katset teha ka teiste mittelineaarsete meetoditega ning mitmeid mudeleid kombineerivate meetodiga.

SUMMARY

Within this master's theses, random forest and multilinear regression models were developed to predict the gas-to-ionic liquid partition coefficient of ionic liquids $[BMPyrr]^+[FAP]^-$, $[BMPyrr]^+[C(CN)_3]^-$ ja $[MeoeMPyrr]^+[FAP]^-$. The experimental partition coefficients of ionic liquids for different organic compounds were divided into five chemically diverse sets for finding cross-validated estimates. Molecular descriptors were calculated for these organic compounds and feature selection algorithms based on the Orthogonal Matching Pursuit and Forward Selection were applied. The hyperparameters of the random forest models were tuned. A comparison of the best cross - validated models showed that the non - linear random forest model can predict gas-to-ionic liquid partition coefficients more accurately than the multilinear regression model. However, the simpler models obtained with multilinear regression were not significantly inferior. The highest value of the cross-validated coefficient of determination for both multilinear regression and random forest regression was obtained for $log\ K_{g[BMPyrr]+[FAP]-}$ models, which were 0.925 and 0.926, respectively. The Concordance Correlation Coefficient (CCC) of the constructed QSPR models is >0.85, which proves the predictive power of these models. In particular, the high predictive power is indicated by the fact that CCC > 0.93 for multilinear regression models and CCC > 0.94 for random forest regression models. The analysis of the selected features in the optimal models showed that the gas - ion liquid partition coefficient can be significantly influenced by the following properties of the gas molecule: number of atoms, presence of aromatic ring, presence of specific functional group, especially the alcohol group, polarity and amount of atoms heavier than carbon. The models of ionic liquids with the cation $[BMPyrr]^+$ differed in that the $log\ K_{g[BMPyrr]+[FAP]-}$ models also clearly took into account the polarity of the molecule, while the $log\ K_{g[BMPyrr]+[C(CN)3]-}$ models used the electrotopological state descriptor in place of polarity. The models of ionic liquids with the anion $[FAP]^-$ had molecular descriptors that seemed to take into account very similar chemical and physical properties, which may have been due to the fact that their cations differed in only one ether group. These results are a good support and basis for further research in gas-to-ionic liquid partition coefficient modeling, and in the following works one could try to construct cation-specific, anion-specific and even more general models. In addition, modelling could also be tried with other non-linear methods and with methods combining several models.

VIITED

[1] T. Welton, "Room-Temperature Ionic Liquids. Solvents for Synthesis and Catalysis.," *Chemical Reviews,* vol. 99, no. 8, pp. 2071-2084, 1999.

[2] J. S. Wilkes, J. A. Levisky, R. A. Wilson and C. L. Hussey, "Dialkylimidazolium chloroaluminate melts: a new class of room-temperature ionic liquids for electrochemistry, spectroscopy and synthesis," *Inorganic Chemistry,* vol. 21, no. 3, pp. 1263-1264, 1982.

[3] V. I. Pârvulescu and C. Hardacre, "Catalysis in Ionic Liquids," *Chemical Reviews,* vol. 107, no. 6, pp. 2615-2665, 2007.

[4] E. G. Yanes, S. R. Gratz, M. J. Baldwin, S. E. Robison and A. M. Stalcup, "Capillary electrophoretic application of 1-Alkyl-3-methylimidazolium-based ionic liquids," *Analytical Chemistry,* vol. 73, no. 16, pp. 3838-3844, 2001.

[5] W. Qin, H. Wei and S. F. Y. Li, "1,3-Dialkylimidazolium-based room-temperature ionic liquids as background electrolyte and coating material in aqueous capillary electrophoresis," *Journal of Chromatography A,* vol. 985, no. 1-2, pp. 447-454, 2003.

[6] M. C. Buzzeo, R. G. Evans and R. G. Compton, "Non-Haloaluminate Room-Temperature Ionic Liquids in Electrochemistry—A Review," *ChemPhysChem,* vol. 5, no. 8, 2004.

[7] J. G. Huddleston, H. D. Willauer, R. P. Swatloski, A. E. Visser and R. D. Rogers, "Room temperature ionic liquids as novel media for 'clean' liquid–liquid extraction," *Chemical Communications,* no. 16, pp. 1765-1766, 1998.

[8] S. Chun, S. V. Dzyuba and R. A. Bartsch, "Influence of structural variation in room-temperature ionic liquids on the selectivity and efficiency of competitive alkali metal salt extraction by a crown ether," *Analytical Chemistry,* vol. 73, no. 15, pp. 3737-3741, 2001.

[9] T.-F. Jiang, Y.-L. Gu, B. Liang, J.-B. Li, Y.-P. Shi and Q.-Y. Ou, "Dynamically coating the capillary with 1-alkyl-3-methylimidazolium-based ionic liquids for separation of basic proteins by capillary electrophoresis," *Analytica Chimica Acta,* vol. 479, no. 2, pp. 249-254, 2003.

[10] H. Qiu, S. Jiang, X. Liu and L. Zhao, "Novel imidazolium stationary phase for high-performance liquid chromatography," *Journal of Chromatography A,* vol. 1116, no. 1-2, pp. 46-50, 2006.

[11] Y. Liu, L. Shi, M. Wang, Z. Li, H. Liu and J. Li, "A novel room temperature ionic liquid sol–gel matrix for amperometric biosensor application," *Green Chemistry,* vol. 7, no. 9, pp. 655-658, 2005.

[12] J.-F. Liu, G.-B. Jiang, J.-F. Liu and J. Å. Jönsson, "Application of ionic liquids in analytical chemistry," *TrAC Trends in Analytical Chemistry,* vol. 24, no. 1, pp. 20-27, 2005.

[13] Z. Ma, J. Yu and S. Dai, "Preparation of Inorganic Materials Using Ionic Liquids," *Advanced Materials,* vol. 22, no. 2, 2009.

[14] D. J. G. Speight, "Chapter 6 - Introduction Into the Environment" in Environmental Organic Chemistry for Engineers, Butterworth-Heinemann, 2017, pp. 263-303.

[15] K.-S. Kim, J. W. Kang and S.-P. Kang, "Tuning ionic liquids for hydrate inhibition," *Chemical Communications,* vol. 47, no. 22, pp. 6341-6343, 2011.

[16] J. L. Anderson and K. D. Clark, "Ionic liquids as tunable materials in (bio)analytical chemistry," *Analytical and Bioanalytical Chemistry*, vol. 410, pp. 4565-4566, 2018.

[17] G. Gonfa, M. A. Bustam, A. M. Sharif, N. Mohamad and S. Ullah, "Tuning ionic liquids for natural gas dehydration using COSMO-RS methodology," *Journal of Natural Gas Science and Engineering*, vol. 27, no. 2, pp. 1141-1148, 2015.

[18] S. Khooshechin, Z. Dashtbozorgi, H. Golmohammadi and W. E. A. Jr., "QSPR prediction of gas-to-ionic liquid partition coefficient of organic solutes dissolved in 1-(2-hydroxyethyl)-1-methylimidazolium tris(pentafluoroethyl)trifluorophosphate using the replacement method and support vector regression," *Journal of Molecular Liquids*, vol. 196, pp. 43-51, 2014.

[19] T. e. a. Puzyn, "Quantitative Structure-Activity Relationships - Applications and Methodology," in *Recent Advances in QSAR Studies*, Springer Science+Business Media B.V., 2010, pp. 3-11.

[20] L. Chin Yee and Y. C. Wei, "Current Modeling Methods Used in QSAR/QSPR," in *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, vol. 2, Wiley-VCH Verlag GmbH & Co. KGaA., 2012, pp. 1-25.

[21] S. Abdolrahimi, B. Nasernejad and G. Pazuki, "Prediction of partition coefficients of alkaloids in ionic liquids based aqueous biphasic systems using hybrid group method of data handling (GMDH) neural network," *Journal of Molecular Liquids*, vol. 191, pp. 79-84, 2014.

[22] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," *Journal of Chemical Information and Modeling*, vol. 43, no. 6, pp. 1947-1958, 2003.

[23] A. L. Teixeira, J. P. Leal and A. O. Falcao, "Random forests for feature selection in QSPR Models - an application for predicting standard enthalpy of formation of hydrocarbons," *Journal of Cheminformatics*, vol. 5, no. 9, 2013.

[24] C.-H. Chen, K. Tanaka and K. Funatsu, "Random Forest Approach to QSPR Study of Fluorescence Properties Combining Quantum Chemical Descriptors and Solvent Conditions," *Journal of Fluorescence*, vol. 28, no. 2, pp. 695-706, 2018.

[25] D. R. MacFarlane, M. Kar and J. M. Pringle, Fundamentals of ionic liquids: From Chemistry to Applications, Wiley-VCH, 2017, pp. 2.5.7-9.

[26] K. Ghandi, "A Review of Ionic Liquids, Their Limits and Applications," *Green and Sustainable Chemistry*, vol. 4, pp. 44-53, 2014.

[27] S. K. Singh and A. W. Savoy, "Ionic liquids synthesis and applications: An overview," *Journal of Molecular Liquids*, pp. 112038-112060, 2020.

[28] F. Zhou, A. Izgorodin, R. K. Hocking, L. Spiccie and D. R. MacFarlane, "Electrodeposited MnOx Films from Ionic Liquid for Electrocatalytic Water Oxidation," *Advanced Energy Materials*, vol. 2, no. 8, pp. 1013-1021, 2012.

[29] M. Nikinmaa, "Chapter 2 - What Causes Aquatic Contamination?" in An Introduction to Aquatic Toxicology, Academic Press, 2014, pp. 19-39.

[30] F. Javed, F. Ullah, M. R. Zakaria and H. M. Akil, "An approach to classification and hi-tech applications of room-temperature ionic liquids (RTILs): A review," *Journal of Molecular Liquids*, vol. 271, pp. 403-420, 2018.

[31] A. Hospido and H. Rodriguez, "Life Cycle Assessment (LCA) of Ionic Liquids," in *Encyclopedia of Ionic Liquids*, Singapore, Springer Singapore, 2019, pp. 1-9.

[32] W. L. Armarego and C. Chai, Purification of Laboratory Chemicals, 7 ed., Butterworth-Heinemann, 2013, pp. 91-102.

[33] J. P. Hallett and T. Welton, "Room-Temperature Ionic Liquids: Solvents for Synthesis and Catalysis," *Chemical Reviews,* vol. 111, no. 5, pp. 3508-3576, 2011.

[34] J. G. Speight, "Chapter 9 - Molecular Interactions, Partitioning, and Thermodynamics" in Reaction Mechanisms in Environmental Engineering, Butterworth-Heinemann, 2018, pp. 307-336.

[35] E. Voutsas, "Chapter 11 - Estimation of the Volatilization of Organic Chemicals from Soil," in *Thermodynamics, Solubility and Environmental Issues*, Elsevier B.V., 2007, pp. 205-227.

[36] A. Voelkel, B. Strzemiecka, K. Adamska and K. Milczewska, "Inverse gas chromatography as a source of physiochemical data," *Journal of Chromatography,* vol. 1216, no. 10, pp. 1551-1566, 2009.

[37] A. Marciniak, "The Solubility Parameters of Ionic Liquids," *International Journal of Molecular Sciences,* vol. 11, no. 5, pp. 1973-1990, 2010.

[38] A. Tropsha, P. Gramatica and V. K. Gombar, "The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models," *QSAR & Combinatorial Science,* vol. 22, no. 1, pp. 69-77, 2003.

[39] M. Krein, T.-W. Huang, L. Morkowchuk, D. K. Agrafiotis and C. M. Breneman, "Developing Best Practices for Descriptor-Based Property Prediction: Appropriate Matching of Datasets, Descriptors, Methods, and Expectations," in *Statistical Modelling of Molecular Descriptors in QSAR/QSPR, Volume 2*, Wiley-Blackwell, 2012, pp. 33-56.

[40] A. Gobraikh and A. Tropsha, "Beware of q2!," *Journal of Molecular Graphics and Modelling,* vol. 20, no. 1, pp. 269-276, 2002.

[41] M. Shahlaei, "Descriptor Selection Methods in Quantitative Structure−Activity Relationship Studies: A Review Study," *Chemical Reviews,* vol. 113, no. 10, pp. 8093-8103, 2013.

[42] N. Chirico and P. Gramatica, "Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient," *Journal of Chemical Information and Modeling,* vol. 51, no. 9, pp. 2320-2335, 2011.

[43] H. Hond, S. Slavov, W. Ge, F. Qian, Z. Su, H. Fang, Y. Cheng, R. Perkins, L. Shi and W. Tong, "Mold2 Molecular Descriptors for QSAR," in *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, vol. 2, Wiley-Blackwell, 2012, pp. 65-105.

[44] Talete, "Dragon molecular descriptors," Talete SRL, [Online]. Available: http://www.talete.mi.it/products/dragon_molecular_descriptors.htm. [Accessed 25 aprill 2020].

[45] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins and W. Tong, "Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics," *Journal of Chemical Information and Modeling,* vol. 48, no. 7, pp. 1337-1344, 2008.

[46] Semichem, Inc., "Semichem, Codessa III, CODESSA TM Features," Semichem, [Online]. Available: http://www.semichem.com/codessa/cfeatures.php. [Accessed 25 aprill 2020].

[47] T. T. Cai and L. Wang, "Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise," *IEEE Transactions on Information Theory,* vol. 57, no. 7, pp. 4680-4688, 2011.

[48] R. Rubinstein, M. Zibulevsky and M. Elad, "Efficient Implementation of the K-SVD Algorithm using Batch Orthogonal Matching Pursuit," Technion - Israel Institute of Technology, Haifa, 2008.

[49] A. C. Rencher and G. B. Schaalje, "Multiple Regression: Estimation," in *Linear Models in Statistics*, John Wiley & Sons, Inc., 2007, pp. 137-174.

[50] J. D. Jobson, "Multiple Linear Regression," in *Applied Multivariate Data Analysis. Volume I: Regression and Experimental Design*, Springer-Verlag, 1991, pp. 219-251.

[51] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 1 10 2001.

[52] G. Louppe, "Random Forests," in *Understanding Random Forests*, 2014, pp. 55-115.

[53] G. Louppe, "Classification and Regression Trees," in *Understanding Random Forests*, University of Liége, 2014, pp. 25-52.

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[55] L. Breiman, "Bagging Predictors," University of California, Berkeley, 1994.

[56] M. H. Abraham, "Scales of Solute Hydrogen-bonding: Their Construction and Application to Physicochemical and Biochemical Processes," *Chemical Society Reviews*, vol. 22, no. 2, pp. 73-83, 1993.

[57] J. William E Acree and M. H. Abraham, "The analysis of solvation in ionic liquids and organic solvents using the Abraham linear free energy relationship," *Journal of Chemical Technology & Biotechnology*, vol. 81, no. 8, pp. 1441-1446, 2006.

[58] L. Sprunger, M. Clark, W. E. Acree Jr. and M. H. Abraham, "Characterization of Room-Temperature Ionic Liquids by the Abraham Model with Cation-Specific and Anion-Specific Equation Coefficients," *Journal of Chemical Information and Modelling*, vol. 47, no. 3, pp. 1123-1129, Apr 2007.

[59] L. M. Sprunger, A. Proctor, W. E. Acree Jr. and M. H. Abraham, "LFER correlations for room temperature ionic liquids: Separation of equation coefficients into individual cation-specific and anion-specific contributions," *Fluid Phase Equilibria*, vol. 265, no. 1-2, pp. 104-111, Mar 2008.

[60] L. M. Sprunger, J. Gibbs, A. Proctor, W. E. Acree Jr., M. H. Abraham, Y. Meng and J. L. Anderson, "Linear Free Energy Relationship Correlations for Room Temperature Ionic Liquids: Revised Cation-Specific and Anion-Specific Equation Coefficients for Predictive Applications Covering a Much Larger Area of Chemical Space," *Industrial & Engineering Chemistry Research*, vol. 48, no. 8, pp. 4145-4154, Mar 2009.

[61] A.-L. Revelli, F. Mutelet and J.-N. Jaubert, "Prediction of Partition Coefficients of Organic Compounds in Ionic Liquids: Use of a Linear Solvation Energy Relationship with Parameters Calculated through a Group Contribution Method," *Industrial & Engineering Chemistry Research*, vol. 49, no. 8, pp. 3883-3892, Apr 2010.

[62] L. Anderson, L. M. Grubbs, S. Ye, M. Saifullah, M. Cornelius, McMillan-Wiggins, W. E. A. Jr., M. H. Abraham, P. Twu and J. Anderson, "Correlations for describing gas-to-ionic liquid partitioning at 323 K based on ion-specific equation coefficient and group contribution versions of the Abraham model," *Fluid Phase Equilibria*, vol. 301, no. 2, pp. 257-266, 2011.

[63] T. W. Stephens, V. Chou, A. N. Quay, C. Shen, N. Dabadge, A. Tian, M. Loera, B. Willis, A. Wilson, W. E. Acree Jr., P. Twu, J. L. Anderson and M. H. Abraham, "Thermochemical investigations of solute transfer into ionic liquid solvents: updated

Abraham model equation coefficients for solute activity coefficient and partition coefficient predictions," *Physics and Chemistry of Liquids,* vol. 52, no. 4, pp. 488-518, 2014.

[64] T. W. Stephens, E. Hart, N. Kuprasertkul, S. Mehta, A. Wadawadigi, W. E. Acree Jr. and M. H. Abraham, "Abraham model correlations for describing solute transfer into ionic liquid solvents: calculation of ion-specific equation coefficients for the 4,5-dicyano-2-(trifluoromethyl)imidazolide anion," *Physics and Chemistry of Liquids,* vol. 52, no. 6, pp. 777-791, Nov/Dec 2014.

[65] F. Mutelet, H. Djebouri, G. A. Baker, S. Ravula, B. Jiang, X. Tong, D. Woods and W. E. Acree Jr., "Study of benzyl- or cyclohexyl-functionalized ionic liquids using inverse gas chromatography," *Journal of Molecular Liquids,* vol. 242, pp. 550-559, Jul 2017.

[66] B. Jiang, M. Y. Horton, W. E. A. Jr. and M. H. Abraham, "Ion-specific equation coefficient version of the Abraham model for ionic liquid solvents: determination of coefficients for tributylethylphosphonium, 1-butyl-1-methylmorpholinium, 1-allyl-3-methylimidazolium and octyltriethylammonium cations," *Physics and Chemistry of Liquids,* vol. 55, no. 3, pp. 358-385, 2017.

[67] W. E. A. Jr. and B. Jiang, "Abraham model correlations for ionic liquid solvents: computational methodology for updating existing ion-specific equation coefficients," *Physics and Chemistry of Liquids,* vol. 55, no. 4, pp. 457-462, 2017.

[68] F. Mutelet, S. Ravula, G. A. Baker, D. Woods, X. Tong and W. E. Acree Jr., "Infinite Dilution Activity Coefficients and Gas-to-Liquid Partition Coefficients of Organic Solutes Dissolved in 1-Benzylpyridinium Bis(Trifluoromethylsulfonyl)Imide and 1-Cyclohexylmethyl-1-Methylpyrrolidinium Bis(Trifluoromethylsulfonyl)Imide," *Journal of Solution Chemistry,* vol. 47, pp. 308-335, Feb 2018.

[69] F. Mutelet, G. A. Baker, S. Ravula, E. Qian, L. Wang and W. E. A. Jr, "Infinite dilution activity coefficients and gas-to-liquid partition coefficients of organic solutes dissolved in 1-sec-butyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide and in 1-tert-butyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide," *Physics and Chemistry of Liquids,* vol. 57, no. 4, pp. 453-472, 2018.

[70] D. Yue, W. E. Acree Jr. and M. H. Abraham, "Development of Abraham model IL-specific correlations for N-triethyl(octyl)ammonium correlations for N-triethyl(octyl)ammonium methylpyrrolidinium bis(fluorosulfonyl)imide," *Physics and Chemistry of Liquids,* vol. 57, no. 6, pp. 733-745, 2019.

[71] F. Mutelet, C. Hussard, G. A. Baker, H. Zhao, B. Churchill and W. E. Acree Jr., "Characterization of the solubilizing ability of short-chained glycol-grafted ammonium and phosphonium ionic liquids," *Journal of Molecular Liquids,* vol. 304, p. 112786, Feb 2020.

[72] H. Moriwaki, Y.-S. Tian, N. Kawashite and T. Takagi, "Mordred: a molecular descriptor calculator," *Journal of Cheminformatics,* vol. 10, no. 4, 2018.

[73] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences,* vol. 28, no. 1, pp. 31-36, 1988.

[74] V. M. Alves, A. Golbraikh, S. J. Capuzzi, K. Liu, W. I. Lam, D. R. Korn, D. Pozefsky, C. H. Andrade and E. N. T. A. Muratov, "Multi-Descriptor Read Across (MuDRA): A Simple and Transparent Approach for Developing Accurate Quantitative Structure–Activity Relationship Models," *Journal of Chemical Information and Modelling,* vol. 58, no. 6, pp. 1214-1223, 2018.

[75] R. B. Patil, E. G. Barbosa, J. N. Sangshetti, V. P. Zambre and S. D. Sawant, "Structural insights of dipeptidyl peptidase-IV inhibitors through molecular dynamics-guided receptor-dependent 4D-QSAR studies," *Molecular Diversity*, vol. 22, no. 1, pp. 575-583, 2018.

[76] T. M. Oshiro, P. S. Perez and J. A. Baranauskas, "How Many Trees in a Random Forest?," in *Lecture Notes in Computer Science*, Sao Paulo, 2012.

[77] C. Strobl, A.-L. Boulesteix, A. Zeileis and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics,* vol. 8, no. 25, 2007.

[78] T. Parr, K. Turgutlu, C. Csiszar and J. Howard, "Beware Default Random Forest Importances," explained.ai, 26 märts 2018. [Online]. Available: https://explained.ai/rf-importance/#corr_collinear. [Accessed 26 aprill 2020].

[79] L. H. Hall and L. B. Kier, "Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information," *Journal of Chemical Information and Computer Sciences,* vol. 35, no. 6, pp. 1039-1045, 1995.

[80] H. Moriwaki, "Mordred 1.2.1a1 documentation. Descriptor List," 2016. [Online]. Available: https://mordred-descriptor.github.io/documentation/master/descriptors.html. [Accessed 16 mai 2020].

[81] V. Consonni and R. Todeschini, Handbook of Molecular Descriptors, vol. 11, WILEY-VCH Verlag GmbH, 2000.

LISAD

Lisa 1. Tunnuse *isPolar* väärtused töösse kaasatud ühenditel.

1,2-dichlorobenzene	1	**2-pentanone**	1	**ethyl phenyl ether**	1
1,2-dimethylbenzene	0	**2-propanol**	1	**ethylbenzene**	0
1,3-dimethylbenzene	0	**3-methylpentane**	0	**heptane**	0
1,4-dimethylbenzene	0	**3-pentanone**	1	**hexane**	0
1,4-dioxane	1	**acetic acid**	1	**methanol**	1
1-bromohexane	1	**acetonitrile**	1	**methyl acetate**	1
1-bromooctane	1	**acetophenone**	1	**methyl butanoate**	1
1-butanol	1	**alpha-methylstyrene**	0	**methyl hexanoate**	1
1-chlorobutane	1	**aniline**	1	**methyl propanoate**	1
1-chlorohexane	1	**benzaldehyde**	1	**methyl tert-pentyl ether**	1
1-chlorooctane	1	**benzene**	0	**methylcyclohexane**	0
1-decene	0	**benzonitrile**	1	**n,n-dimethylformamide**	1
1-heptene	0	**benzyl alcohol**	1	**naphthalene**	0
1-heptyne	0	**bromoethane**	1	**nitrobenzene**	1
1-hexene	0	**butanal**	1	**nitrous oxide**	1
1-hexyne	0	**carbon dioxide**	0	**nonane**	0
1-iodobutane	1	**cycloheptane**	0	**octanal**	1
1-nitropropane	1	**cyclohexane**	0	**octane**	0
1-octanol	1	**cyclohexanol**	1	**p-cresol**	1
1-octene	0	**cyclohexanone**	1	**pentane**	0
1-octyne	0	**cyclohexene**	0	**phenol**	1
1-pentanol	1	**cyclooctane**	0	**propanone**	1
1-pentene	0	**cyclopentane**	0	**propionic acid**	1
1-pentyne	0	**decane**	0	**propionitrile**	1
1-propanol	1	**decyl alcohol**	1	**pyridine**	1
2,2,4-trimethylpentane	0	**dibutyl ether**	1	**pyrrole**	1
2,2-dimethylbutane	0	**diethyl ether**	1	**styrene**	0
2-butanol	1	**diisopropyl ether**	1	**tert-butyl ethyl ether**	1
2-chloroaniline	1	**dipropyl ether**	1	**tert-butyl methyl ether**	1
2-methyl-1-propanol	1	**ethane**	0	**tetrahydrofuran**	1
2-methyl-2-propanol	1	**ethanol**	1	**thiophene**	1
2-nitrophenol	1	**ethyl acetate**	1	**toluene**	0

Lisa 2. Optimaalsete multilineaarse regressiooni ja juhumetsa regresiooni mudelite LR1 ja RF1 ennustatud [BMPyrr]$^{+}$[FAP]$^{-}$ log K_{giv} väärtuste ennustusvead.

	Multilineaarne regressioon					Juhumetsa regressioon				
	1	2	3	4	5	1	2	3	4	5
ethane	0.4553	0.3017	0.3168	0.3073	0.3449	-0.7210	0.0000	0.0000	0.0000	0.0000
carbon dioxide	-0.4677	-0.5371	-0.5101	-0.5308	-0.4918	0.0000	0.0102	0.0000	0.0000	0.0000
nitrous oxide	0.0610	0.0707	-0.0973	0.1118	-0.0890	0.0000	0.0000	-0.3182	0.0000	0.0000
pentane	0.1961	0.0982	0.1079	0.1039	0.1346	0.0000	0.0000	0.0000	-0.4132	0.0000
2,2-dimethylbutane	0.0746	-0.0155	-0.0096	-0.0100	0.0182	0.0000	0.0000	0.0000	0.0000	-0.2030
3-methylpentane	0.2076	0.1180	0.1228	0.1235	0.1510	0.0402	0.0000	0.0000	0.0000	0.0000
hexane	0.2300	0.1407	0.1448	0.1461	0.1733	0.0000	0.0048	0.0000	0.0000	0.0000
1-hexene	0.0100	-0.0513	-0.0398	-0.0459	-0.0151	0.0000	0.0000	-0.0398	0.0000	0.0000
cyclohexane	-0.1777	-0.2243	-0.2178	-0.2184	-0.1935	0.0000	0.0000	0.0000	-0.5099	0.0000
heptane	0.2076	0.1258	0.1231	0.1309	0.1538	0.0000	0.0000	0.0000	0.0000	-0.0991
diethyl ether	-0.5483	-0.5212	-0.5312	-0.5073	-0.5211	-0.6893	0.0000	0.0000	0.0000	0.0000
2,2,4-trimethylpentane	-0.0191	-0.0956	-0.1027	-0.0907	-0.0705	0.0000	-0.1051	0.0000	0.0000	0.0000
methylcyclohexane	-0.2258	-0.2711	-0.2714	-0.2655	-0.2442	0.0000	0.0000	-0.5270	0.0000	0.0000
1-heptene	0.0611	0.0034	0.0071	0.0084	0.0346	0.0000	0.0000	0.0000	-0.0726	0.0000
diisopropyl ether	-0.8089	-0.7803	-0.7960	-0.7680	-0.7811	0.0000	0.0000	0.0000	0.0000	-0.0690
methanol	-0.0012	-0.0209	-0.0352	0.0010	0.0394	-0.3440	0.0000	0.0000	0.0000	0.0000
cyclohexene	-0.3459	-0.3640	-0.3477	-0.3585	-0.3272	0.0000	-0.5672	0.0000	0.0000	0.0000
tert-butyl ethyl ether	-0.8501	-0.8202	-0.8387	-0.8080	-0.8228	0.0000	0.0000	0.0687	0.0000	0.0000
1-hexyne	0.0338	0.0010	0.0223	0.0059	0.0432	0.0000	0.0000	0.0000	-0.3601	0.0000
octane	0.1596	0.0847	0.0739	0.0895	0.1075	0.0000	0.0000	0.0000	0.0000	0.2225
tert-butyl methyl ether	-0.3759	-0.3484	-0.3598	-0.3353	-0.3477	-0.5002	0.0000	0.0000	0.0000	0.0000
cycloheptane	0.0540	0.0089	0.0083	0.0144	0.0356	0.0000	0.0939	0.0000	0.0000	0.0000
ethanol	-0.1585	-0.1696	-0.1708	-0.1497	-0.0906	0.0000	0.0000	0.1469	0.0000	0.0000
dipropyl ether	-0.6077	-0.5779	-0.5964	-0.5657	-0.5805	0.0000	0.0000	0.0000	-0.1878	0.0000
1-octene	0.0568	0.0030	-0.0021	0.0077	0.0287	0.0000	0.0000	0.0000	0.0000	0.4042
2-propanol	-0.3817	-0.3889	-0.3856	-0.3701	-0.2997	-0.2460	0.0000	0.0000	0.0000	0.0000
nonane	0.0792	0.0110	-0.0089	0.0154	0.0280	0.0000	0.0354	0.0000	0.0000	0.0000
methyl tert-pentyl ether	-0.3350	-0.3060	-0.3226	-0.2938	-0.3074	0.0000	0.0000	0.3775	0.0000	0.0000
1-heptyne	0.2565	0.2231	0.2351	0.2279	0.2593	0.0000	0.0000	0.0000	-0.1284	0.0000
1-propanol	-0.1349	-0.1419	-0.1393	-0.1232	-0.0549	0.0000	0.0000	0.0000	0.0000	0.0648
cyclooctane	0.2136	0.1705	0.1624	0.1756	0.1927	-0.2009	0.0000	0.0000	0.0000	0.0000
methyl acetate	0.0528	0.1096	0.0990	0.1249	0.1048	0.0000	-0.4221	0.0000	0.0000	0.0000
2-butanol	-0.3500	-0.3546	-0.3516	-0.3366	-0.2608	0.0000	0.0000	0.1166	0.0000	0.0000
decane	-0.0336	-0.0951	-0.1250	-0.0911	-0.0846	0.0000	0.0000	0.0000	0.0595	0.0000
tetrahydrofuran	0.0272	0.0723	0.0689	0.0849	0.0783	0.0000	0.0000	0.0000	0.0000	-0.5811
1-octyne	0.2124	0.1799	0.1819	0.1843	0.2098	0.3002	0.0000	0.0000	0.0000	0.0000
2-methyl-1-propanol	-0.2309	-0.2354	-0.2327	-0.2176	-0.1440	0.0000	0.1056	0.0000	0.0000	0.0000
propanone	0.0020	0.0853	0.0862	0.1011	0.0860	0.0000	0.0000	-0.3038	0.0000	0.0000
benzene	-0.4845	-0.4431	-0.3950	-0.4403	-0.3825	0.0000	0.0000	0.0000	-0.6236	0.0000
dibutyl ether	-0.6584	-0.6245	-0.6578	-0.6138	-0.6346	0.0000	0.0000	0.0000	0.0000	-0.3445
thiophene	-0.5776	-0.5265	-0.4455	-0.5273	-0.4372	-0.3719	0.0000	0.0000	0.0000	0.0000
butanal	-0.1379	-0.0553	-0.0581	-0.0403	-0.0555	0.0000	-0.4141	0.0000	0.0000	0.0000
ethyl acetate	0.0044	0.0668	0.0530	0.0814	0.0610	0.0000	0.0000	0.0735	0.0000	0.0000
1-butanol	-0.0261	-0.0303	-0.0284	-0.0126	0.0599	0.0000	0.0000	0.0000	-0.0890	0.0000
1-decene	-0.0340	-0.0787	-0.1039	-0.0748	-0.0657	0.0000	0.0000	0.0000	0.0000	0.4748
methyl propanoate	0.0427	0.1050	0.0915	0.1196	0.0994	0.0525	0.0000	0.0000	0.0000	0.0000
1-chlorohexane	0.2261	0.2457	0.2368	0.2559	0.2540	0.0000	-0.2156	0.0000	0.0000	0.0000
acetonitrile	0.5709	0.6508	0.6437	0.6687	0.6422	0.0000	0.0000	0.1846	0.0000	0.0000
toluene	-0.1108	-0.0813	-0.0487	-0.0780	-0.0317	0.0000	0.0000	0.0000	-0.4720	0.0000
1-bromohexane	0.6247	0.6296	0.6293	0.6379	0.6485	0.0000	0.0000	0.0000	0.0000	0.2192
methyl butanoate	-0.0191	0.0476	0.0282	0.0615	0.0392	0.3755	0.0000	0.0000	0.0000	0.0000
acetic acid	0.3342	0.3580	0.3573	0.3781	0.4332	0.0000	0.3936	0.0000	0.0000	0.0000
1-pentanol	0.0336	0.0318	0.0305	0.0488	0.1226	0.0000	0.0000	0.4852	0.0000	0.0000
1,4-dioxane	0.5194	0.5647	0.5483	0.5781	0.5605	0.0000	0.0000	0.0000	0.3575	0.0000
3-pentanone	0.2097	0.2910	0.2839	0.3053	0.2893	0.0000	0.0000	0.0000	0.0000	-0.0013
2-pentanone	0.2012	0.2826	0.2753	0.2969	0.2808	0.0013	0.0000	0.0000	0.0000	0.0000
ethylbenzene	-0.0295	-0.0067	0.0106	-0.0034	0.0327	0.0000	0.0948	0.0000	0.0000	0.0000
1,4-dimethylbenzene	0.1226	0.1443	0.1641	0.1477	0.1853	0.0000	-0.0169	-0.0337	-0.0169	-0.0169
pyridine	0.1353	0.2025	0.2225	0.2109	0.2316	0.0000	0.0000	0.0000	0.2113	0.0000
propionic acid	0.1712	0.2070	0.2070	0.2265	0.2866	0.0000	0.0000	0.0000	0.0000	0.4398
1,3-dimethylbenzene	0.1518	0.1736	0.1933	0.1770	0.2145	0.0337	0.0169	0.0000	0.0169	0.0169
1-chlorooctane	0.2785	0.2948	0.2698	0.3034	0.2954	0.0000	-0.2091	0.0000	0.0000	0.0000
1,2-dimethylbenzene	0.2448	0.2666	0.2860	0.2701	0.3074	0.0000	0.0000	0.0706	0.0000	0.0000
1-nitropropane	0.1794	0.3084	0.2618	0.3309	0.2611	0.0000	0.0000	0.0000	0.5696	0.0000
1-bromooctane	0.6489	0.6513	0.6332	0.6583	0.6606	0.0000	0.0000	0.0000	0.0000	0.2844
methyl hexanoate	-0.3025	-0.2272	-0.2653	-0.2148	-0.2461	-0.2499	0.0000	0.0000	0.0000	0.0000
styrene	-0.0451	0.0005	0.0323	0.0024	0.0513	0.0000	0.1138	0.0000	0.0000	0.0000
alpha-methylstyrene	0.1011	0.1388	0.1562	0.1411	0.1796	0.0000	0.0000	0.2932	0.0000	0.0000
octanal	0.0080	0.0895	0.0589	0.1017	0.0758	0.0000	0.0000	0.0000	0.1879	0.0000
cyclohexanone	0.3024	0.3974	0.3900	0.4097	0.3985	0.0000	0.0000	0.0000	0.0000	-0.0393
ethyl phenyl ether	-0.2267	-0.1604	-0.1668	-0.1544	-0.1448	-0.1401	0.0000	0.0000	0.0000	0.0000
1-octanol	-0.1007	-0.0941	-0.1155	-0.0791	-0.0112	0.0000	0.2086	0.0000	0.0000	0.0000
benzaldehyde	-0.3245	-0.2073	-0.1822	-0.2014	-0.1727	0.0000	0.0000	-0.1467	0.0000	0.0000
benzonitrile	-0.1423	-0.0289	-0.0022	-0.0235	0.0078	0.0000	0.0000	0.0000	0.1871	0.0000
phenol	0.1868	0.2176	0.2580	0.2271	0.3542	0.0000	0.0000	0.0000	0.0000	0.3171
nitrobenzene	-0.4431	-0.2849	-0.3018	-0.2722	-0.2944	0.0959	0.0000	0.0000	0.0000	0.0000
decyl alcohol	-0.2524	-0.2387	-0.2796	-0.2249	-0.1661	0.0000	0.4896	0.0000	0.0000	0.0000
benzyl alcohol	0.1932	0.2265	0.2503	0.2367	0.3495	0.0000	0.0000	0.1041	0.0000	0.0000
acetophenone	0.0995	0.2142	0.2236	0.2209	0.2367	0.0000	0.0000	0.0000	0.5155	0.0000
p-cresol	0.3354	0.3675	0.3941	0.3780	0.4943	0.0000	0.0000	0.0000	0.0000	0.0969
naphthalene	-0.1133	-0.0326	-0.0111	-0.0330	0.0137	0.6468	0.0000	0.0000	0.0000	0.0000

TREENING

TEST

Lisa 3. Optimaalsete multilineaarse regresiooni ja juhumetsa regresiooni mudelite LR2 ja RF2 ennustatud [BMPyrr]$^+$[C(CN)$_3$]$^-$ log K_{giv} väärtuste ennustusvead.

	Multilineaarne regressioon					Juhumetsa regressioon				
	1	2	3	4	5	1	2	3	4	5
pentane	0.1436	0.1618	0.1218	0.1822	0.1550	-0.6637	0.0000	0.0000	0.0000	0.0000
2,2-dimethylbutane	0.0622	0.1072	0.0337	0.1481	0.1403	-0.1573	-0.2360	-0.1573	-0.1295	-0.1065
1-pentene	-0.2371	-0.2079	-0.2300	-0.1843	-0.2035	0.0000	0.0000	-0.2236	0.0000	0.0000
3-methylpentane	0.2909	0.3306	0.2647	0.3673	0.3558	0.0557	-0.0230	0.0557	0.0835	0.1065
hexane	0.0911	0.1128	0.0651	0.1150	0.1069	0.1017	0.0230	0.1017	0.1295	0.1525
cyclopentane	-0.2372	-0.2467	-0.1973	-0.2622	-0.3052	-0.2156	0.0000	0.0000	0.0000	0.0000
2,2,4-trimethylpentane	0.0210	0.0898	-0.0210	0.1248	0.1537	-0.2240	-0.4480	-0.2240	0.0000	-0.2240
1-hexene	-0.2036	-0.1708	-0.2070	-0.1665	-0.1650	0.0000	0.0000	-0.1829	0.0000	0.0000
heptane	0.0406	0.0653	0.0086	0.0477	0.0595	0.0000	0.0000	0.0000	0.0930	0.0000
diethyl ether	-0.1527	-0.1110	-0.1097	-0.0631	-0.0898	0.0000	0.0000	0.0000	0.0000	-0.2993
diisopropyl ether	-0.3903	-0.3204	-0.3802	-0.2768	-0.2599	-0.0820	0.0000	0.0000	0.0000	0.0000
cyclohexane	-0.2522	-0.2603	-0.2264	-0.3005	-0.3209	0.0000	0.2649	0.0000	0.0000	0.0000
tert-butyl ethyl ether	-0.3199	-0.2493	-0.3102	-0.2060	-0.1880	0.0000	0.0000	0.0820	0.0000	0.0000
1-heptene	-0.2097	-0.1739	-0.2238	-0.1903	-0.1678	0.0000	0.0000	0.0000	-0.4076	0.0000
bromoethane	-0.5987	-0.6060	-0.3532	-0.5115	-0.6626	0.0000	0.0000	0.0000	0.0000	-0.5767
1-pentyne	-0.1399	-0.0996	-0.1039	-0.0728	-0.0841	0.1818	0.0000	0.0000	0.0000	0.0000
methylcyclohexane	-0.0568	-0.0382	-0.0434	-0.0517	-0.0568	0.0000	-0.3850	-0.1925	-0.1925	-0.1925
tert-butyl methyl ether	-0.0332	0.0267	-0.0098	0.0740	0.0736	0.0000	0.0000	-0.2144	0.0000	0.0000
octane	-0.0106	0.0168	-0.0498	-0.0218	0.0104	0.2240	0.0000	0.2240	0.4480	0.2240
dipropyl ether	-0.1618	-0.1099	-0.1490	-0.0948	-0.0793	0.0000	0.0000	0.0000	0.0000	0.2194
cyclohexene	-0.3735	-0.3742	-0.3220	-0.4111	-0.4279	-0.6300	-0.3150	-0.3150	-0.3150	0.0000
1-octene	-0.2280	-0.1897	-0.2528	-0.2276	-0.1839	0.0000	-0.2536	0.0000	0.0000	0.0000
nonane	-0.0700	-0.0403	-0.1173	-0.1006	-0.0477	0.0000	0.0000	-0.0638	0.0000	0.0000
1-hexyne	-0.0392	0.0046	-0.0202	0.0110	0.0220	0.0000	0.0000	0.0000	0.0852	0.0000
methyl tert-pentyl ether	0.1651	0.2357	0.1748	0.2790	0.2970	0.0000	0.0000	0.0000	0.0000	0.4582
cycloheptane	-0.0868	-0.0953	-0.0743	-0.1637	-0.1611	0.3850	0.0000	0.1925	0.1925	0.1925
1-chlorobutane	-0.2360	-0.2385	-0.0376	-0.1782	-0.2910	0.0000	-0.5466	0.0000	0.0000	0.0000
methyl acetate	-0.9953	-0.9102	-0.9024	-0.8551	-0.8383	0.0000	0.0000	-0.4531	0.0000	0.0000
decane	-0.1230	-0.0912	-0.1789	-0.1737	-0.1000	0.0000	0.0000	0.0000	0.2722	0.0000
1-heptyne	0.0171	0.0638	0.0209	0.0488	0.0819	0.0000	0.0000	0.0000	0.0000	0.2970
ethyl acetate	-0.8506	-0.7549	-0.7827	-0.7056	-0.6676	-0.3030	0.0000	0.0000	0.0000	0.0000
propanone	0.0938	0.1537	0.1742	0.2105	0.1953	0.0000	-0.1023	0.0000	0.0000	0.0000
tetrahydrofuran	0.2044	0.2225	0.3012	0.2428	0.2003	0.0000	0.0000	0.0446	0.0000	0.0000
dibutyl ether	-0.1179	-0.0590	-0.1320	-0.0837	-0.0253	0.0000	0.0000	0.0000	-0.6063	0.0000
cyclooctane	0.0165	0.0087	0.0163	-0.0857	-0.0606	0.0000	0.3150	0.3150	0.3150	0.6300
methyl propanoate	-0.7166	-0.6209	-0.6488	-0.5716	-0.5337	-0.1845	0.0000	0.0000	0.0000	0.0000
methanol	-0.1397	-0.1946	-0.1023	-0.1107	-0.2004	0.0000	0.3225	0.0000	0.0000	0.0000
butanal	0.0389	0.0974	0.0944	0.1246	0.1336	0.0000	0.0000	-0.2367	0.0000	0.0000
benzene	-0.6801	-0.6659	-0.5772	-0.6965	-0.7058	0.0000	0.0000	0.0000	-0.3663	0.0000
1-decene	-0.2863	-0.2437	-0.3326	-0.3262	-0.2399	0.0000	0.0000	0.0000	0.0000	-0.0644
1-octyne	0.0345	0.0838	0.0242	0.0467	0.1018	0.3040	0.0000	0.0000	0.0000	0.0000
ethanol	-0.0637	-0.1081	-0.0455	-0.0289	-0.0990	0.0000	0.5921	0.0000	0.0000	0.0000
2-methyl-2-propanol	-0.4613	-0.4790	-0.4771	-0.4096	-0.4338	0.0000	0.0000	0.3085	0.0000	0.0000
2-propanol	-0.2050	-0.2363	-0.2044	-0.1621	-0.2095	0.0000	0.0000	0.0000	0.2169	0.0000
1-chlorohexane	-0.0301	-0.0273	0.1309	-0.0094	-0.0779	0.0000	0.0000	0.0000	0.0000	-0.2341
thiophene	-0.0145	-0.0055	0.1588	0.0289	-0.0460	-1.3342	0.0000	0.0000	0.0000	0.0000
methyl butanoate	-0.5738	-0.4721	-0.5283	-0.4382	-0.3778	0.0000	-0.3256	0.0000	0.0000	0.0000
1-iodobutane	0.1652	0.1601	0.3979	0.2266	0.1015	0.0000	0.0000	0.1561	0.0000	0.0000
acetonitrile	0.4880	0.5328	0.6115	0.5990	0.5517	0.0000	0.0000	0.0000	0.3491	0.0000
2-pentanone	0.4411	0.5175	0.4759	0.5535	0.5814	0.0000	0.0000	0.0000	0.0000	0.2004
toluene	-0.0981	-0.0585	-0.0203	-0.0645	-0.0569	0.3546	0.0000	0.0000	0.0000	0.0000
3-pentanone	0.5252	0.6040	0.5604	0.6450	0.6719	0.0000	0.1384	0.0000	0.0000	0.0000
propionitrile	0.7279	0.7785	0.8144	0.8270	0.8057	0.0000	0.0000	0.3461	0.0000	0.0000
1-propanol	0.0874	0.0492	0.0895	0.1140	0.0648	0.0000	0.0000	0.0000	0.0986	0.0000
2-butanol	-0.0295	-0.0502	-0.0426	0.0201	-0.0090	0.0000	0.0000	0.0000	0.0000	-0.1530
1-bromohexane	-0.3073	-0.2991	-0.1064	-0.2754	-0.3467	0.3534	0.0000	0.0000	0.0000	0.0000
1,4-dioxane	0.1903	0.2380	0.2966	0.2655	0.2539	0.0000	-0.4167	0.0000	0.0000	0.0000
ethylbenzene	-0.0320	0.0108	0.0232	-0.0172	0.0142	-0.1171	-0.1122	-0.1496	-0.1122	-0.0488
2-methyl-1-propanol	0.1415	0.1208	0.1284	0.1911	0.1620	0.0000	0.0000	0.0000	0.1710	0.0000
1,4-dimethylbenzene	0.4055	0.4695	0.4622	0.4862	0.5101	-0.0298	-0.0249	-0.0623	-0.0249	0.0385
1,3-dimethylbenzene	0.3965	0.4594	0.4530	0.4737	0.4981	-0.0194	-0.0145	-0.0519	-0.0145	0.0488
1-chlorooctane	0.0908	0.0978	0.2169	0.0699	0.0464	0.0000	-0.2675	0.0000	0.0000	0.0000
1-butanol	0.1880	0.1541	0.1757	0.2010	0.1731	0.0000	0.0000	0.1092	0.0000	0.0000
pyridine	0.0297	0.0650	0.1528	0.0672	0.0571	0.0000	0.0000	0.0000	-0.1800	0.0000
1,2-dimethylbenzene	0.5262	0.5872	0.5825	0.5972	0.6224	0.1468	0.1517	0.1143	0.1517	0.2151
methyl hexanoate	-0.3392	-0.2296	-0.3323	-0.2352	-0.1295	0.3971	0.0000	0.0000	0.0000	0.0000
1-nitropropane	-0.4237	-0.3085	-0.3515	-0.2722	-0.2000	0.0000	0.0062	0.0000	0.0000	0.0000
styrene	-0.0300	0.0237	0.0396	-0.0035	0.0393	0.0000	0.0000	-0.2220	0.0000	0.0000
1-pentanol	0.2343	0.2040	0.2086	0.2310	0.2246	0.0000	0.0000	0.0000	0.2474	0.0000
1-bromooctane	-0.2820	-0.2692	-0.1063	-0.2897	-0.3180	0.0000	0.0000	0.0000	0.0000	0.3257
alpha-methylstyrene	0.2426	0.3155	0.2904	0.2995	0.3611	0.1287	0.0000	0.0000	0.0000	0.0000
octanal	0.3657	0.4353	0.3488	0.3740	0.4745	0.0000	0.1661	0.0000	0.0000	0.0000
1,2-dichlorobenzene	0.6588	0.6715	0.8763	0.7157	0.6323	0.0000	0.0000	0.0069	0.0000	0.0000
cyclohexanone	0.7006	0.7580	0.7582	0.7460	0.7835	0.0000	0.0000	0.0000	0.3280	0.0000
ethyl phenyl ether	0.1403	0.2118	0.1995	0.1886	0.2525	0.0000	0.0000	0.0000	0.0000	0.3139
cyclohexanol	0.0781	0.0445	0.0805	0.0540	0.0532	-0.1811	0.0000	0.0000	0.0000	0.0000
acetic acid	-0.3457	-0.3380	-0.2833	-0.2594	-0.2685	0.0000	0.4107	0.0000	0.0000	0.0000
pyrrole	0.3138	0.2583	0.4076	0.3062	0.2361	0.0000	0.0000	0.2270	0.0000	0.0000
benzaldehyde	-0.1348	-0.0522	-0.0523	-0.0816	-0.0019	0.0000	0.0000	0.0000	0.2906	0.0000
benzonitrile	0.0217	0.1007	0.1116	0.0731	0.1460	0.0000	0.0000	0.0000	0.0000	0.1370
propionic acid	-0.0834	-0.0653	-0.0457	0.0077	0.0196	0.6044	0.0000	0.0000	0.0000	0.0000
1-octanol	0.2799	0.2574	0.2170	0.2194	0.2777	0.0000	0.7391	0.0000	0.0000	0.0000

TREENING

TEST

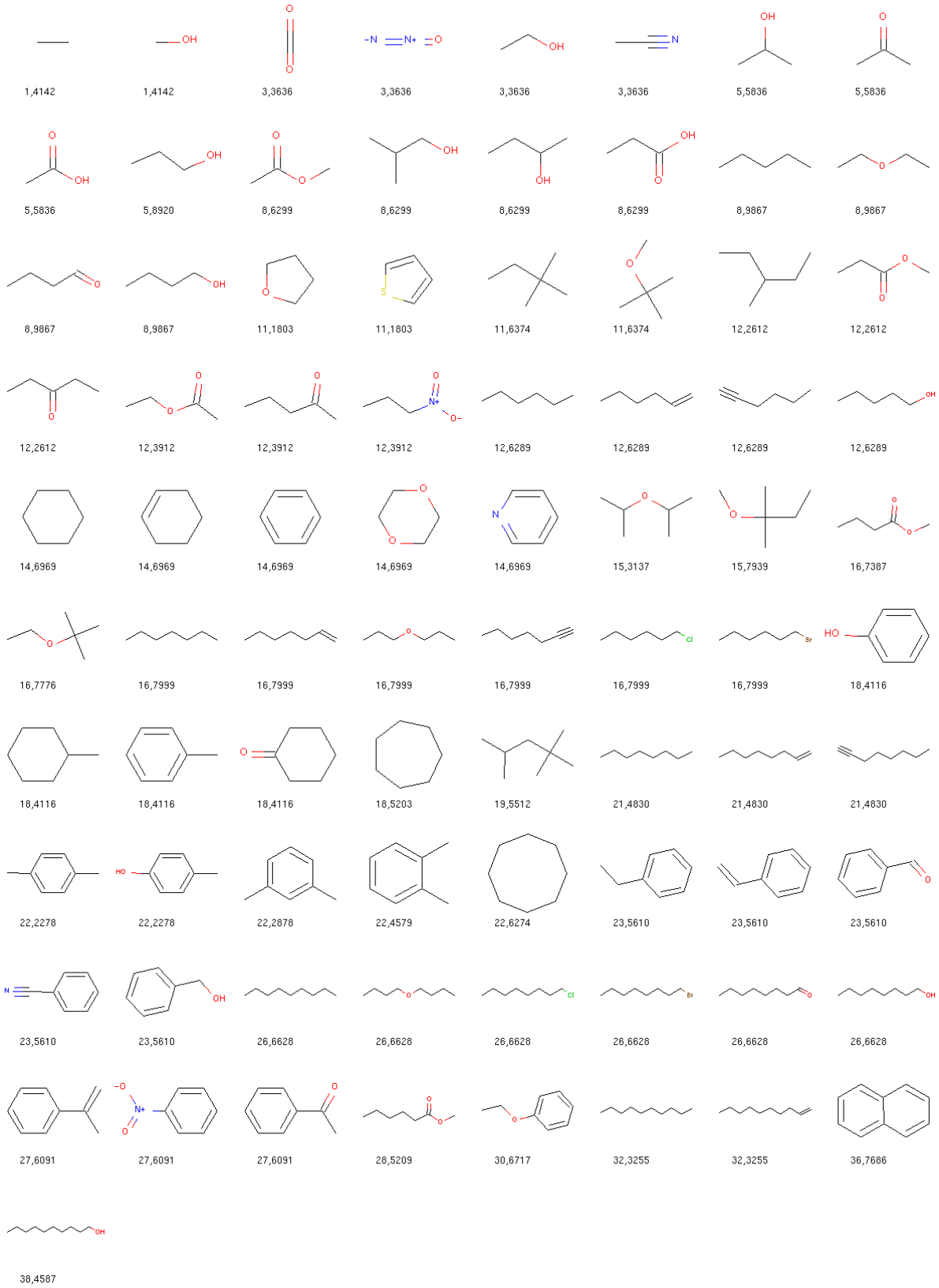
Lisa 4. Optimaalsete multilineaarse regresiooni ja juhumetsa regresiooni mudelite LR3 ja RF3 ennustatud [MeoeMPyrr]$^+$[FAP]$^-$ $\log K_{giv}$ väärtuste ennustusvead.

	Multilineaarne regressioon					Juhumetsa regressioon				
	1	2	3	4	5	1	2	3	4	5
pentane	0.2284	0.1168	0.2373	0.1690	0.1947	-0.2572	0.0000	0.0000	-0.0009	-0.0009
2,2-dimethylbutane	-0.5430	-0.6259	-0.5503	-0.5963	-0.5721	-0.0052	-0.1573	0.0000	-0.0036	-0.0027
1-pentene	-0.1693	-0.2704	-0.1807	-0.2490	-0.2097	-0.0039	-0.0017	-0.3972	-0.0028	-0.0146
3-methylpentane	-0.0458	-0.1352	-0.0430	-0.0904	-0.0721	-0.0004	-0.0025	-0.0001	-0.0281	0.0039
hexane	0.1338	0.0416	0.1409	0.0929	0.1087	0.0074	0.0033	0.0001	0.0046	0.0269
cyclopentane	-0.3235	-0.4095	-0.3340	-0.3853	-0.3560	-0.3509	-0.0070	-0.0023	-0.0004	-0.0043
1-hexene	-0.1098	-0.1941	-0.1190	-0.1676	-0.1407	-0.0039	-0.3635	-0.0088	-0.0133	-0.0146
heptane	0.0913	0.0175	0.0980	0.0701	0.0750	-0.0029	-0.0009	-0.1610	-0.0014	-0.0034
cyclohexane	-0.2899	-0.3593	-0.2980	-0.3296	-0.3128	-0.0022	0.0003	-0.0009	0.1227	0.0017
2,2,4-trimethylpentane	-0.7307	-0.7771	-0.7380	-0.7440	-0.7418	-0.0065	-0.0003	-0.0093	-0.0131	-0.1574
1-pentyne	0.1056	0.0079	0.0849	0.0150	0.0614	-0.3857	-0.0128	-0.0061	-0.0133	-0.0002
diethyl ether	-0.0051	0.0814	0.0543	0.0137	0.0055	-0.0582	-0.4137	-0.0380	-0.0732	-0.0563
methylcyclohexane	-0.4936	-0.5443	-0.5022	-0.5138	-0.5076	-0.0116	0.0013	-0.0355	-0.0100	-0.0101
1-heptene	-0.0678	-0.1353	-0.0747	-0.1036	-0.0891	-0.0500	-0.0131	-0.0223	-0.1319	-0.0356
diisopropyl ether	-0.5465	-0.4142	-0.4843	-0.4873	-0.5135	-0.0488	-0.1084	-0.0288	-0.0703	-0.1211
cyclohexene	-0.3406	-0.4047	-0.3609	-0.3937	-0.3681	-0.7647	0.0010	0.0028	-0.0012	0.0107
tert-butyl ethyl ether	-0.5829	-0.4492	-0.5237	-0.5261	-0.5509	-0.0084	-0.0629	0.0044	-0.0131	-0.0596
octane	0.0838	0.0280	0.0911	0.0830	0.0767	0.0103	0.0018	0.2254	0.0164	0.0021
methanol	-0.1761	-0.2297	-0.1916	-0.2639	-0.2586	-0.0668	-0.0599	-0.0471	-0.2387	-0.0687
1-hexyne	0.2119	0.1298	0.1952	0.1448	0.1777	-0.0095	0.0046	0.0123	0.0282	0.3090
tert-butyl methyl ether	-0.4490	-0.3339	-0.4013	-0.4207	-0.4317	0.0749	-0.0221	0.0262	0.0396	0.0115
cycloheptane	-0.1028	-0.1556	-0.1084	-0.1204	-0.1160	-0.0026	0.0142	-0.0274	0.0030	-0.0135
dipropyl ether	-0.0343	0.0911	0.0337	0.0305	-0.0008	-0.0412	-0.0278	0.1069	-0.0159	-0.0226
1-octene	-0.0202	-0.0710	-0.0247	-0.0340	-0.0318	-0.0074	-0.0154	-0.0240	-0.3333	-0.0499
ethanol	0.2420	0.1964	0.2421	0.1876	0.1727	-0.0238	-0.0062	0.0161	-0.1013	-0.0102
nonane	0.0671	0.0289	0.0753	0.0871	0.0692	0.0874	0.0108	0.0169	-0.0176	-0.0002
2-propanol	-0.2211	-0.2446	-0.2269	-0.2605	-0.2830	0.0070	-0.0490	-0.0316	-0.0400	0.0466
methyl tert-pentyl ether	-0.2231	-0.0915	-0.1664	-0.1678	-0.1930	0.0574	0.0932	0.5283	0.0764	0.0962
1-heptyne	0.2908	0.2247	0.2776	0.2467	0.2666	0.0520	0.0134	0.0246	0.1215	0.0415
2-methyl-2-propanol	-0.6954	-0.6970	-0.7069	-0.7198	-0.7499	0.0590	0.0243	-0.0076	-0.0158	0.0456
cyclooctane	0.0435	0.0074	0.0404	0.0481	0.0400	0.3955	-0.0098	0.0132	-0.0034	0.0093
1-propanol	0.2509	0.2225	0.2526	0.2179	0.1910	0.0181	-0.1248	0.0108	0.0911	0.0118
decane	0.0747	0.0540	0.0842	0.1159	0.0861	0.0062	0.0143	0.0162	0.0183	0.0293
2-butanol	-0.0694	-0.0785	-0.0693	-0.0837	-0.1207	-0.0067	-0.0169	0.0046	0.0241	-0.0338
tetrahydrofuran	0.2367	0.3481	0.2823	0.2574	0.2509	-0.0140	-0.1076	-0.0112	-0.0214	-0.8936
methyl acetate	-0.3325	-0.2375	-0.3160	-0.3472	-0.3400	-0.5939	-0.0899	-0.1738	-0.0974	-0.0720
1-iodobutane	0.3094	0.3790	0.4079	0.5358	0.3547	-0.0107	0.2977	0.0000	-0.0017	0.0000
1-octyne	0.3594	0.3095	0.3494	0.3380	0.3450	0.0058	-0.0004	0.0001	0.0292	0.0079
2-methyl-1-propanol	-0.0967	-0.1034	-0.1003	-0.1142	-0.1490	0.0237	0.0211	0.0627	-0.1029	0.0210
benzene	-0.1374	-0.1919	-0.1806	-0.2159	-0.1736	-0.0058	-0.0117	-0.0151	-0.0060	0.1249
dibutyl ether	0.0554	0.2175	0.1306	0.1655	0.1105	-0.4070	0.0172	0.0168	0.0149	0.0103
propanone	-0.1609	-0.1780	-0.1813	-0.2163	-0.2283	-0.0371	-0.2255	-0.0723	-0.0378	-0.1436
thiophene	-0.8167	-0.7756	-0.8103	-0.7661	-0.8396	-0.0017	-0.0024	-0.2716	0.0032	-0.0041
1-decene	0.0557	0.0383	0.0559	0.0860	0.0633	0.0419	0.0324	0.0331	0.2452	0.0413
butanal	-0.0001	-0.0087	-0.0056	-0.0226	-0.0545	0.0104	-0.0338	0.0031	0.0021	-0.5959
1-chlorohexane	-0.1683	-0.1355	-0.1402	-0.0928	-0.1826	-0.4548	-0.0909	-0.0624	-0.0464	-0.0310
ethyl acetate	0.0610	0.1714	0.0940	0.0791	0.0695	-0.0244	-0.0532	-0.0116	-0.0205	-0.0249
1-butanol	0.3353	0.3238	0.3391	0.3241	0.2849	0.0350	0.0429	0.2452	0.0796	0.0309
methyl propanoate	0.0323	0.1407	0.0625	0.0487	0.0388	-0.0129	-0.0528	0.0076	-0.1217	-0.0001
1-bromohexane	-0.1831	-0.1105	-0.1272	-0.0203	-0.1614	-0.0310	-0.0034	0.0013	0.0053	0.5492
acetonitrile	0.2311	0.1991	0.1996	0.1420	0.1499	0.0214	0.0218	0.0022	0.0151	-0.0179
toluene	0.0362	-0.0031	-0.0023	-0.0183	0.0102	-0.0010	-0.3361	-0.0003	-0.0006	-0.0035
methyl butanoate	0.1795	0.3035	0.2174	0.2221	0.1978	0.0144	0.0504	-0.0978	0.0333	0.0276
1-pentanol	0.3849	0.3902	0.3909	0.3956	0.3441	0.0000	0.0103	0.0053	0.3493	0.0144
acetic acid	-0.3610	-0.3631	-0.3999	-0.4289	-0.4318	0.0027	0.0091	0.0133	0.0040	-0.2575
propionitrile	0.6110	0.5868	0.5955	0.5556	0.5431	0.0810	0.1144	0.0827	0.0500	0.0791
1,4-dioxane	0.1569	0.2995	0.1946	0.1992	0.1818	0.0123	-0.0781	0.0876	0.0317	0.0065
3-pentanone	0.2654	0.2731	0.2629	0.2654	0.2208	0.0074	0.0393	0.0222	0.0170	0.0314
2-pentanone	0.1957	0.2050	0.1907	0.1936	0.1504	0.0111	0.0492	0.0364	0.0312	0.0396
ethylbenzene	0.2314	0.2052	0.2010	0.2039	0.2167	-0.0127	-0.0004	-0.0035	-0.0046	-0.0123
1,4-dimethylbenzene	0.1460	0.1223	0.1118	0.1152	0.1302	-0.0802	-0.0120	-0.0116	-0.0178	-0.0001
1-chlorooctane	-0.2271	-0.1582	-0.1985	-0.1111	-0.2233	0.0103	-0.2670	0.0244	0.0044	-0.0019
1,3-dimethylbenzene	0.1812	0.1575	0.1469	0.1503	0.1654	-0.0076	-0.0012	-0.1269	0.0008	0.0016
pyridine	-0.7822	-0.7453	-0.8154	-0.7986	-0.8305	0.0593	0.0171	-0.0127	-0.1231	0.0142
1,2-dimethylbenzene	0.3052	0.2813	0.2710	0.2744	0.2894	0.0166	0.0135	0.0136	0.0187	0.1371
propionic acid	0.1521	0.1582	0.1285	0.1173	0.0943	0.5903	0.0729	0.0675	0.0478	0.0314
1-bromooctane	-0.4075	-0.2956	-0.3560	-0.2085	-0.3690	0.0135	0.0304	0.0036	0.0101	0.0296
cyclohexanol	-0.0641	-0.0216	-0.0672	-0.0270	-0.0920	0.0008	0.0033	-0.1573	-0.0031	-0.0139
1-nitropropane	-0.0097	0.0178	-0.0259	-0.0096	-0.0526	0.0027	0.0084	-0.0112	0.0363	-0.0007
styrene	0.3397	0.3171	0.2996	0.3009	0.3210	0.0010	-0.0015	0.0009	-0.0031	-0.2184
1,2-dichlorobenzene	-1.4079	-1.3010	-1.4169	-1.3096	-1.4075	-0.1455	-0.0009	0.0012	0.0015	0.0012
methyl hexanoate	0.4673	0.6232	0.5168	0.5593	0.5077	0.0004	-0.2327	-0.0007	0.0000	0.0009
alpha-methylstyrene	0.3154	0.3083	0.2796	0.3003	0.3069	0.0032	0.0025	0.2594	0.0011	-0.0020
octanal	0.3326	0.3882	0.3409	0.4019	0.3179	0.0033	0.0002	-0.0001	0.9086	0.0030
pyrrole	0.1912	0.2131	0.1573	0.1574	0.1351	0.0238	0.0147	0.0365	0.0107	0.6250
cyclohexanone	0.2965	0.3393	0.2882	0.3258	0.2658	0.4763	0.0048	0.0007	0.0064	0.0072
ethyl phenyl ether	0.3913	0.5764	0.4346	0.4801	0.4394	0.0000	-0.5067	0.0004	-0.0006	-0.0026
1-octanol	0.5045	0.5600	0.5176	0.5813	0.4926	0.0000	0.0000	0.4970	0.0000	0.0000
benzaldehyde	-0.3152	-0.2481	-0.3448	-0.2929	-0.3466	-0.0007	-0.0152	-0.0535	-0.5109	-0.0136
benzonitrile	-0.1240	-0.0597	-0.1555	-0.1076	-0.1578	0.0007	0.0082	-0.0176	-0.0252	-0.0385
phenol	0.1374	0.1906	0.1058	0.1414	0.0981	-0.3447	-0.0089	-0.0295	-0.0060	0.0009
n,n-dimethylformamide	1.1001	1.1127	1.0719	1.0651	1.0428	0.0020	1.5820	0.0074	0.0038	0.0430
2-nitrophenol	-0.9789	-0.8614	-1.0089	-0.9018	-0.9857	-0.0004	-0.0124	0.1169	-0.0093	0.0033
decyl alcohol	0.5986	0.6875	0.6165	0.7195	0.6060	0.0000	0.0000	0.0000	0.7045	0.0000
nitrobenzene	-0.4942	-0.3979	-0.5241	-0.4403	-0.5114	0.0004	0.0071	0.0008	0.0007	0.2088
acetophenone	0.0542	0.1361	0.0298	0.1011	0.0332	0.5123	0.0040	0.0134	0.0095	0.0016
p-cresol	0.2991	0.3671	0.2726	0.3275	0.2701	-0.0004	0.0512	0.0123	-0.0008	-0.0005
benzyl alcohol	0.3407	0.4081	0.3152	0.3699	0.3119	0.0001	0.0011	0.0232	0.0053	0.0026
aniline	0.6024	0.6531	0.5725	0.6064	0.5628	-0.0003	-0.0010	0.0063	0.6996	0.0039
2-chloroaniline	0.0980	0.1868	0.0785	0.1601	0.0833	0.0006	0.0163	0.0568	0.0223	0.3573
naphthalene	0.8802	0.9026	0.8431	0.8954	0.8854	1.0432	0.0008	0.0114	0.0067	0.0045

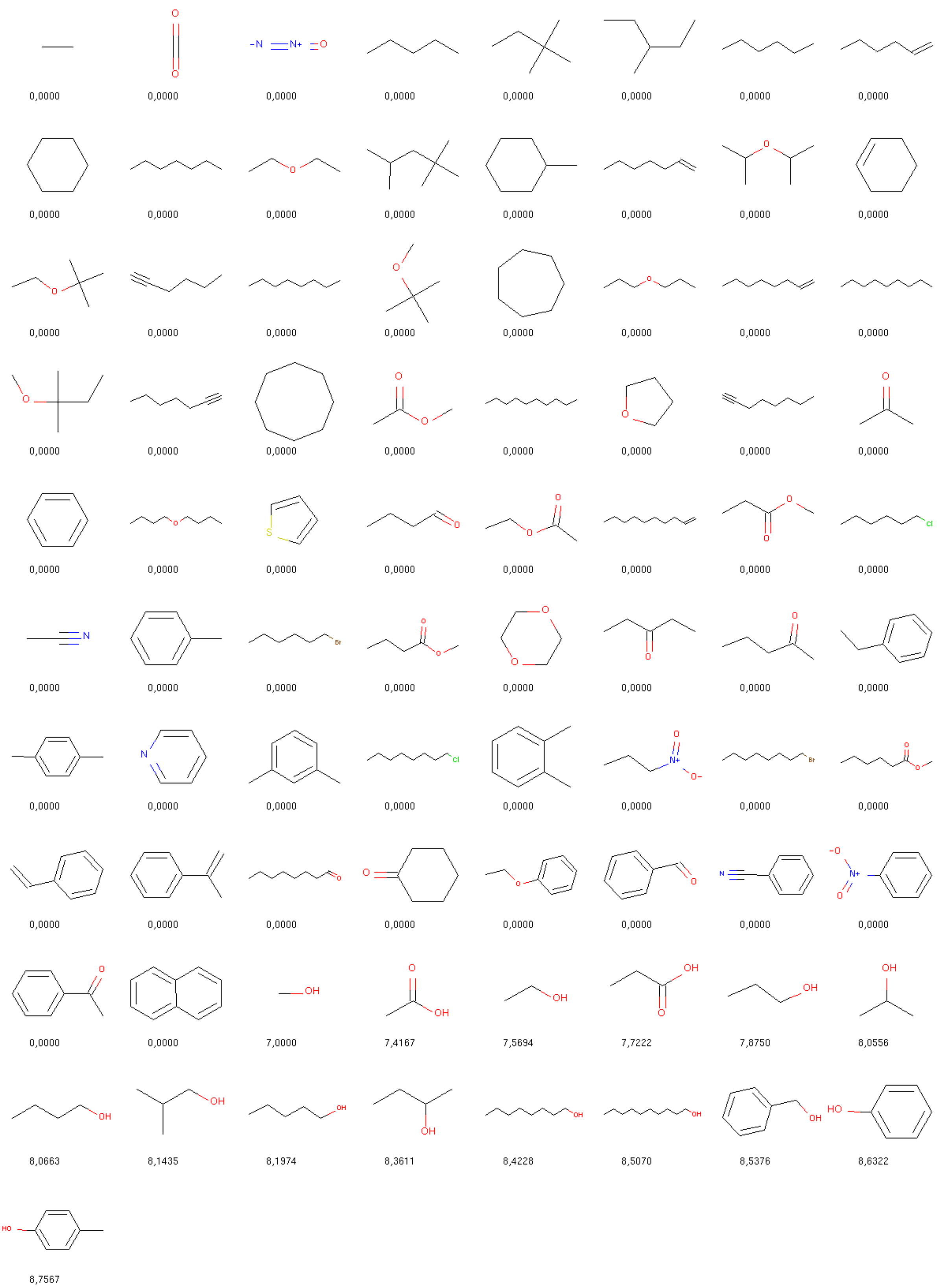
TREENING

TEST

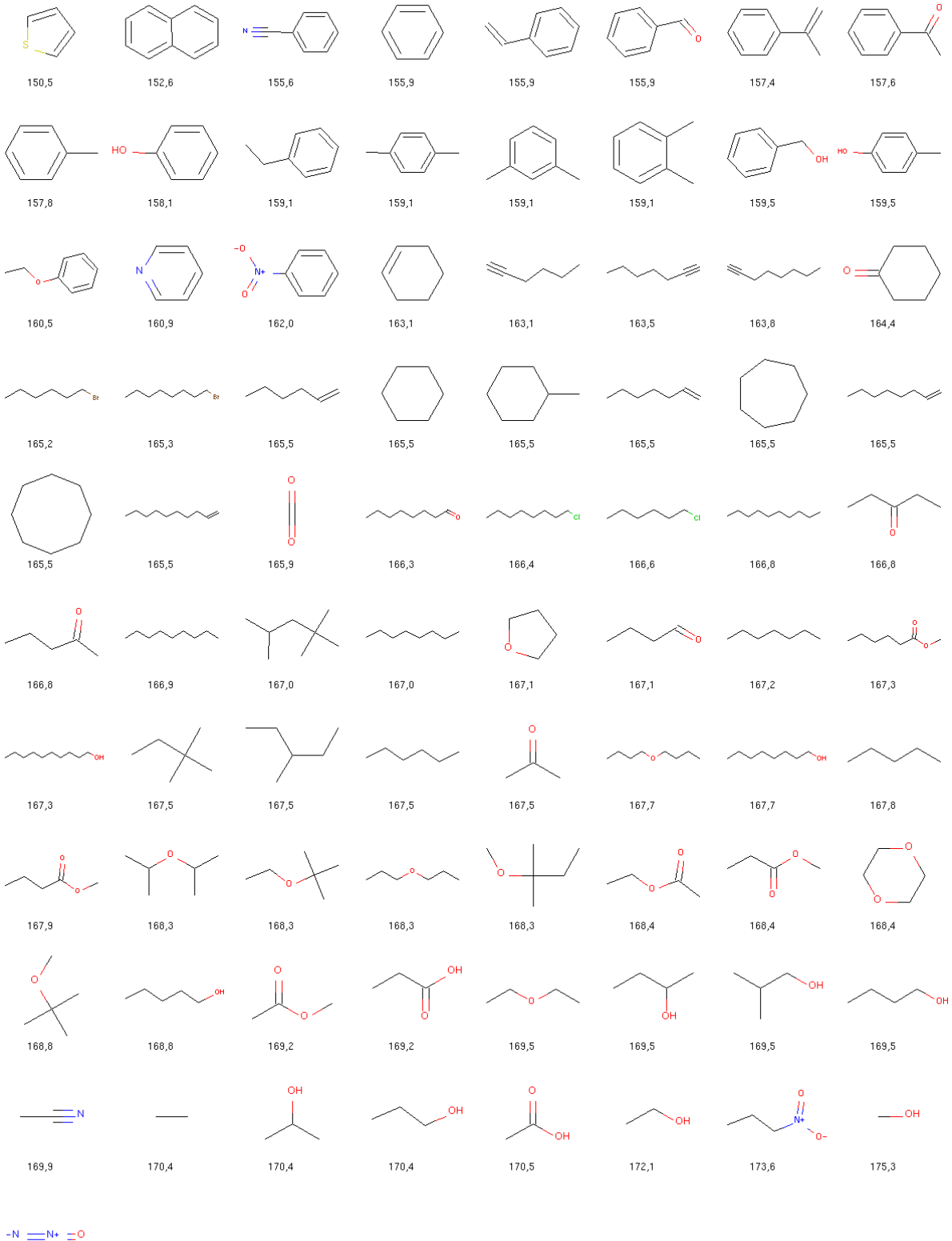
Lisa 5. Tunnuse *VR1_A* väärtused kasvavas järjestuses.



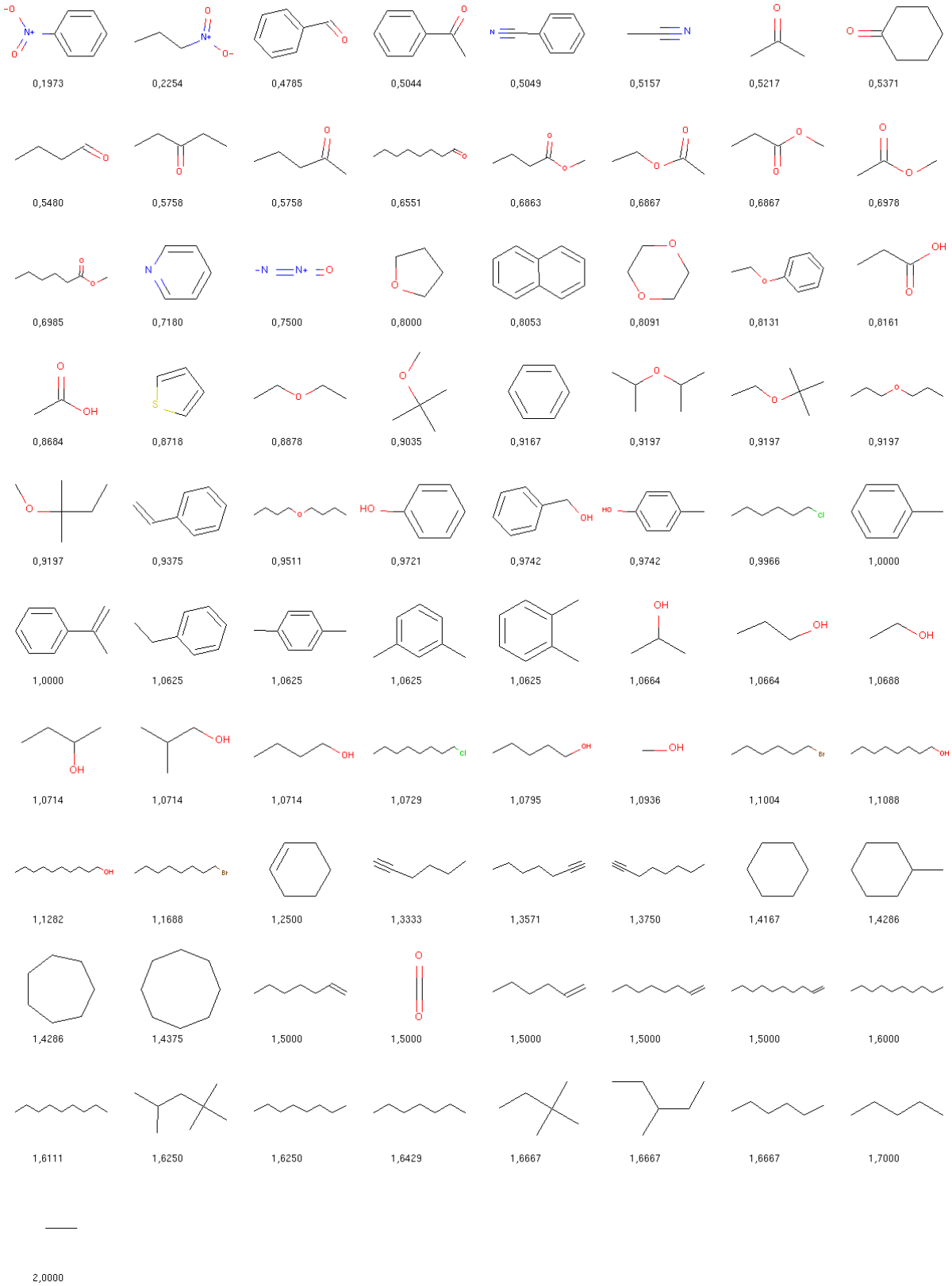
Lisa 6. Tunnuse *SsOH* väärtused kasvavas järjestuses.



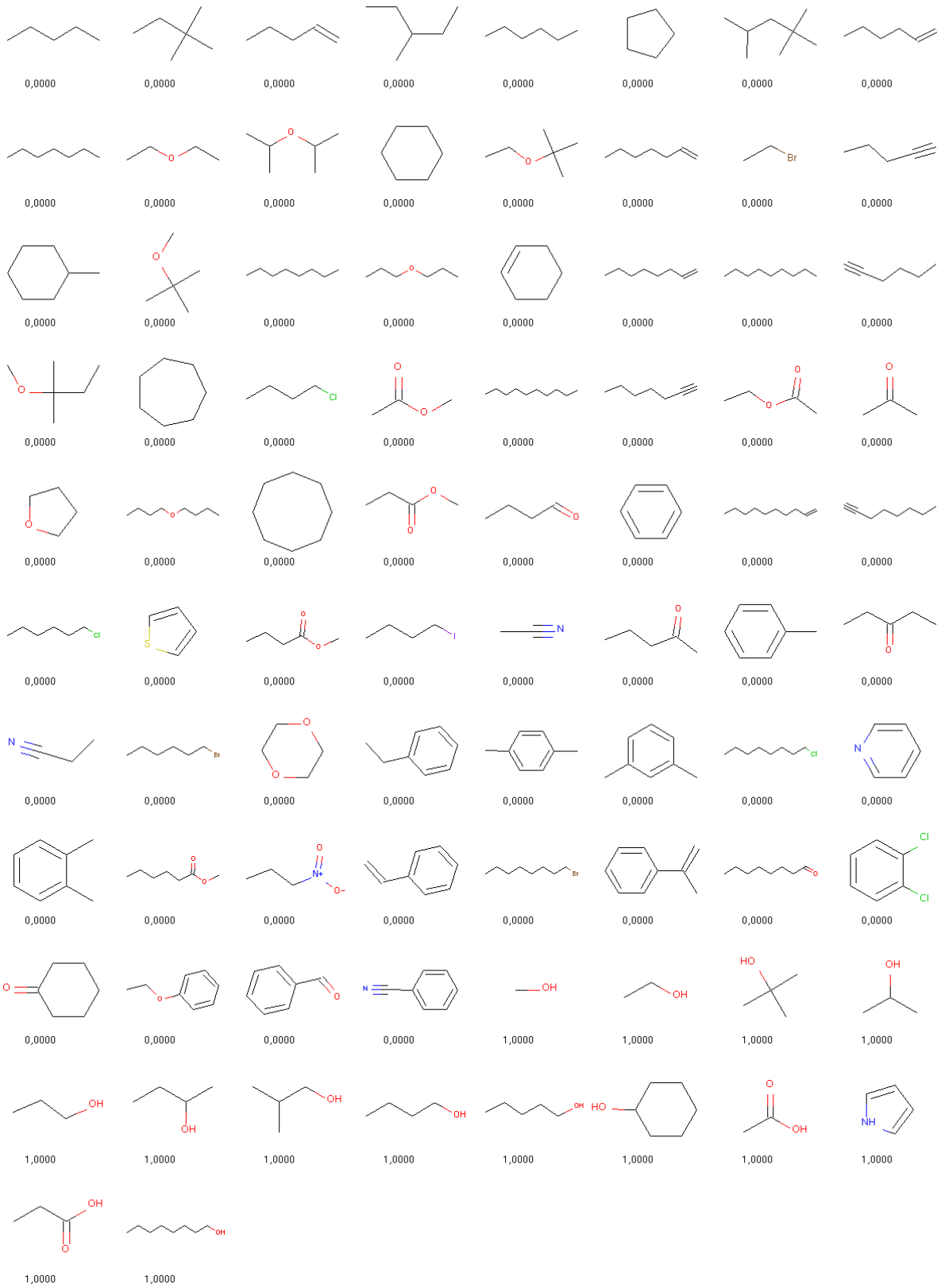
Lisa 7. Tunnuse *AATS0i* väärtused kasvavas järjestuses.



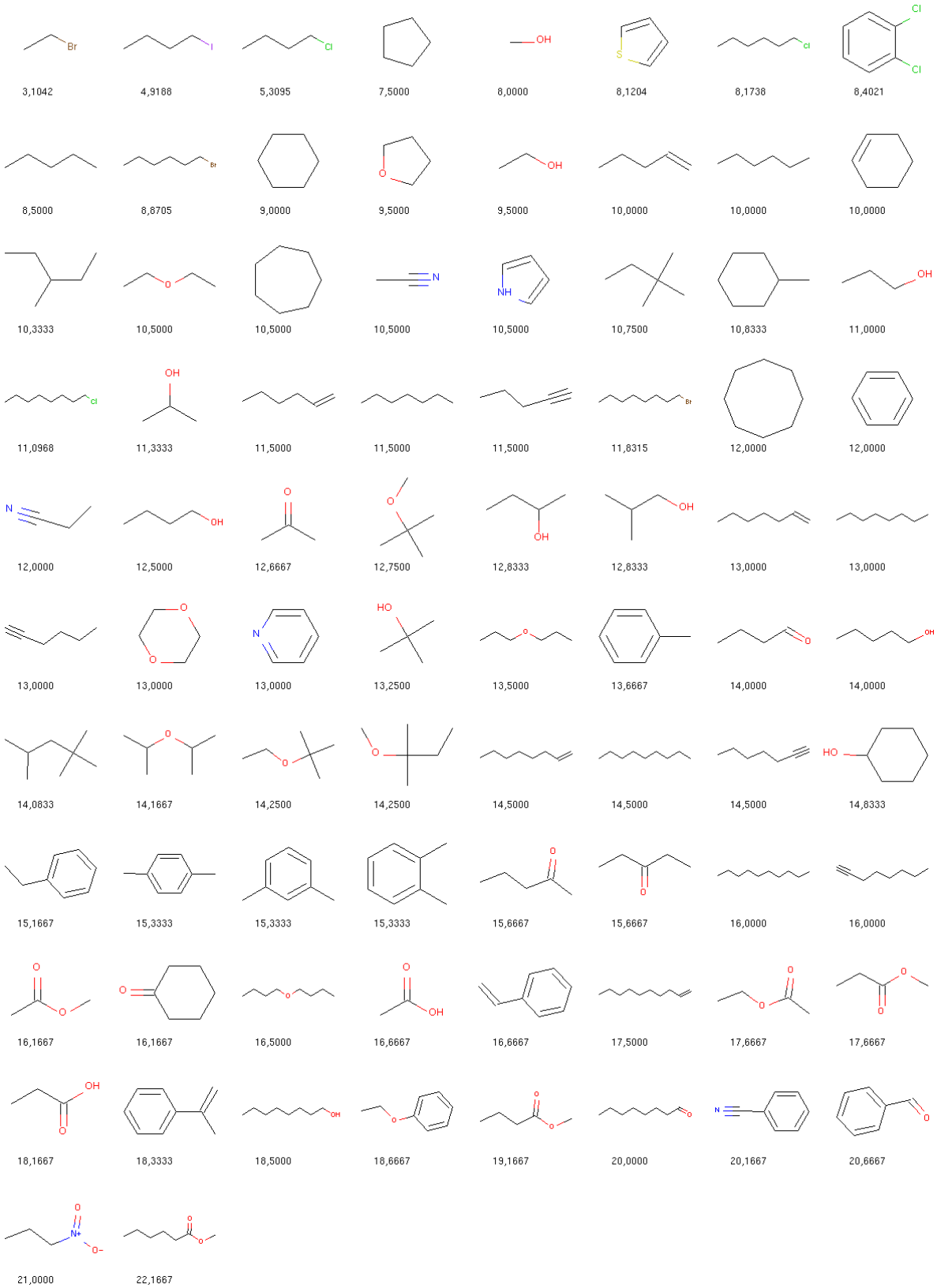
Lisa 8. Tunnuse *GATS1are* väärtused kasvavas järjestuses.



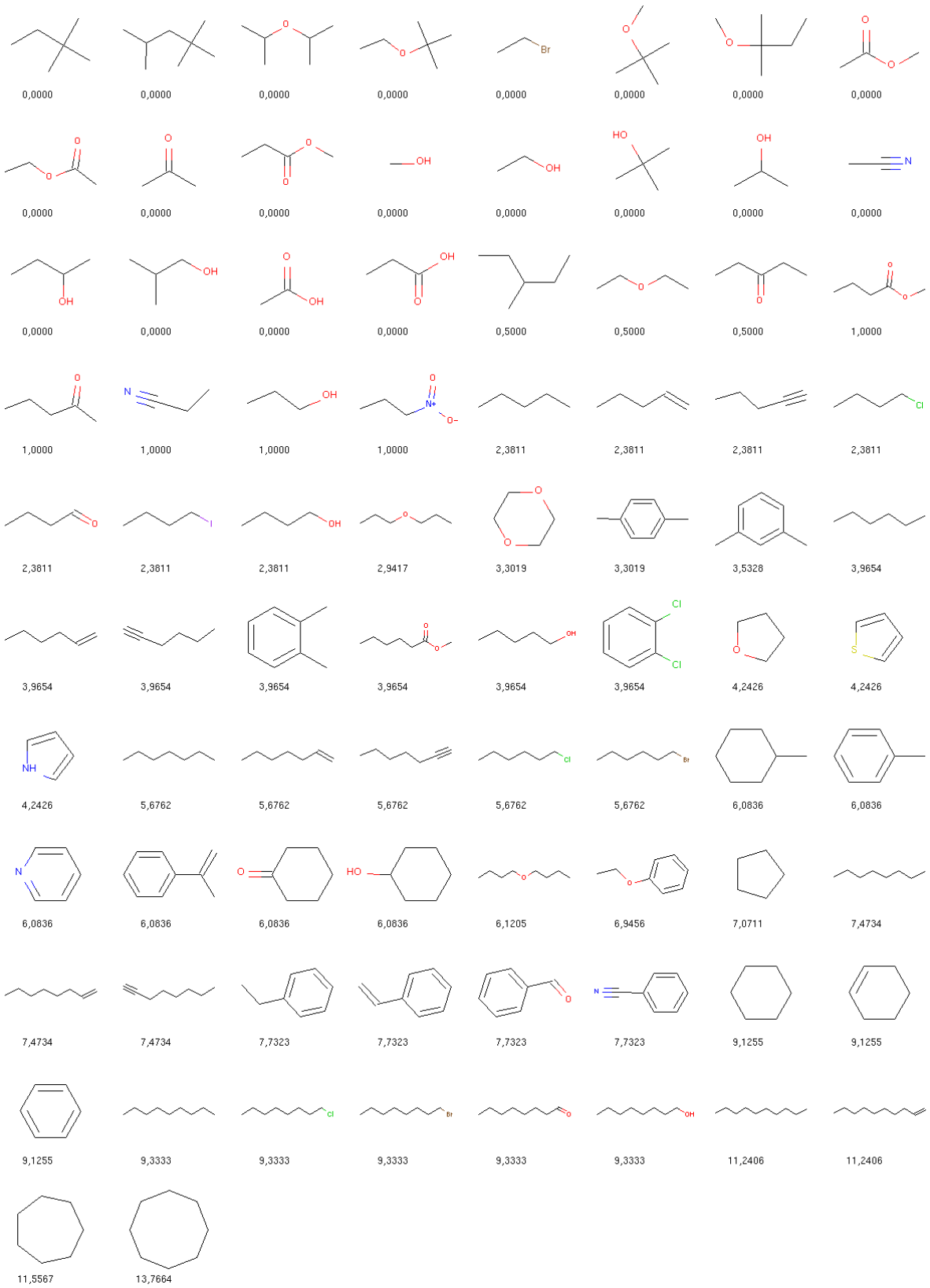
Lisa 9. Tunnuse *nHBDon* väärtused kasvavas järjestuses.



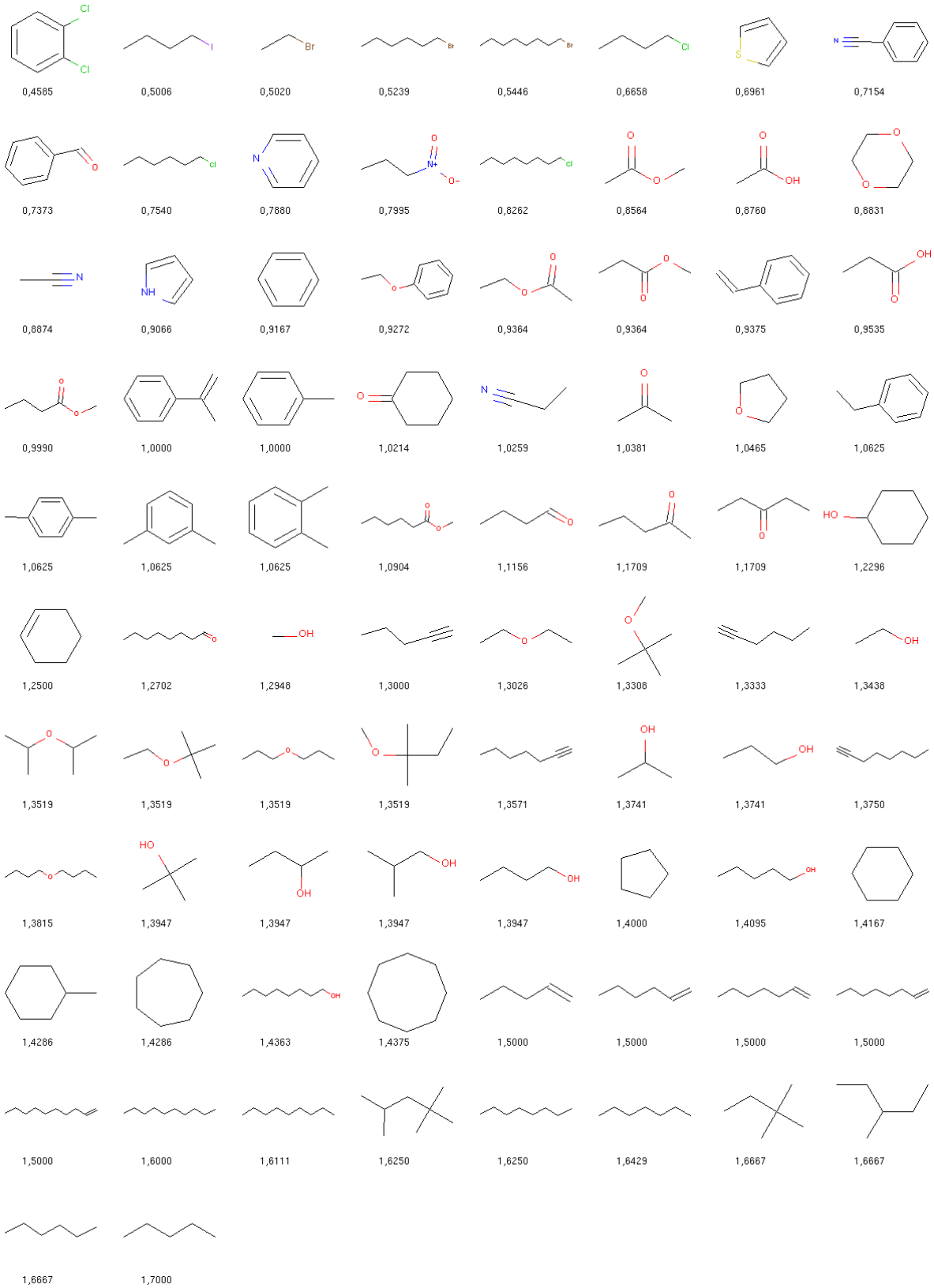
Lisa 10. Tunnuse *VSA_EState9* väärtused kasvavas järjestuses.



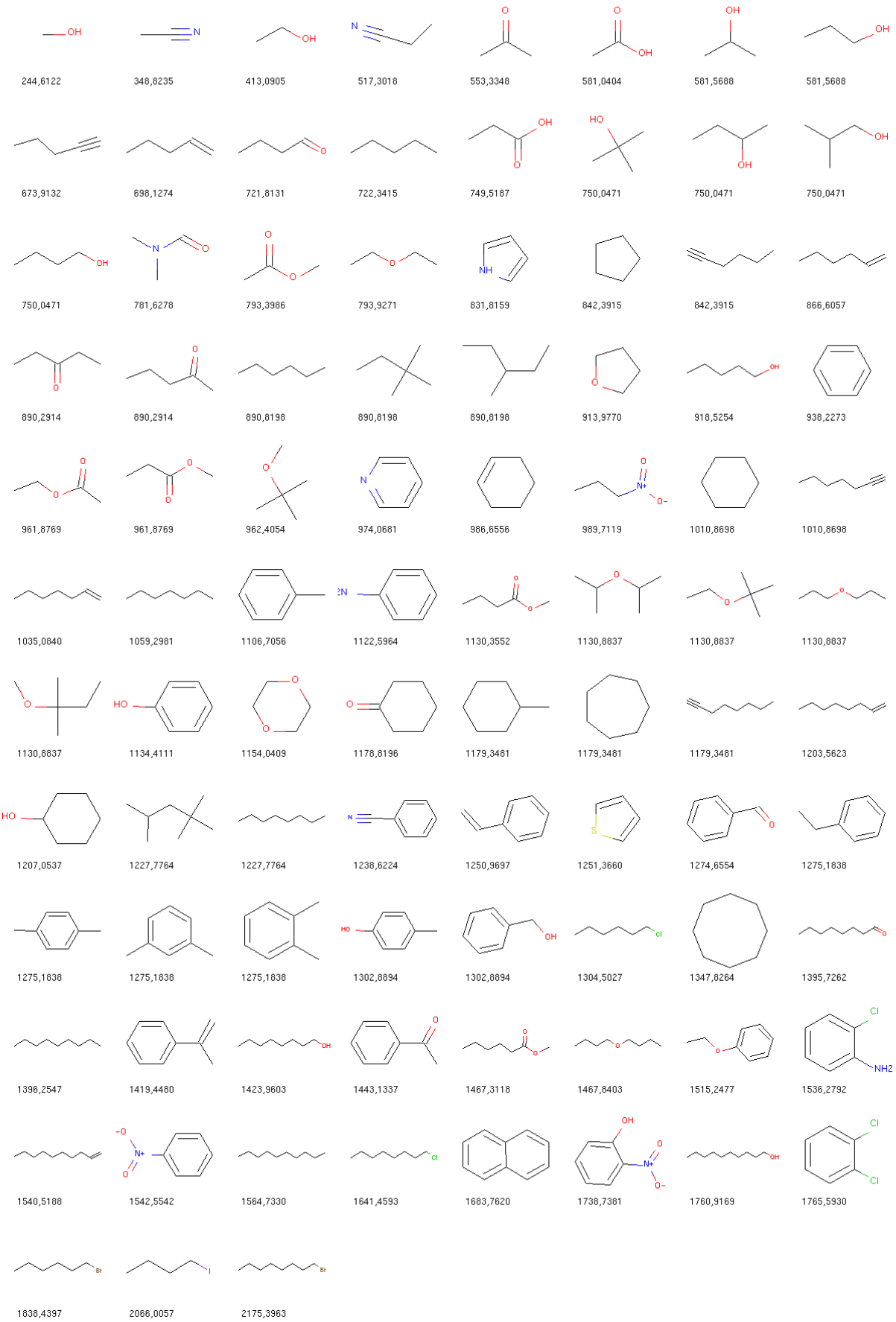
Lisa 11. Tunnuse *MDEC-22* väärtused kasvavas järjestuses.



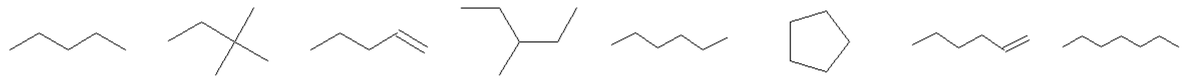
Lisa 12. Tunnuse *GATS1m* väärtused kasvavas järjestuses.



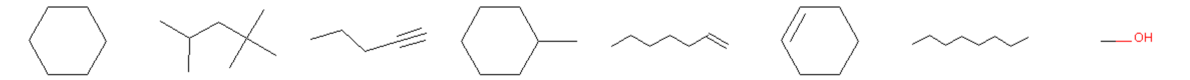
Lisa 13. Tunnuse *ATS1m* väärtused kasvavas järjestuses.



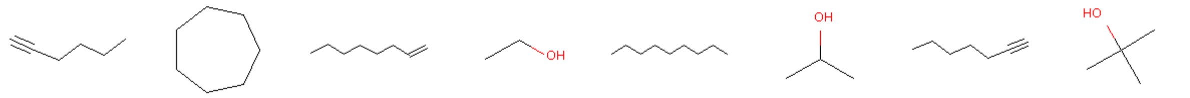
Lisa 14. Tunnuse *MAXssO* väärtused kasvavas järjestuses.



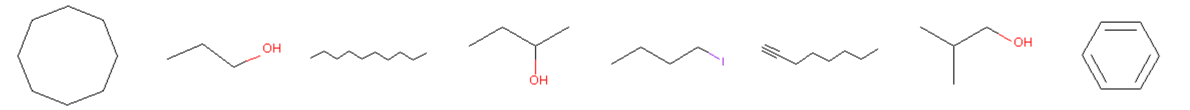
0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000



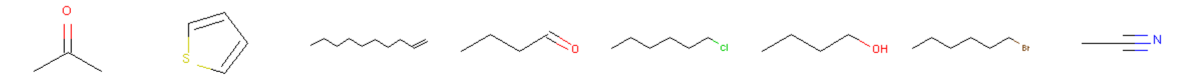
0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000



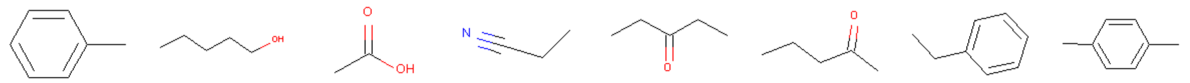
0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000



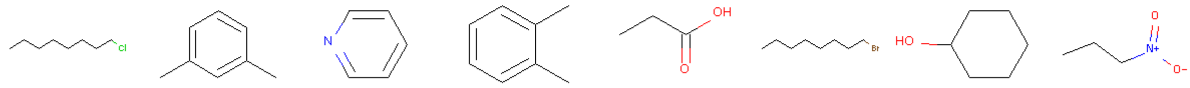
0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000



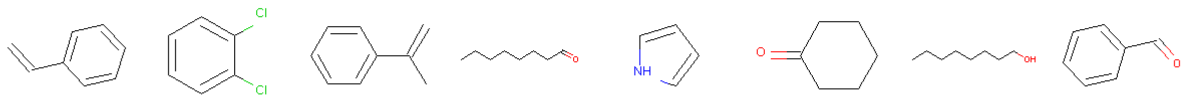
0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000



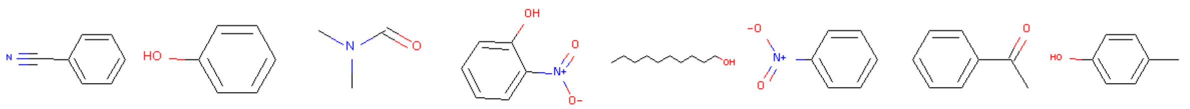
0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000



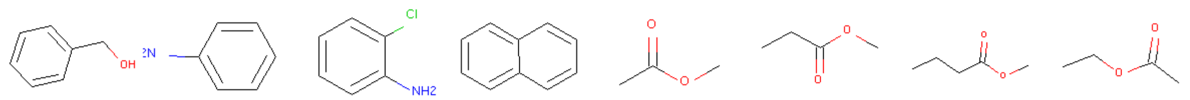
0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000



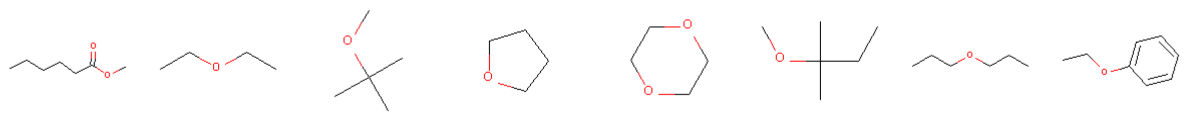
0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000



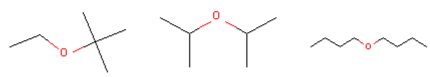
0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000 0,0000



0,0000 0,0000 0,0000 0,0000 4,1111 4,2604 4,3517 4,4028

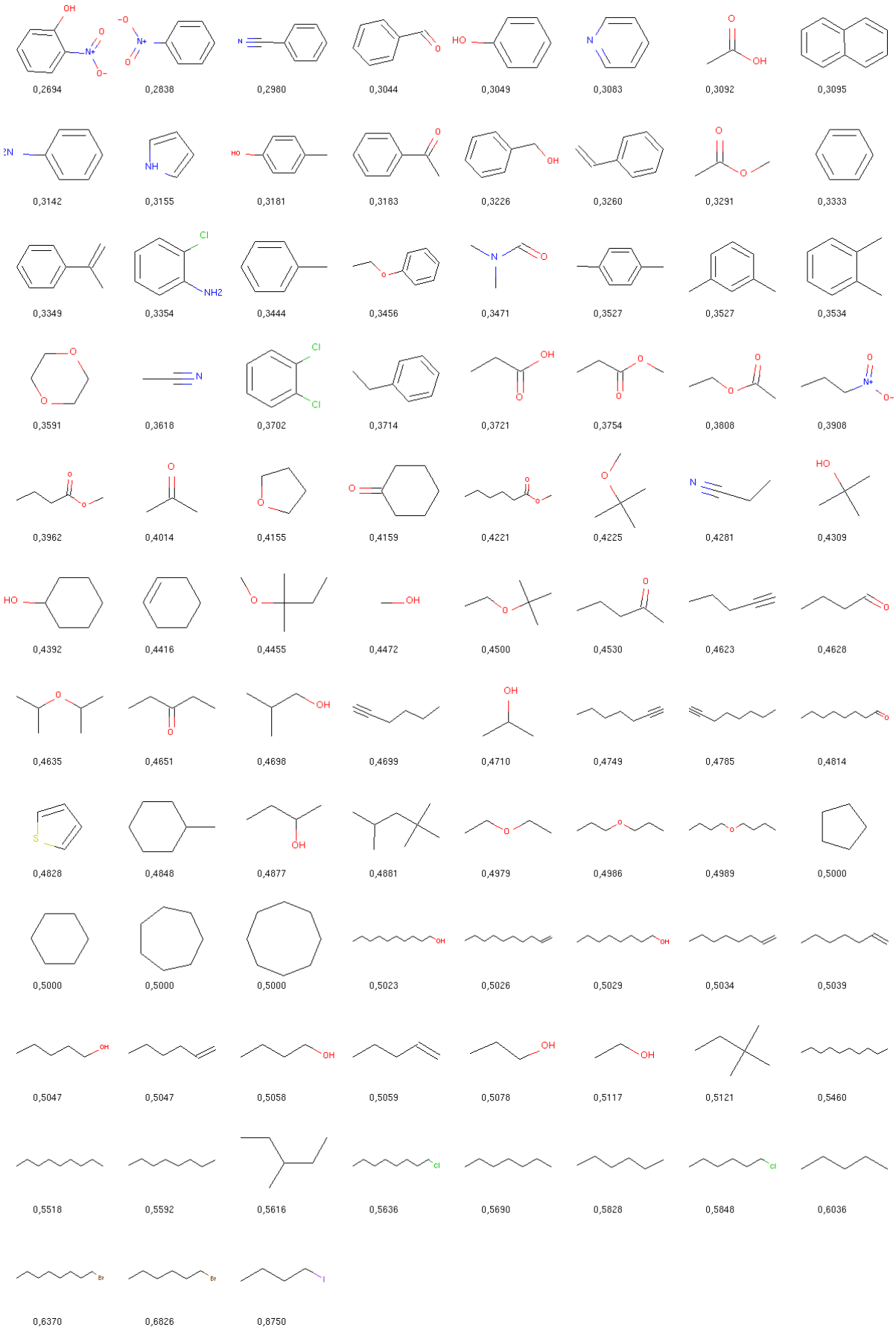


4,4578 4,8333 4,9375 4,9444 4,9444 5,0868 5,1319 5,2058

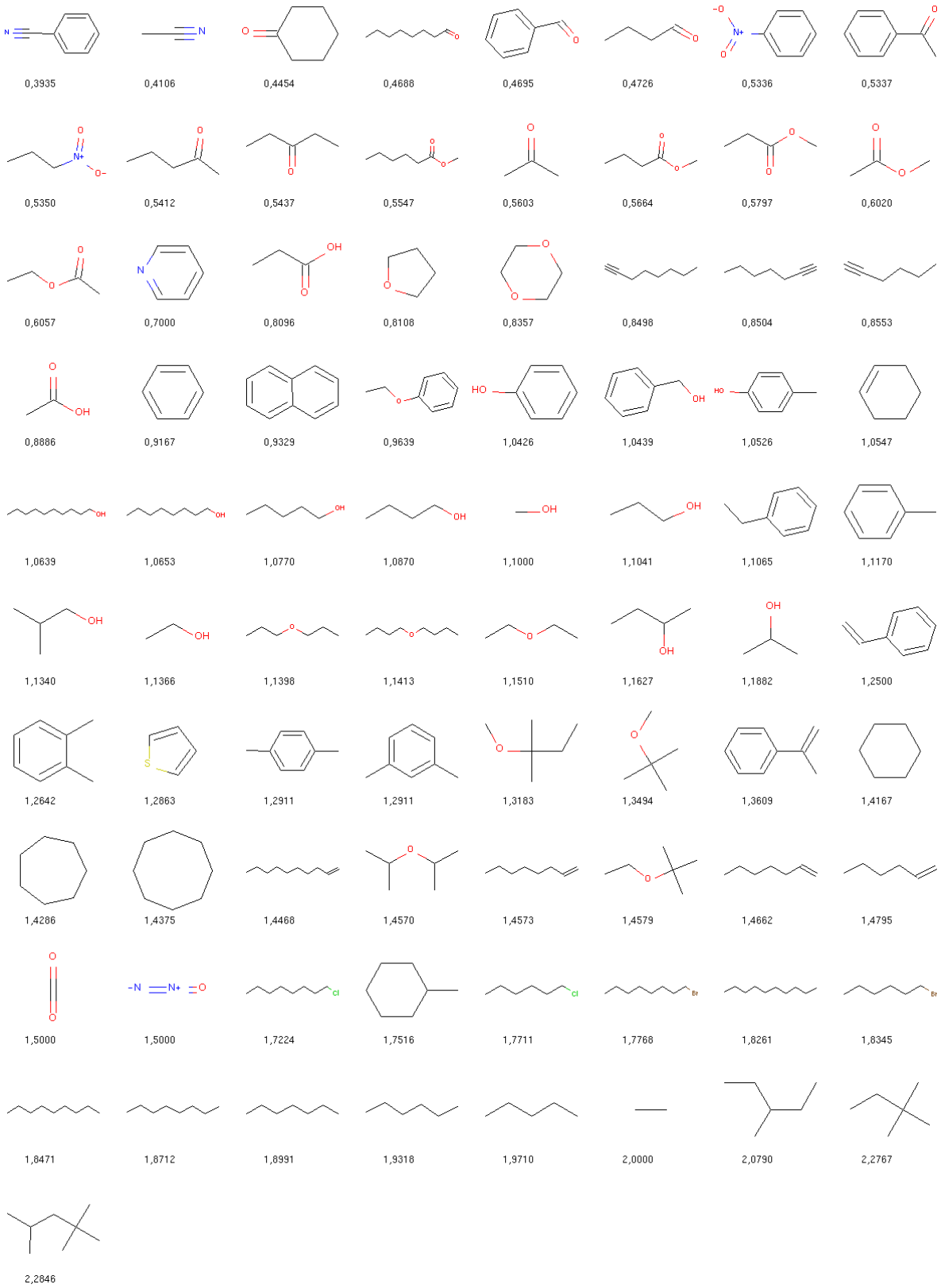


5,2292 5,2500 5,3144

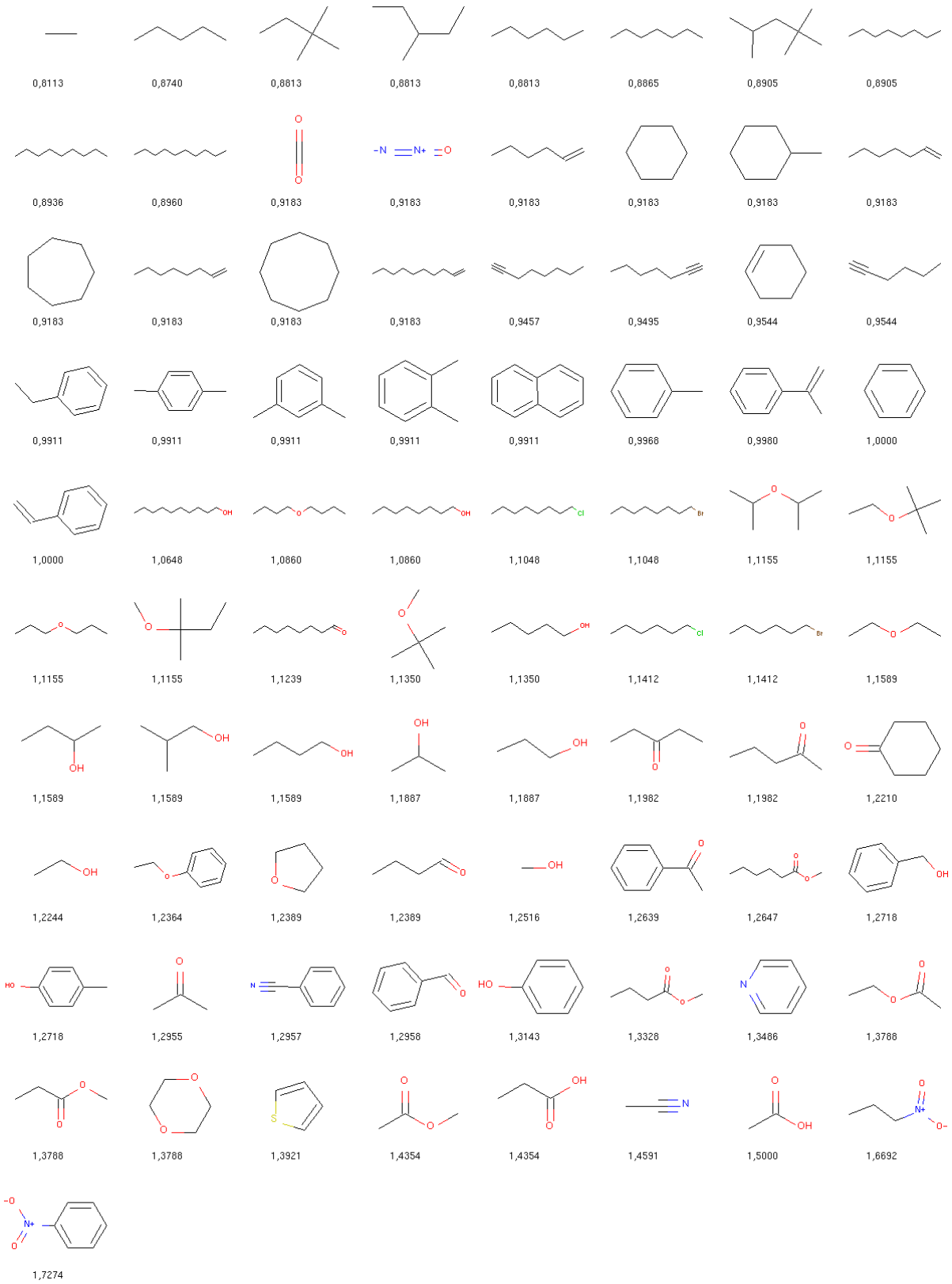
Lisa 15. Tunnuse *AXp-1dv* väärtused kasvavas järjestuses.



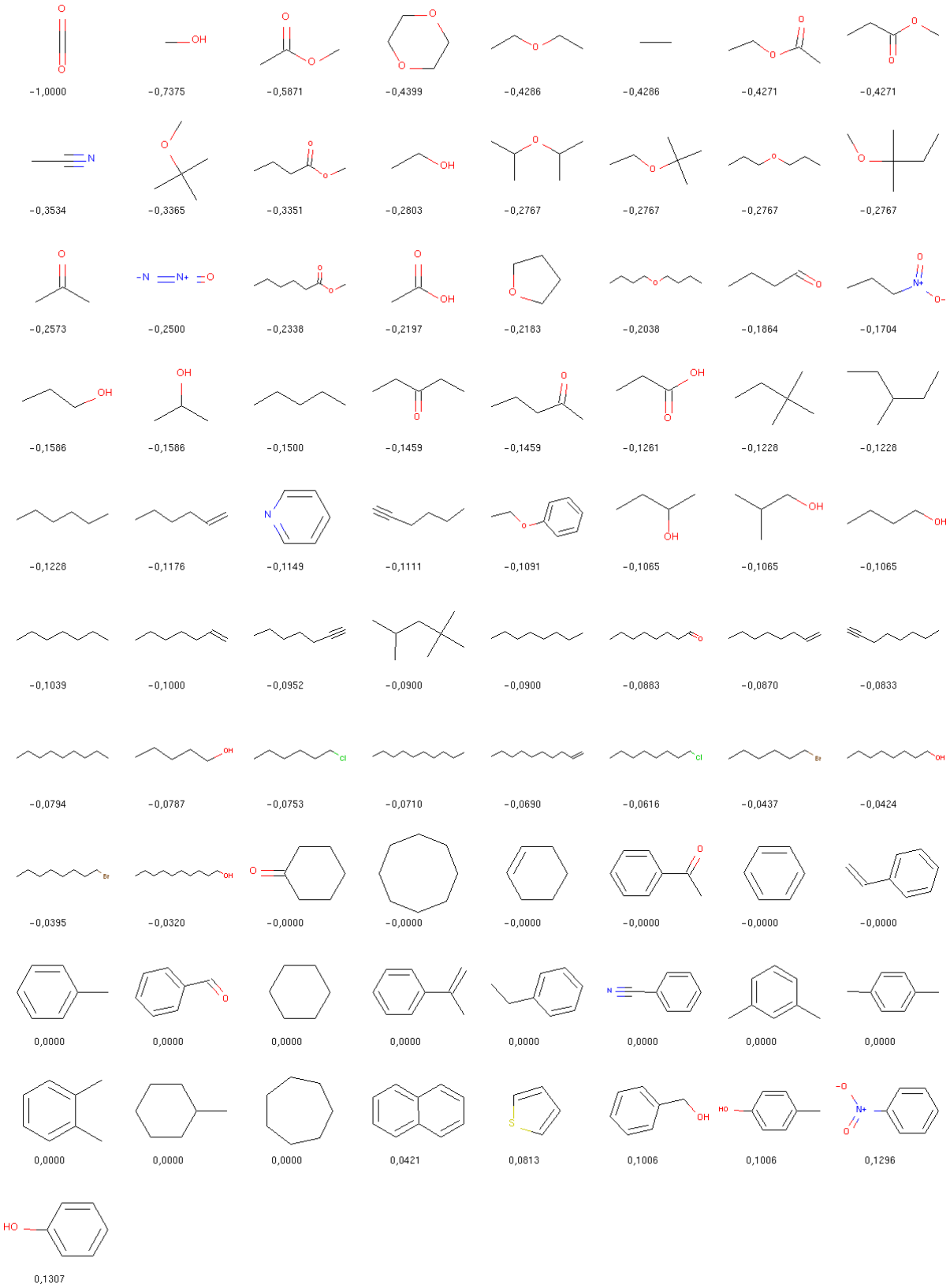
Lisa 16. Tunnuse *GATS1s* väärtused kasvavas järjestuses.



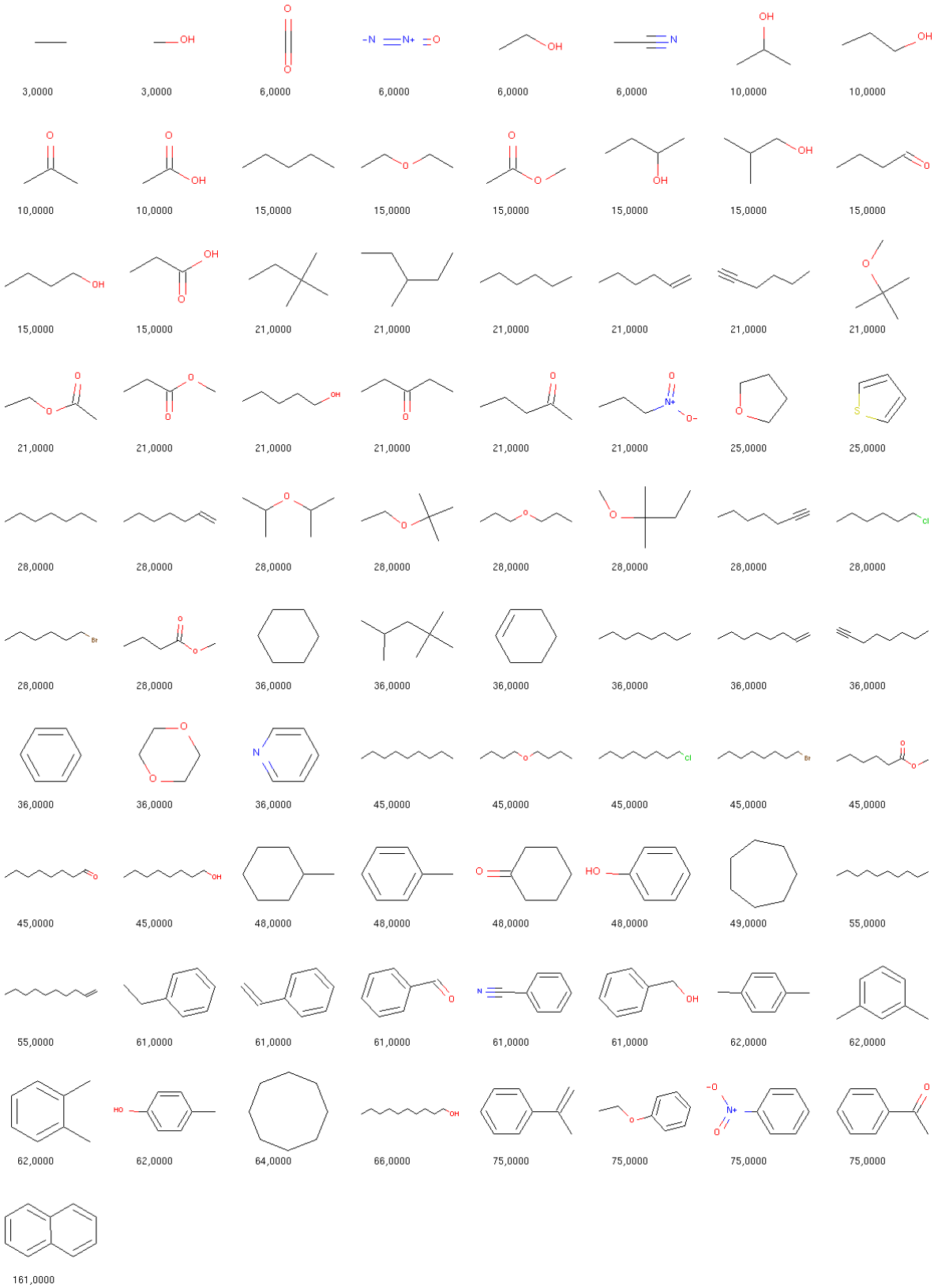
Lisa 17. Tunnuse *IC0* väärtused kasvavas järjestuses.



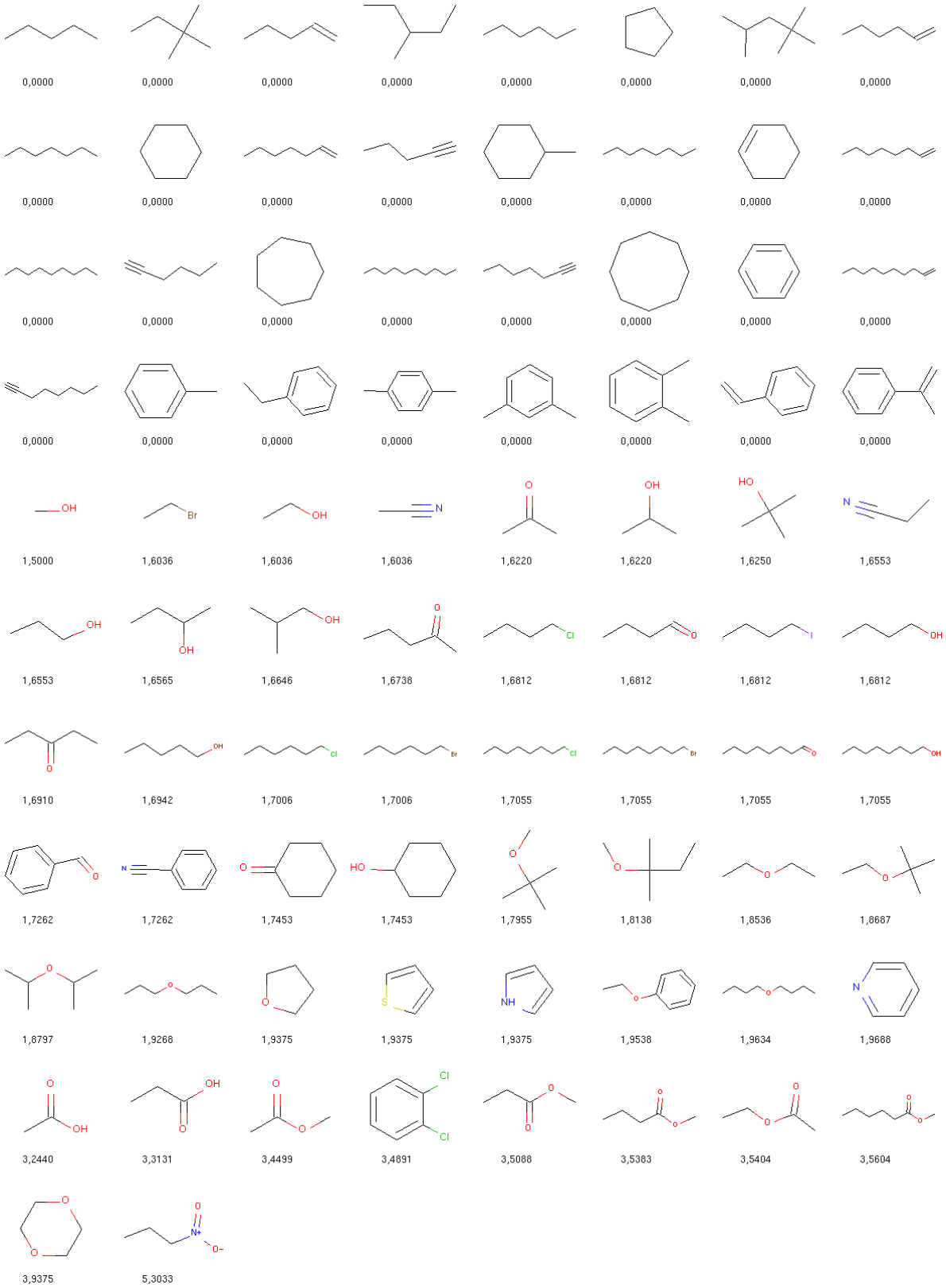
Lisa 18. Tunnuse *MATS1p* väärtused kasvavas järjestuses.



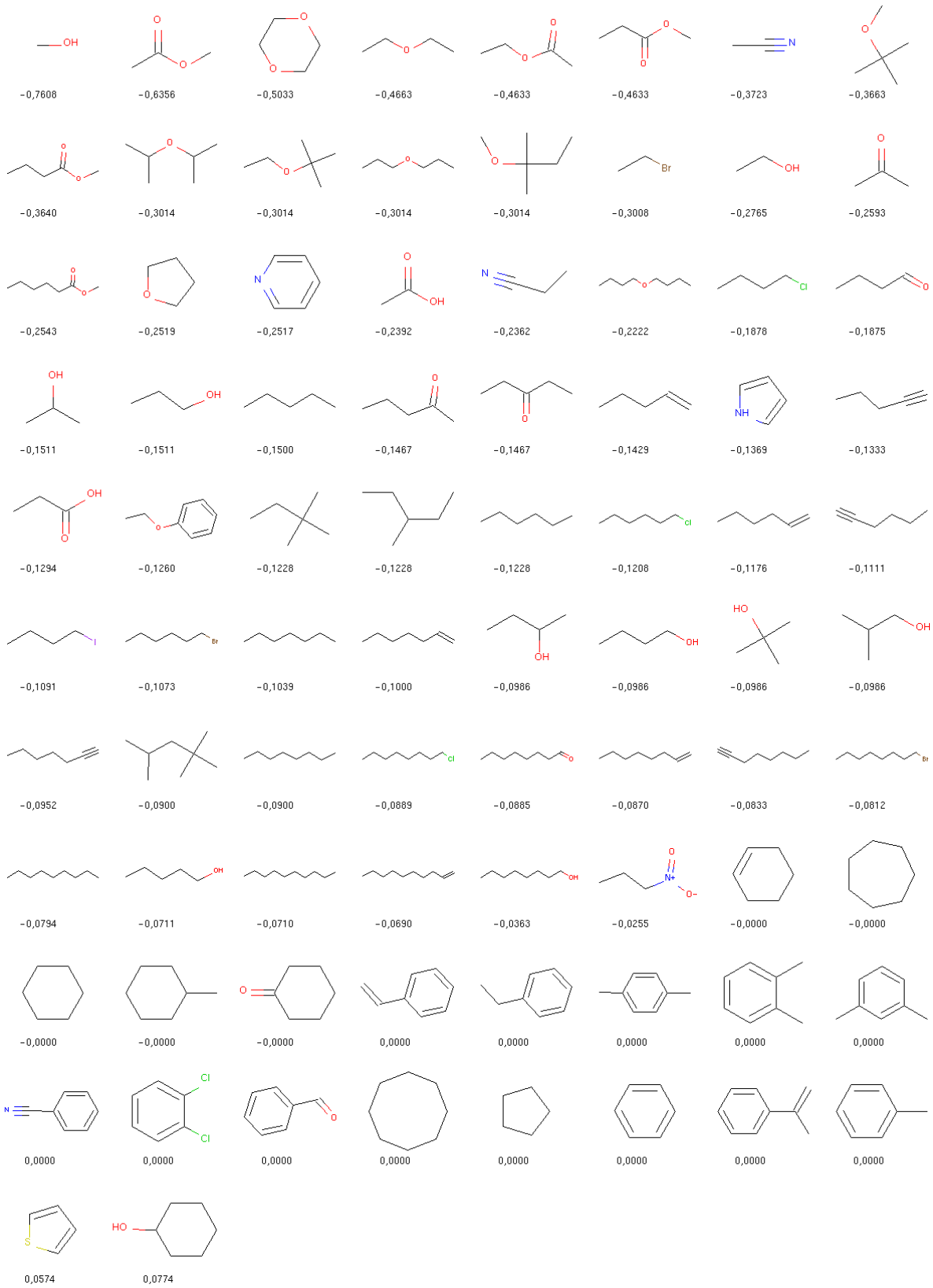
Lisa 19. Tunnuse *TMPC10* väärtused kasvavas järjestuses.



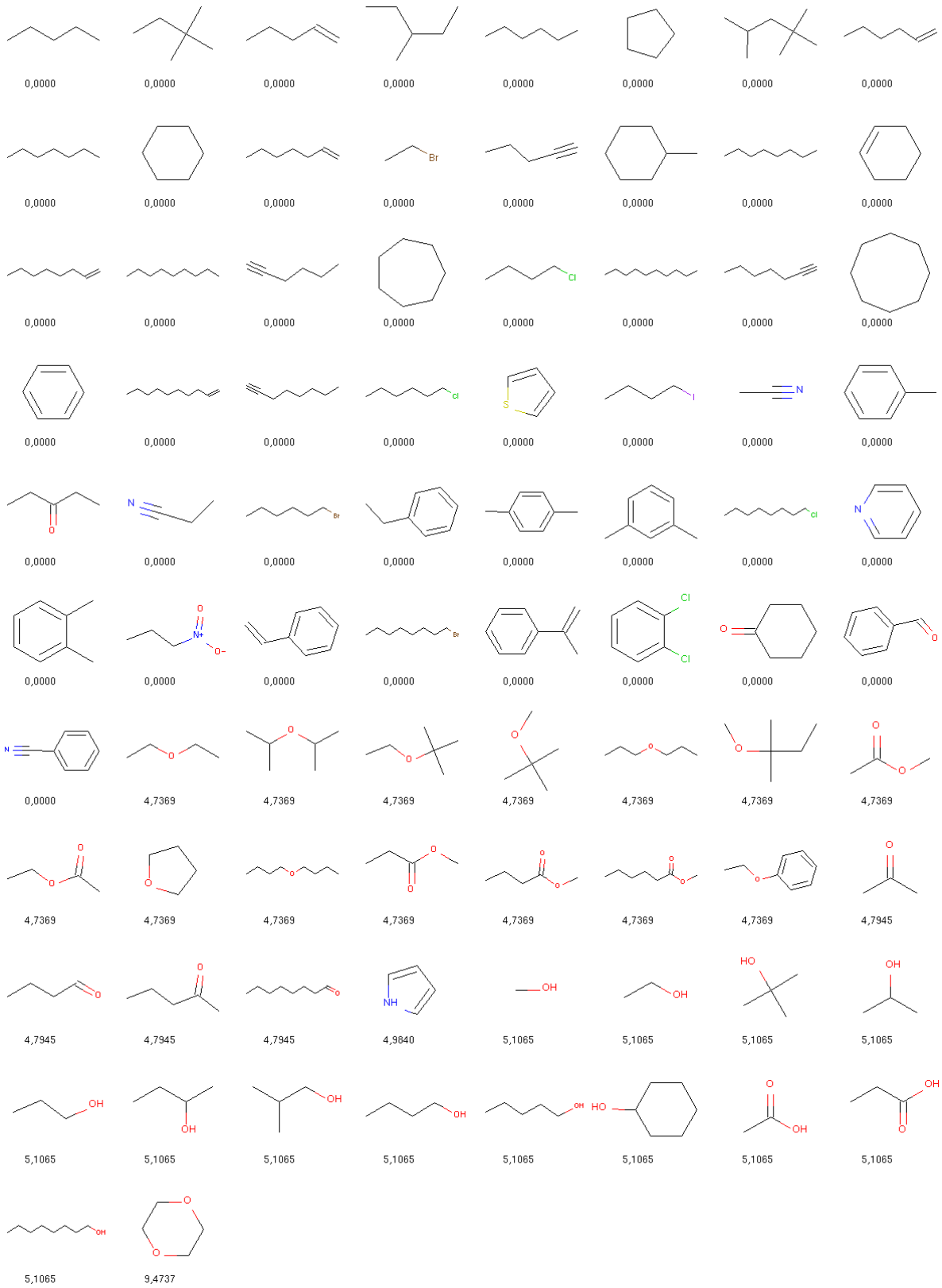
Lisa 20. Tunnuse *MID_h* väärtused kasvavas järjestuses.



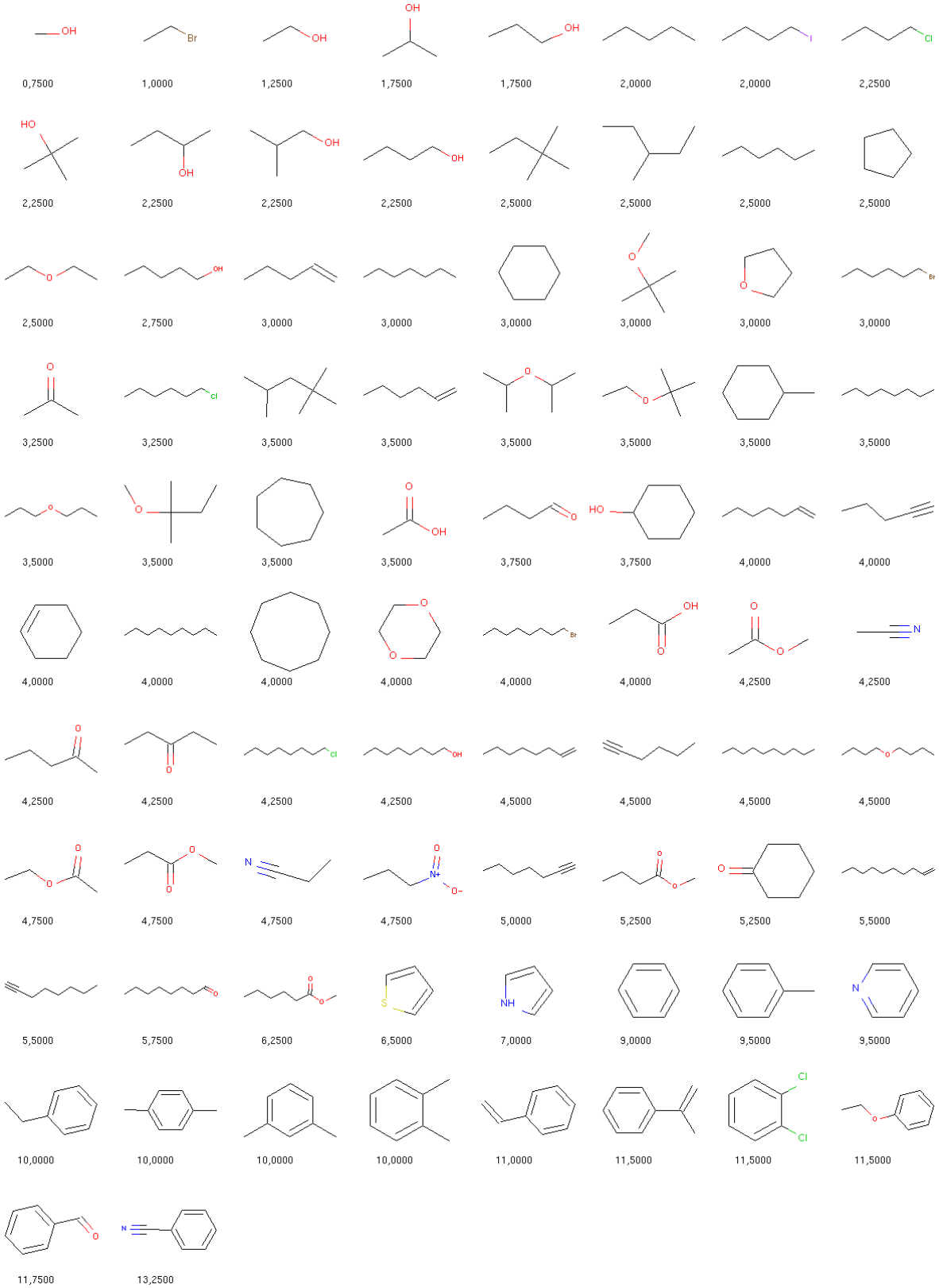
Lisa 21. Tunnuse *MATS1i* väärtused kasvavas järjestuses.



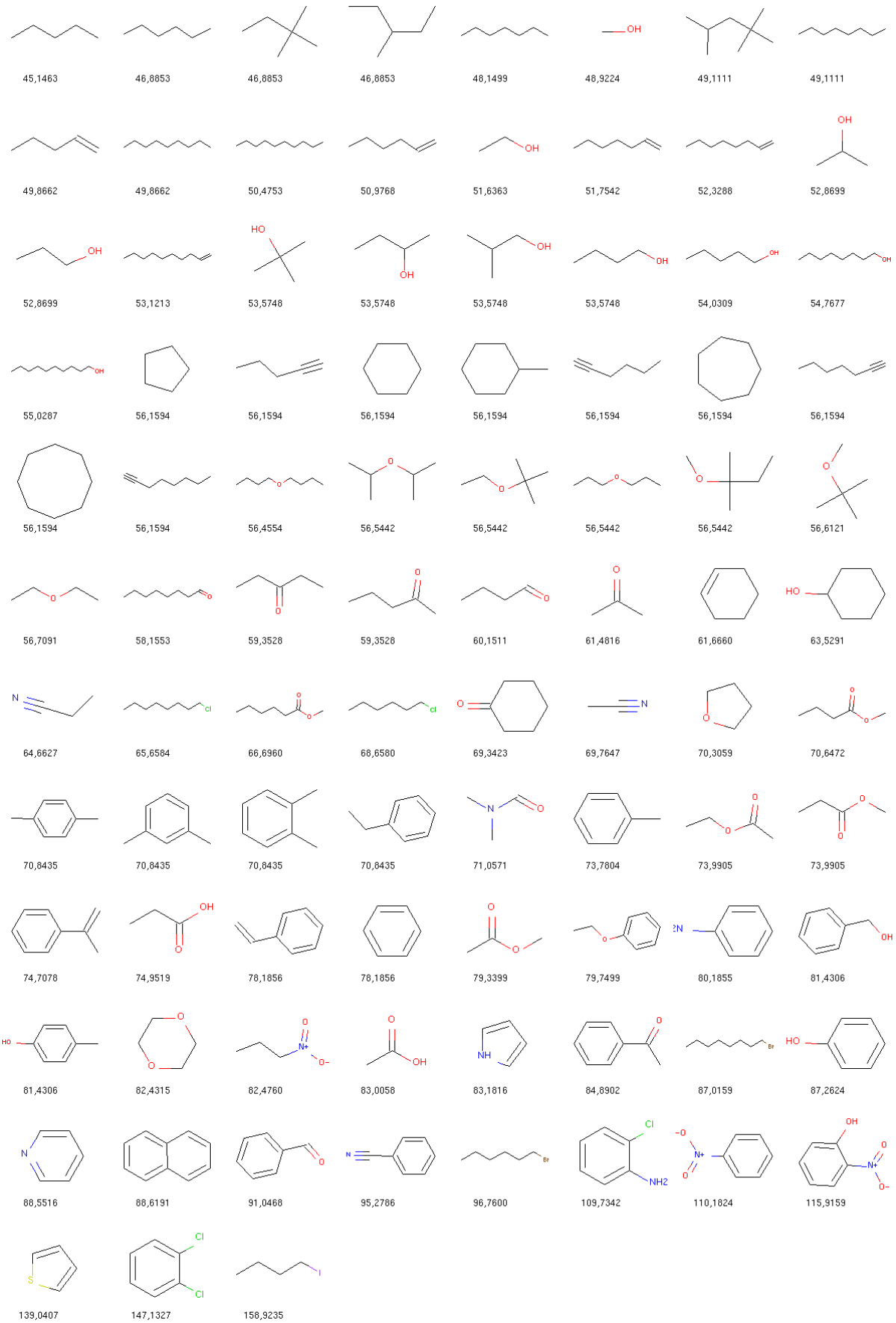
Lisa 22. Tunnuse *PEOE_VSA1* väärtused kasvavas järjestuses.



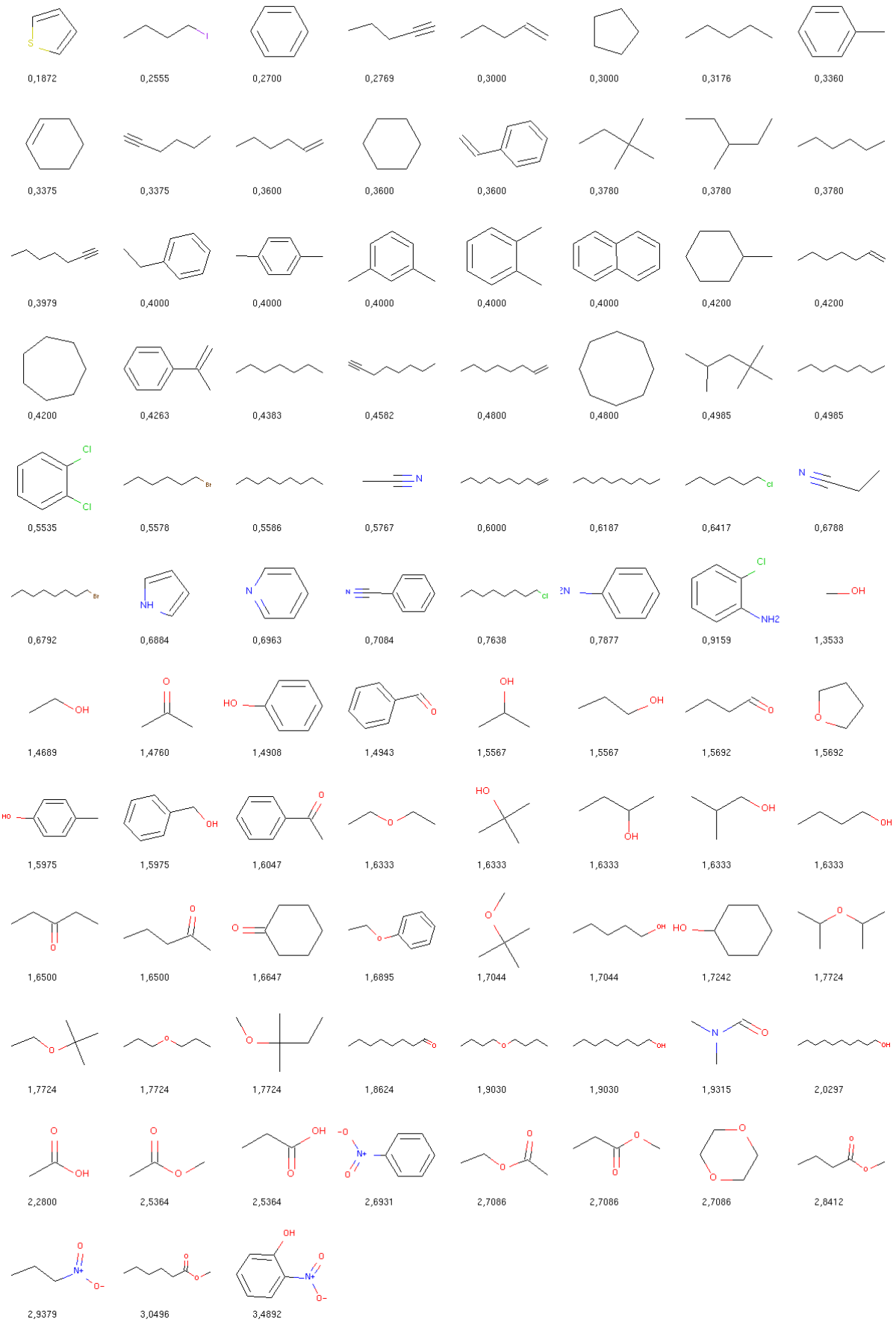
Lisa 23. Tunnuse *ETA_beta* väärtused kasvavas järjestuses.



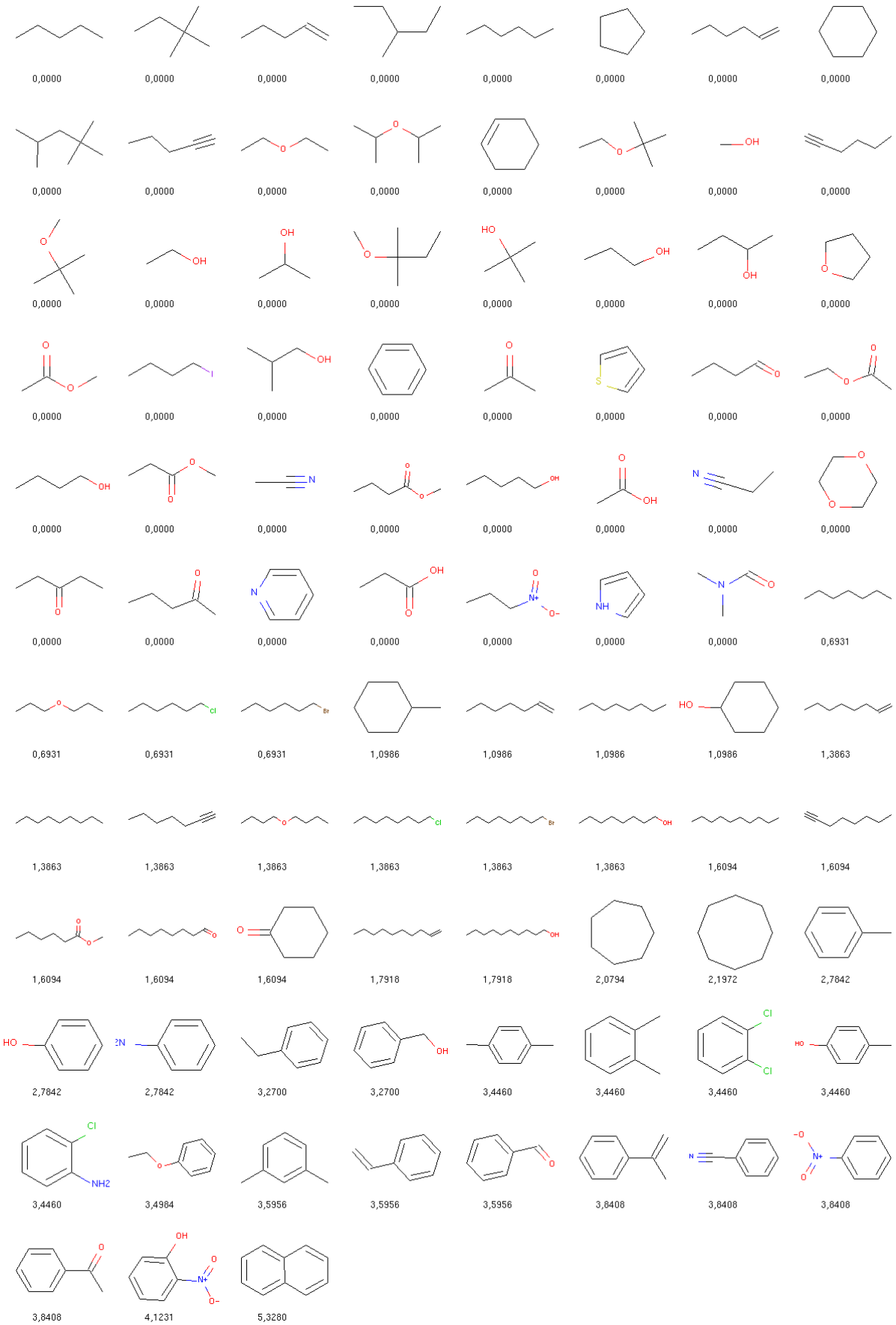
Lisa 24. Tunnuse *AATS1m* väärtused kasvavas järjestuses.



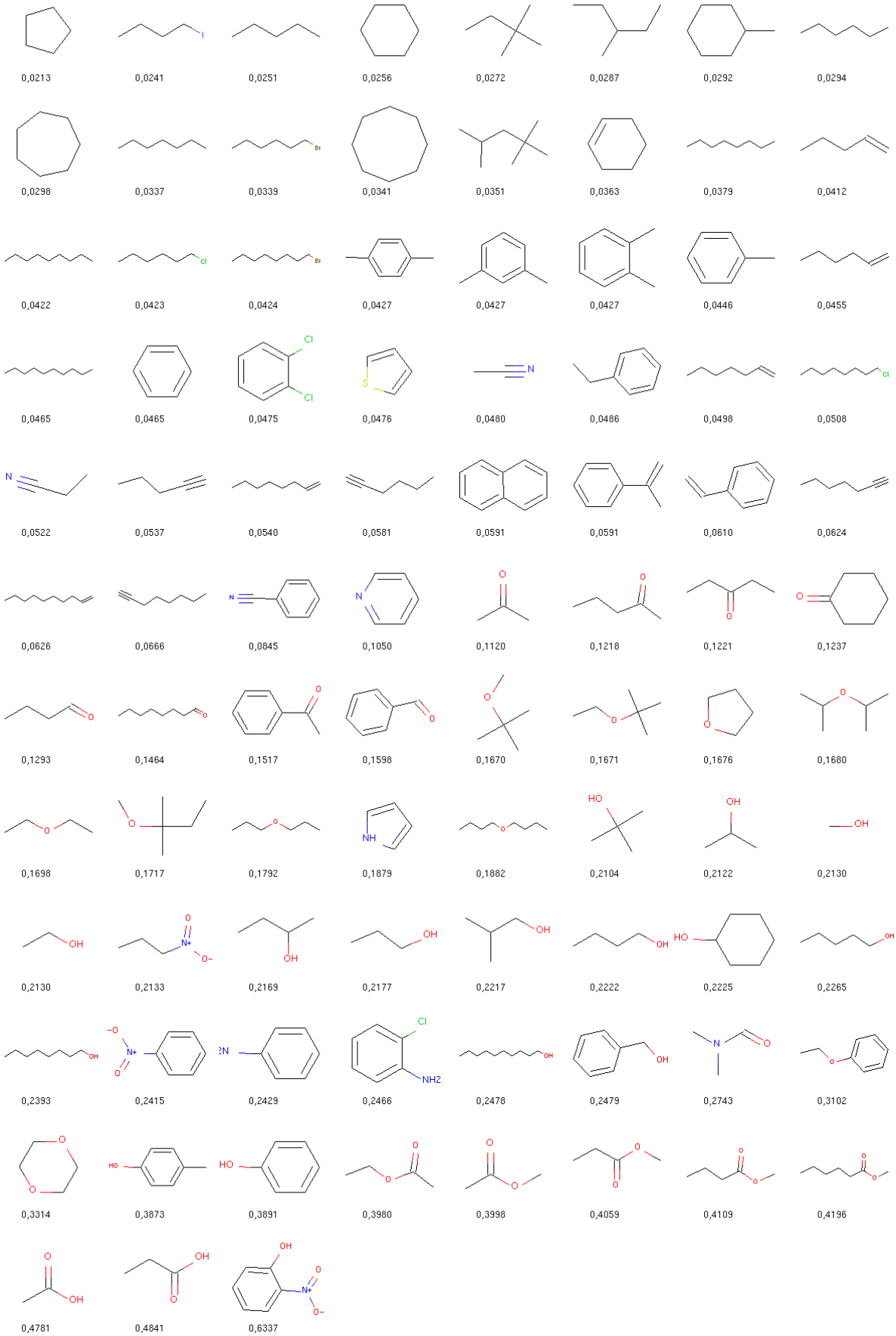
Lisa 25. Tunnuse *ATSC0are* väärtused kasvavas järjestuses.



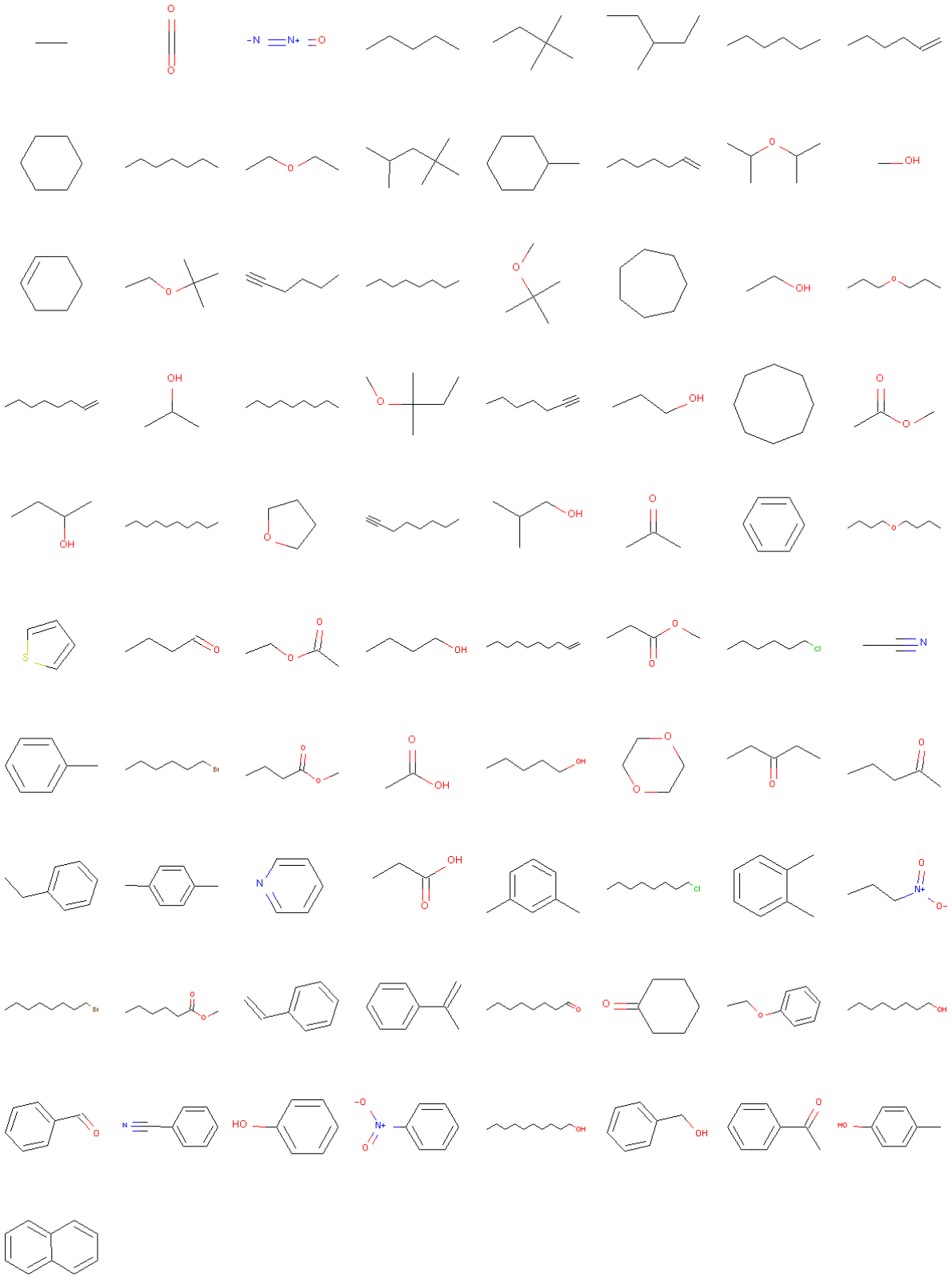
Lisa 26. Tunnuse *piPC6* väärtused kasvavas järjestuses.



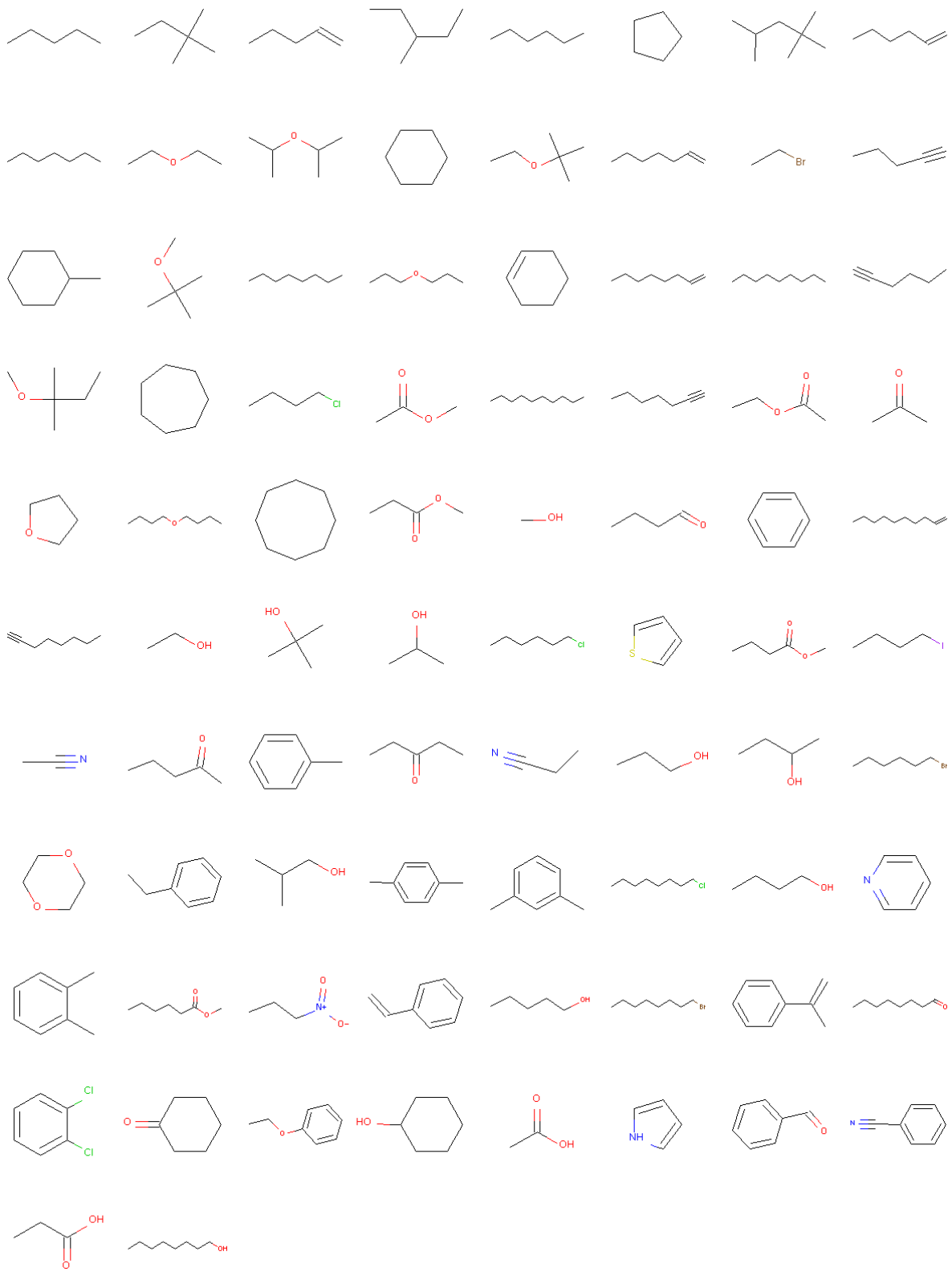
Lisa 27. Tunnuse *ATSC0c* väärtused kasvavas järjestuses.



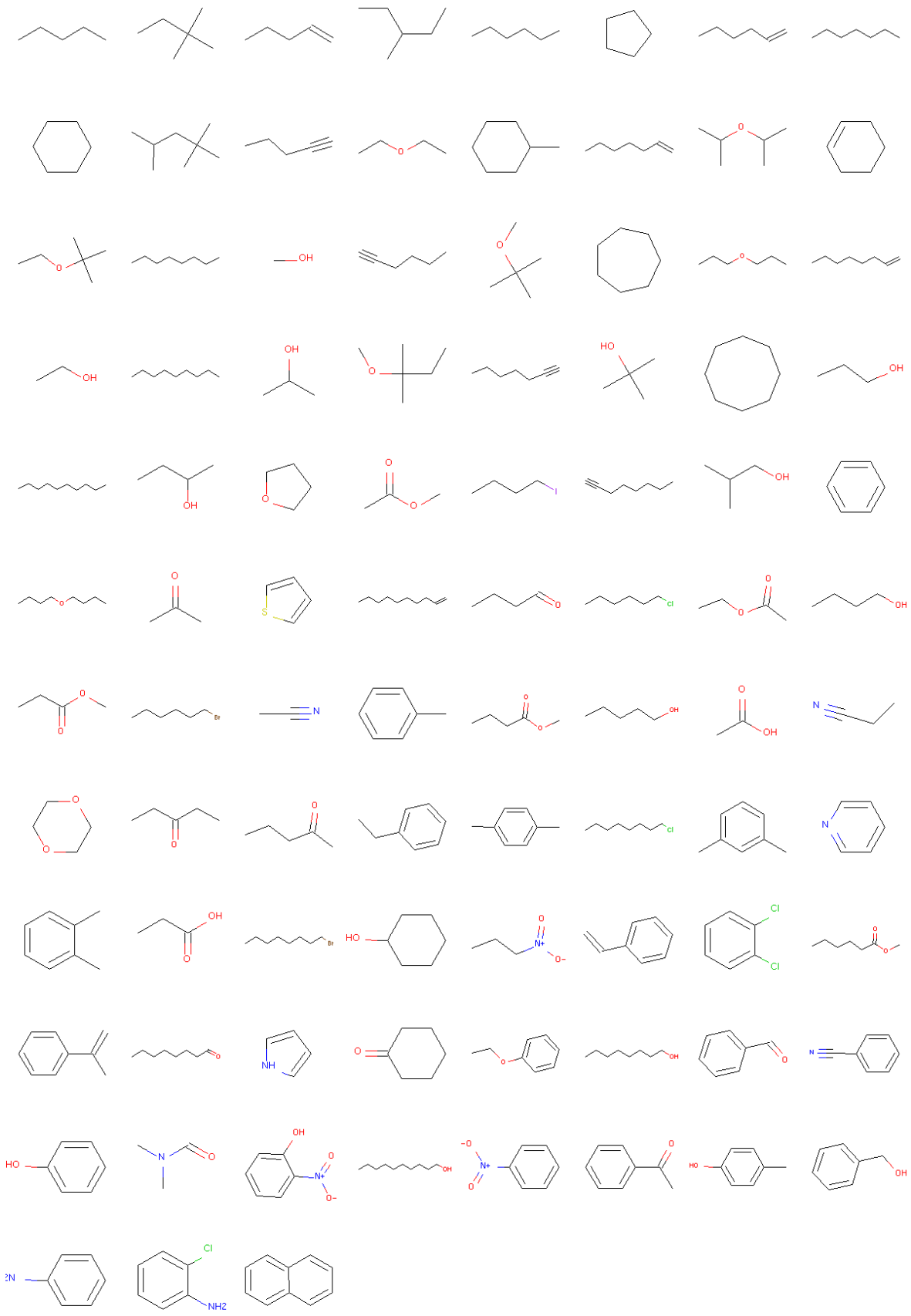
Lisa 28. log $K_{g[BMPyrr]+[FAP]-}$ väärtuste kasvav järjestus.



Lisa 29. log $K_{g[BMPyrr]+[C(CN)3]-}$ väärtuste kasvav järjestus.



Lisa 30. log $K_{g[MeoeMPyrr]+[FAP]-}$ väärtuste kasvav järjestus.



Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Karl Marti Toots

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose
   „Gaas-ioonne vedelik jaotuskoefitsiendi modelleerimine“

   mille juhendajad on Uko Maran, Sulev Sild, Jaan Leis

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, alates ***05.06.2023*** kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Karl Marti Toots
29.05.2020