

Tartu Ülikool

Füüsika-keemiateaduskond

Keemilise Füüsika Instituut

**Lahustuvuste määramine kasutades
kvantitatiivseid struktuur-omadus sõltuvusi ja
põhikomponentide analüüsi**

Magistritöö
Teoreetilise ja arvutikeema erialal

Indrek Tulp

Juhendaja: Uko Maran

Tartu, 2005

SISUKORD

Sissejuhatus.	4
Magistritöö eesmärk ja sisu	6
Kirjanduse ülevaade.	7
Lahustuvuse määramine kvantitatiivsete struktuur-omadus sõltuvustega.....	7
Lahustuvuse määramise empiirilised mudelid	8
Lahustuvuse määramise teoreetilised mudelid.....	10
Lahustuvusega seotud omaduste kirjeldamine põhikomponentide analüüsiga	11
Lahustuvust kirjeldavad eksperimentaalsed suurused	12
Modelleritavad andmed ja metoodika	15
Eksperimendi andmed	15
QSPR mudelite tuletamine	15
Põhikomponentide analüüs.....	15
Kvantkeemilised arvutused.....	16
Molekulaardeskriptorid	16
Arvutuslik eksperiment	18
Tulemused ja diskussioon.....	21
QSPR mudelite tuletamine (etapp I).....	21
Väikese maatriksi täitmine (etapp II)	23
PCA tagasiarvutus ja mudelid PCA komponentidele (etapp III)	25
Suure maatriksi täitmine (etapp IV)	29
Põhikomponentide analüüs (etapp V)	32
Kokkuvõte.	33
Summary.....	34
Kasutatud kirjandus.	35
Lisad	39
Avaldatud artikli koopia	57

Kasutatud lühendid

BMLR parim multi-lineaarne regressioon (Best Multi linear Regression)

EPA elektronpaari aktseptor (Electron Pair Acceptor)

EPD elektronpaari doonor (Electron Pair Donor)

GA geneetiline algoritm

HBA vesiniksideme aktseptor (Hydrogen Bonding Aceptor)

HBD vesiniksideme doonor (Hydrogen Bonding Donor)

HM suur maatriks (Huge matrix)

HOMO kõrgeim okupeeritud molekulaarorbital (Highest Occupied Molecular Orbital)

IUPAC (International Union of Pure and Applied Chemistry)

LSEER lineaarne solvatatsioonienergia sõltuvus (Linear Solvation Energy Relationships)

LUMO madalaim mitteokupeeritud molekulaarorbital (Lowest Unoccupied Molecular Orbital)

MC Monte Carlo

MD molekulaardünaamika (Molecular Dynamics)

MLR multi-lineaarne regressioon (Multi linear Regression)

MO molekulaarorbital (Molecular Orbital)

NN närvivõrk (Neural Network)

PCA põhikomponent analüüs (Principal Component Analysis)

QSPR kvantitatiivne struktuur-omadus sõltuvus (Quantitative Structure Property Relationship)

SA tehislõõmutamin (Simulated Annealing)

SCF kooskõlaline väli (Self-Consistent Field)

SFS iteratiivne järjestikune valik (Stepwise Forward Selection)

SM väike maatriks (Small Matrix)

SISSEJUHATUS.

Orgaaniliste ühendite lahustuvuse tundmine on olnud oluline läbi terve inimkonna ajaloo. Juba vanad kreeklased püüdsid mõista, miks vein seguneb veega aga mitte oliivõliga. Lahustuvusega vees kaasnevad olulisied probleemid keskkonnas, nagu näiteks saastatus ja erosiooni. Lahustuvus orgaanilistes solventides moodustab ühe olulise osa keemiatööstusest. Lahustuvus meditsiinis on kriitiliselt seotud bioaktiivsusega ja seeläbi ravimite efektiivsusega, biodegradatsiooniga, gaasiliste anesteetikute sobivusega, vere asendajatega, hapniku kandjatega, jne. Keemiliste ühendite toksilisus on otseselt seotud lahustuvusega. Seetõttu on lahustuvuse korrektne ennustamine ja soluut-solvent interaktsioonide mõistmine olulisel kohal keemias [1] tehnoloogias [2] ja meditsiinis [3].

Teatavas tähenduses vaadeldakse lahust kui homogeenet vedelfaasi, mis sisaldab rohkem kui ühte ainet. Üldiselt nimetatakse komponenti, mida on liias, solvendiks ja komponenti (komponente) mida on vähem soluudiks (soluutideks).

Solvendi ja soluudi vaheliste interaktsioonijõudude olemus viitab sajandeid tuntud põhimõttele, et „sarnane lahustab sarnast“ (*similia similibus solvuntur*). Paljudel juhtudel on vajalik sarnase funktsionaalgrupi olemasolu molekulis. Kui ainete vahel on keemiline sarnasus, siis kahe komponendi vaheline lahus on enamasti sarnase struktuuriga võrreldes puhaste ainetega (nagu alkohol-vesi segud [4]). „Sarnane lahustab sarnast“ reegel kehtib vaid osaliselt, kuna küllaldaselt tuntakse näiteid erinevate ühendite omavahelisest lahustuvusest. Näiteks metanol ja benseen, vesi ja *N,N*-dimetüülformamiid, aniliin ja dietüüleeter ning polüstüreen ja kloroform on omavahel täielikult segunevad toatemperatuuril. Samas lahustumatus võib esineda väga sarnastes ainetes, näiteks polüvinüülalkohol ei lahustu etanolis, atsetüülselluloos on lahustumatu etüülatsetaadis ja polüakrüülnitriil on lahustumatu akrülonitriilis [5]. Toodud ekstreemsuste vahel on terve rida võimalusi, kus kaks ainet omavahel lahustuvad limiteeritud hulgal. Vesi/dietüüleeter süsteem on selliseks näiteks, kus puhas dietüüleeter lahustab vett 15 mg/g-ni temperatuuril 25 °C, samas kui vesi lahustab dietüületrit 60 mg/g. Kui üks solventidest on suures ülekaalus, saadakse homogeenne lahus, kui aga suhe on üle lahustuvuse piiri, tekib kahefaasiline süsteem. Samuti on „sarnane lahustab sarnast“ reegli kinnituseks kolme dimensionaalse delokaliseeritud 60π -elektronidega süsteemi, fullereen (C_{60}) erinev lahustuvus metanolis ($s = 0,01$ mg/mL) ja 1-kloronaftaleenis ($s = 50$ mg/mL) [6].

Siiski, olulisemad kui „sarnane lahustab sarnast“ reegel, on intermolekulaarsed interaktsioonid solvendi ja soluudi molekulide vahel, mis määravad ära ainete omavahelise

lahustuvuse. Ühend A lahustub solvendis B ainult juhul, kui puhta komponendi intermolekulaarsed külgetõmbejõud K_{AA} ja K_{BB} on ületatud K_{AB} jõudude poolt lahuses [1, 7].

Solvendi ja soluudi vaheliste interaktsioonijõudude summat saab siduda nn. A ja B polaarsusega. Nimetades suure interaktsiooniga ühendeid $A \cdots A$ või $B \cdots B$ vastavalt polaarseks ning väikse interaktsiooniga mittepolaarseks.

Intermolekulaarsed jõud on tuntud ka nn. van der Waals'i jõududena, kuna van der Waals täheldas neid reaalsete gaaside mitteideaalse käitumise põhjusena. Intermolekulaarsed jõud on enamasti klassifitseeritud kahte kategooriasse. Esimene kategooria sisaldab endas orienteerimis-, indutseerimis- ja dispersioonjõude, mis on mittespetsiifilised ja neid ei saa täielikult küllastada (nagu ka Columb'i jõude kahe iooni vahel). Teine grupp sisaldab vesiniksidemest tingitud jõude ja elektronpaari doonor-aktseptor laengu ülekande jõude, mis võivad küllastuda ja viitavad stõhhiomeetrilistele molekulaarsetele ühenditele. Reichardt jagab intermolekulaarsed jõud laiali järgnevalt: a) ioon-dipooljõud, b) dipool-dipooljõud, c) dipool-indutseeritud dipooljõud, d) hetkelised dipoolindutseeritud dipooljõud, e) vesinikside, f) elektronpaar doonor / elektronpaar aktseptor (EPD/EPA) interaktsioonid, ja g) solvofoobilised interaktsioonid [1].

Lahustuvuse, kui komplekse omaduse uurimine, on viinud mitmete empiiriliste ja füüsikaliste mudelite tuletamiseni. Hästi tuntud on lahustuvuse lineaarne vabaenergia sõltuvus (LSER), mis põhimõtteliselt koondab kaks peamist protsessi, esiteks mittespetsiifilised dielektrilised polarisatsiooniefektid ja teiseks spetsiifilised semikeemilised sidemete (nagu vesiniksideme) tekke solvendi ja soluudi molekulide vahel. LSER mudelid põhinevad empiirilistel andmetel, erinevatel solvendiefektidel. Teoreetiliselt täendatud solvendi efektide käsitus on viinud füüsikaliste mudelite tuletuseni, kus solvatatsiooni protsesse vaadeldakse järgnevalt: 1) augu teke solvendi keskkonda, 2) soluudi asetus auku ja elektrostaatiline ja orjentatsiooniline polarisatsioon ning dispersiooni interaktsioonid soluudi ja solvendi molekulide vahel, 3) vesiniksidemete või komplekside teke soluudi ja solvendi molekulide vahel. Augu tekke efekte kirjeldatakse enamasti statistilise mehaanika meetoditega, polarisatsiooni ja dispersiooni efekte statistilise ja kvantmehaanika teooriaga ning vesiniksidemete teket kvantmehaanika meetoditega [8].

MAGISTRITÖÖ EESMÄRK JA SISU

Magistritöö eesmärk on ennustada suure hulga soluutide lahustuvus solventides kasutades koos QSPR-i ja PCA meetodeid, välja töötada vastav metodoloogiline lahend ning täita puuduvate väärtustega $154 \text{ solvendi} \times 397 \text{ soluudi}$ andmematriks. Kirjanduse ülevaade annab kättesaadava kirjanudse baasil laiendatud kokkuvõtte mõlema eespool mainitud meetodi kasutamisest lahustuvuse määramisel. Metoodika osa annab ülevaate eksperimendi andmetest millele uurimustöö tugineb ja kasutatud meetoditest. Arvutusliku eksperimendi peatüki all on kirjeldatud kasutatud lahendust ja analüüsi käiku ning diskussiooni osa teeb kokkuvõtte töö tulemustest.

Magistritöö toetub publitseerimisel olevate teadusartiklite seeria kolmandal artiklil, mis on ära toodud ka magistritöö lisas:

Katritzky, A. R.; Tulp, I.; Fara, D. C.; Lauria, A.; Maran, U.; Acree, W. E. Jr. General Treatment of Solubility. 3. Principal Component Analysis (PCA) of Solubilities of Diverse Solutes in Diverse Solvents. *J. Chem. Inf. Model.* **2005**, *45*, (in press).

KIRJANDUSE ÜLEVAADE.

Lahustuvuse määramine kvantitatiivsete struktuur-omadus sõltuvustega

Sõltuvalt mudelites kasutatavatest deskriptoritest klassifitseerisid Yalkowsky ja Banerjee erinevad lahendused vees lahustuvuse määramiseks kolme kategooriasse [9, 10]:

- a) korelatsioon eksperimentaalselt määratud füsikokeemiliste väärtustega või eksperimendist tuletatavate empiiriliste parameetritega;
- b) grupi (fragmendi) kontributsioonidel baseeruv korelatsioon;
- c) korelatsioon otseselt molekuli struktuurist arvatud parameetritega e. molekulaardeskriptoritega.

Samasuguse klassifikatsiooni võib laiendada ka kõigile teistele solventidele peale vee.

Esimese kategooria on ajalooliselt kõige rohkem uuritu. Siia alla võib tuua tööd, kus vastavaid meetodeid ja empiirilistele mõõtmistele baseeruvaid deskriptoreid on välja töötanud Katritzky *et al.* [11], Koppel ja Palm [12], Kamlet ja Taft [13], Krygowsky ja Fawcett [14], Sawin *et al.* [15], Mayer [16], Dougherty [17], jpt.. Suurimat edu on selles kategoorias saavutanud LSER meetod, mida on hiljem täiendatud Abraham'i [18] poolt, ja millel allpool ka pikemalt peatume.

Teise kategooriasse kuuluvaid grupi (fragmendi) kontributsiooni meetodeid on samuti kasutatud lahustuvuse ennustamiseks, kuid oluliselt vähemal määral. Kuigi see tehnoloogia võib anda häid tulemusi, sisaldab ta vähe informatsiooni molekuli struktuuri ja lahustuvuse protsessi vahelise füüsikalise olemuse mõistmiseks. Ülevaade kuidas antud meetodid on kasutatud lahustuvuse määramisel on atud hiljuti ilmunud ülevaateartiklis [3] ja uurimustöös [19].

Kolmas kategooria võtab kokku korelatsioonid otseselt molekuli struktuurist tulenevate deskriptoritega. Need deskriptorid kannavad informatsiooni ühendi ülesehitusest (konstitutsioonilised deskriptorid), topoloogiast, geomeetriast, elektronstruktuurist (laengujaotusest) ja energeetilistest partameetritest [20]. Molekulaardeskriptorid sõltuvad täpselt ühendi struktuurist, neid saab arvutada mistahes struktuurile ja see võimaldab neid ulatuslikult kasutada [3]. Molekulaardeskriptorid ja nednest tuletatud QSPR mudelid on viimastel aastatel kogunud palju populaarsust ja neid on kasutatud koos erinevate QSPR mudelite tuletamise algoritmidega, mistõttu nedel ka pikemalt peatume.

Lahustuvuse määramise empiirilised mudelid

Kirjanduses tuntakse mitmeid soluut/solvent interaktsioonide kirjeldamise ja lahustuvuse määramise empiirilisi mitmeparameetrilisi mudeleid. Neist tuntumateks on Koppel-Palm'i (KP) [1, 12, 21] ja Kamlet-Abboud-Taft'i (KAT) [1, 13 22, 23, 24, 25] võrrandid. Esimene neist, KP mudel (võrrand 1), kirjeldab nelja-parameetrilist võrrandit, mis seob uuritava omaduse A kahe mittespetsiifilise (Y ja P) ja kahe spetsiifilise (E ja B) solvendi karakteristikuga.

$$A = A_0 + yY + pP + eE + bB, \quad (1)$$

kus A on solvendist sõltuva füsikokeemilise omaduse väärtus antud solvendi jaoks, A_0 on antud omaduse statistiline väärtus gaasifaasi või inertsesse solvendi suhtes. Mittespetsiifilised parameetrid Y ja P määravad vastavalt solvendi polaarsuse ja polariseeritavuse, E ja B on spetsiifilised parameetrid, mis määravad solvendi Lewise happelisuse (elektrofiilne solvatatsioonijõud) ja Lewise aluselisuse (nukleofiilne solvatatsioonijõud) ning parameetrid y, p, e ja b on vastavad regressiooni koefitsiendid.

Teine, KAT mudel, on sarnane eelpool toodud KP võrrandile 1. Samuti võetakse eraldi arvesse mittespetsiifilisi ja spetsiifilisi soluut/solvent interaktsioone ning ülejäänu jagatakse solvendi Lewise happe ja aluse interaktsioonideks. Kasutades solvatokromaatilisi solvendi parameetreid, saab mitmeparameetrilise võrrandi kirjutada välja kujul, mida tuntakse laiemalt kui LSER mudelit:

$$A = A_0 + s(\pi^* + d\delta) + a\alpha + b\beta, \quad (2)$$

kus π^* on solvendi dipolaarsuse/polariseeritavuse indeks, mis määrab ära solvendi võime stabiliseerida laengut või dipooli tänu ta dielektrilisele vastasmõjule. δ on polariseeritavuse parandusliige, mis on 0 mitteklореeritud alifaatsete solventide puhul, 0,5 polükloreeritud alifaatsete ja 1 aromaatssete solventide puhul. Parameeter α ja β kirjeldavad vastavalt vesiniksideme doonor-happesust ja aktseptor-aluselisust ning vastavad KT võrrandi (1) liikmetele E ja B . s, d, a ja b on regressiooni koefitsiendid. Võrrandit 2 on täiustatud edukalt tuues sisse kaks lisa solvendi parameetrit [1, 13]:

$$A = A_0 + s(\pi^* + d\delta) + a\alpha + b\beta + h\delta_H^2 + e\xi, \quad (3)$$

kus δ_H^2 on Hildebrand'i lahustuvuse parameeter ruudus ja vastab kohesiooni rõhule c (võrrand 4) [1, 26, 27, 28, 29, 30, 31], mis kirjeldab intermolekulaarset solvent/solvent interaktsiooni energiat, ja ξ on koordinatiivne kovalentsuse parameeter, mis koreleerub nn. perekond-sõltuvate aluseliste omadustega.

$$c = \delta_H^2 = \frac{\Delta U_v}{V_m} = \frac{\Delta H_v - RT}{V_m}, \quad (4)$$

kus ΔU_v ja ΔH_v on vastavalt solvendi nullrõhu aurustumise energia ja entalpia ning V_m on solvendi molaarruumala.

Kamlet, Taft, Abraham *et al.* modifitseerisid võrrandit 3 et koreleerida lahustuvust solvendi omadustega [1, 32, 33].

$$A = A_0 + A'(\delta_H^2)_1(V_2/100) + B\pi_1^*\pi_2^* + C\alpha_1\beta_2 + D\beta_1\alpha_2, \quad (5)$$

kus alaindeks 1 viitab solvendile ja alaindeks 2 soluudile ning A' , B , C ja D on regressiooni koefitsiendid, mis vastavad augu tekke liikmele, dipolaarsuse/polariseeritavuse liikmele ja vesiniksideme liikmetele, kus vesinikside tekib HBD solvendi ja HBA soluudi vahel (määratud kui α_1 ja β_2) ning vastupidi HBA solvendi ja HBD soluudi vahel (määratud kui β_1 ja α_2). V_2 on soluudi molaarruumala. Kusjuures $(\delta_H^2)_1$ kirjeldab solvendi panust augu tekkesse ja $V_2/100$ väljendab soluudi panust augu tekkesse. V_2 asemel on kasutatud $V_2/100$ et vastav parameeter kataks üldiselt sama arvulist vahemikku kui teised sõltumatud muutujad α , β ja π^* (umbes 0,0...1,0).

Võrrandit 5 on võimalik lihtsustada:

- a) kui tegeleme erinevate solvendite efektidega ühes kindlas soluudis, siis soluudiga seotud faktord võib summeerida regressiooni koefitsienti ning saame

$$A = A_0 + h(\delta_H^2)_1 + s\pi_1^* + a\alpha_1 + b\beta_1 \quad (6)$$

- b) või kui tegeleme erinevate soluutide lahustuvuse või muude omadustega ühes kindlas solvendis, siis saame

$$A = A_0 + m(V_2/100) + s\pi_2^* + a\alpha_2 + b\beta_2 \quad (7)$$

LSER-id, nagu KAT võrrand (2) ja nende modifikatsioonid, võib summeerida üldisesse vormi (võrrand 9):

$$A = A_0 + \text{maht/augu teke} + \text{dipolaarsus/polariseeritavus} + \text{vesiniksidemed}, \quad (8)$$

kus A ja A_0 on vastavalt solvendist sõltuv füsikokeemiline omadus vaadeldavas solvendis ja gaasifaasis (või inertses solvendis). Nagu eelpool mainitud, dipolaarsus/polariseeritavuse liige kirjeldab soluut/solvent dipool- ja indutseeritud dipoolinteraktsioone. Vesiniksideme liige kirjeldab interaktsioone, kui tekivad vesiniksidemega HBA soluut/HBD solvent ja/või HBD soluut/HBA solvent kompleksid. Mahu või augu tekke on liige, mis on vajalik kohesiivse solvent/solvent interaktsioonide ületamiseks, et saada õõnsused soluudi molekulide jaoks.

LSER meetodit on edukalt kasutatud kirjeldamaks mitmeid solvendi efekte ja lahustuvust. Pikema ülevaate meetodi rakendamisest lahustuvuse määramisel leiab kahest ülevaate artiklist [2, 3] ja kahest eelmisest sama seeria artiklist [34, 35].

Solvendi efektide mitmeparametiline käsitus on kritiseeritud vähemalt kolmest vaatevinklist. Esiteks, solvendi efektide jagamine erinevateks liidetavateks osadeks on omavoliline, kuna erinevad soluut/solvent interaktsioonide mehhanismid võivad koostoimida teineteisest sõltumatult. Teiseks, parimate parameetrite korrektne valik erinevat tüüpi soluut/solvent interaktsioonide kirjeldamiseks on kriitiline, tänu vastavate empiiriliste solvendi parameetrite keerukusele ning nende nõrkusele vastata enam kui ühele mitmest solvendi polaarsuse osast. Kolmandaks, et saada statistiliselt korrektsel viisil mitmeparametrist regressiooni on tarvis piisaval hulgal eksperimetaalseid andmepunkte, et oleks võimalik ekstrapoleerides või interpoleerides hinnata solvendi efekti või lahutusvust [1].

Kokkuvõtvalt LSER mudel kirjeldab erinevaid karakteristikuid: solvent/soluut polariseeritavust, dipolaarsust, suurust, vesiniksideme happelisust ja aluselisust. Antud meetodi tugevuseks on, et ta koondab vastavad karakteristikud ühte mudelisse, koostades seega tugeva baasi arendamiseks diskussiooni soluut-solvent interaktsioonide üle lahustuvuse määramisel. LSER korelatsiooni võrrandit võib interpreteerida osade kaupa hästi tuntud keemia põhitõdedel. Kahjuks ei saa LSER-i kasutada *a priori* ennustamiseks, kuna deskriptorid tulenevad eksperimentaalsetest mõõtmistest ning muudab raskeks nende leidmise suurte andmebaaside ja laia ühendite hulga puhul. Samuti saadud korelatsioon ei sisalda otsest informatsiooni molekuli struktuuri kohta. Seega on raske kirjeldada kuidas molekuli struktuur mõjutab antud omadust. Kuid samas omavad LSER mudelid enamasti väga head ennustamiskvaliteeti.

Lahustuvuse määramise teoreetilised mudelid

Antud käsitluses peame teoreetiliste mudelite all silmas neid QSPR-e mis on tuletatud kasutades (teoreetilisi) molekulaar deskriptoreid, arvutatuna molekuli struktuurist, ja mudelid on tuletatud kasutades selleks spetsiaalseid deskriptorite valiku meetodeid. MLR on vaieldamatult populaarseim meetod QSPR-i kaudu lahustuvuse modelleerimiseks. Viimasel kümnendil on ka kasutatud palju erinevaid tehislikke NN tehnikaid. Samuti on arendatud erinevaid deskriptorite valiku meetodeid konjugeeritult MLR-i ja NN-iga nagu SFS, GA, ja SA lähenemised. Põhjaliku ülevaate QSPR mudelitest lahustuvuse määramisel annavad kaks hiljutist ülevaate artiklit [2, 3].

Nii soluudi kui solvendi struktuur määravad ära interaktsioonid, mis kaasnevad lahustuvuse protsessis. Järelikult võib soluut-solvent interaktsioonide kvantitatiivse analüüsi väljendada üldise valemiga (9).

$$Lahustuvus = C_0 + \sum C_{el} D_{el-solvent} D_{el-soluut} + \sum C_{disp} D_{disp-solvent} D_{disp-soluut} + \sum C_{auk} D_{auk-solvent} D_{auk-soluut} + \sum C_{VS} D_{VS-solvent} D_{VS-soluut}, \quad (9)$$

kus,

- a) C_0 , C_{el} , C_{disp} , C_{auk} , ja C_{VS} on vabaliige ja üldised koefitsiendid elektrostaatiliselle interaktsioonile, dispersiooni põhjustavale interaktsioonile, solvendis augu moodustamisele, ja vesiniksidemete moodustumisele;
 - b) $D_{el/disp/auk/VS-solvent}$ on sobivad molekulaardeskriptorid, mis kirjeldavad solvendi struktuuri mõju lahustuvusele;
 - c) $D_{el/disp/auk/VS-soluut}$ on deskriptorid mis peegeldavad soluutide struktuuri lahustumisel.
- Summeerimine näitab, et igas liikmes võib esineda mitu deskriptorit, mis kirjeldavad sama tüüpi interaktsiooni. Seeriates, kus soluut on konstantne, võib vastavad deskriptorid liita vastavate koefitsientidega ja sellisel juhul on lahustuvuse kirjeldamine määratletud ainult solvendi struktuurist tulenevate omadustega. Sama kehtib vastupidisel juhul kui solvent on konstantne, siis on lahustuvus kirjeldatud soluudi struktuurist lähtuvate omadustega.

Järelikult saab lahustuvust modelleerida kahel viisil. Esimene, ka kõige levinum vatenurk, on kindlas solvendis lahustuvuse uurimine läbi soluutide seeria. Kõige enam on uuritud erinevate ainete lahustuvust vees. Teine võimalus on uurida kindla soluudi lahustuvust erinevates solventides. Põhjalikult on neid mõlemaid viise kirjeldanud koos vastava kirjanduse ülevaatega kahes eelmises selle seeria artiklis [34, 35].

Lahustuvusega seotud omaduste kirjeldamine põhikomponentide analüüsiga

Põhikomponentide analüüsiga on peamsielt uuritud solvendid ja nende erinevaid omadusi. Nii näiteks uurisid Katritzky *et al.* 40 erinevat polaarsuse skaalat, kus polaarsuse skaalad olid muutujad (deskriptorid) 40 solvendi jaoks [36]. Diagonaliseerimisprotseduur 40×40 ruutmaatriksil jagas solvendid viide gruppi ja solvendi skaalad seitsmesse gruppi, järgides solvendi tüüpe ja polaarsuse skaalade füüsikalist tähendust. PCA-d on kasutatud ka teiste solvatatsiooniga seotud füüsiko-keemiliste omaduste klassifitseerimiseks. Fawcett ja Krygowski uurisid lahuste termodünaamilisi entalpiaid [14]. Cramer analüüsis 114 kemikaali vees lahustuvust koos viie teise eksperimentaalse suurusega [37]. Chastrette *et*

al. teostas PCA analüüsi 83 solvendi kuuel empiirilisel solvendiskaalal ja poolempiirilisel saadud HOMO ja LUMO energiatel [38].

Eelpool mainitud uurimused käitlevad solvendi omadusi läbi erinevate optiliste ja keemiliste aktiivsuste (solvendi skaalad), mis üldiselt võttes ei ole seotud otseselt soluut/solvent interaktsioonidega. Seega oleks ideaalseks skaalaks lahustuvuste põhikomponentide analüüsil lahustuvuse vabaenergia või veel üldisemalt tasakaalulised lahustuvused üksteises. Meile on teada vaid kaks põhikomponent analüüsi, mis on tehtud lahustuvuste analüüsiks. Dunn *et al.* analüüsis 6 solvendi $\times$ 50 soluudi andmematriksit ja leidis kaks põhikomponenti, kus esimene koreleerus hästi isotroopse pinna pindalaga [39]. Hilisem gaas-vedelik jaotuskoefitsientide uurimus Reta *et al.* poolt 11 soluudi $\times$ 67 solvendi matrikis andis samuti tulemuseks kaks olulist põhikomponenti [40]. Kaks solventi võeti arvesse kui „testi faktorid“ kuna nad koreleerusid enamike eksperimentaalsete andmetega.

Lahustuvust kirjeldavad eksperimentaalsed suurused

Lahustuvust saab defineerida kahel peamisel viisil: vedelike ja tahkete ainete lahustuvus või gaaside ja aurude lahustuvus. Esimene neist, on defineeritud kui soluudi kontsentratsioon (lahustuva aine mooli või kaalu ühikut lahuse kaalu või ruumala kohta) solvendi faasis. Teine on defineeritud kui ühendi kontsentratsioonide suhe lahuses ja gaasifaasis.

Tegelikuses kasutatakse lahustuvuste esitamiseks mitmeid eksperimentaalseid suurusi. Fogg *et al.* [41] toob välja peamised IUPAC-i nomenklatuuris kasutatavat lahustuvuse esitusviisid, mis lühidalt oleks järgmised:

1) *moolosa*

$$x_1 = \frac{n_1}{\sum_{s=1}^c n_s}, \quad (10)$$

kus n_1 on antud aine hulk, n_s on s aine hulk ja c on ainete arv süsteemis;

2) *massiosa*

$$w_1 = \frac{g_1}{\sum_{s=1}^c g_s}, \quad (11)$$

kus g_1 on antud aine mass, g_s on s aine mass ja c on ainete arv süsteemis;

3) *molaalsus* [mol kg⁻¹]

$$m_1 = \frac{n_1}{n_2 M_2}, \quad (12)$$

kus n_1 on antud aine hulk, n_2 on solvendi hulk ja M_2 on solvendi molaarmass;

4) *molaarsus* [mol m⁻³]

$$c_1 = \frac{n_1}{V}, \quad (13)$$

kus n_1 on antud aine hulk ja V on lahuse ruumala;

5) *masskontsentratsioon* [kg m⁻³]

$$\rho_1 = \frac{g_1}{V}, \quad (14)$$

kus g_1 on antud aine mass ja V on lahuse ruumala;

6) *moolsuhe* [dimensioonitu]

$$r_{A,B} = \frac{n_1}{n_2}, \quad (15)$$

kus n_1 ja n_2 on vastavalt erinevate ainete hulgad;

7) *ioontugevus* (I_m – molaalsuse baasil; I_c – kontsentratsiooni baasil)

$$I_m = \frac{1}{2} \sum_i m_i z_i^2; \quad I_c = \frac{1}{2} \sum_i c_i z_i^2, \quad (16)$$

kus m_i ja c_i on vastavalt i -nda iooni molaalsus ja molaarsus, z_i tema laeng;

8) *Bunsen'i koefitsient* [dimensioonitu]

$$\alpha = \frac{V_1^\circ}{V_2^*} = \left(\frac{V_1}{V_2^*} \right) \left(\frac{T^\circ}{T} \right), \quad (17)$$

kus V_1° on lahustuva gaasi ruumala V_1 viidud temperatuurile $T^\circ = 273,15$ K ja rõhule $p^\circ = 1$ bar (standardtingimustele) ning mis on absorbeeritud puhta solvendi poolt ruumalaga V_2^* antud temperatuuril T ;

9) *Kuenen'i koefitsient* [m³ kg⁻¹]

$$S = \frac{V_1 T^\circ}{g_2 T} = \frac{\alpha V_{m,2}}{M_2}, \quad (18)$$

kus g_2 ja M_2 on vastavalt solvendi mass ja molaarmass ning $V_{m,2}$ solvendi molekulaarruumala;

10) *Ostwald'i koefitsient* [dimensioonitu]

$$L = \frac{V_1}{V_2^*} = \left(\frac{\alpha T}{T^\circ} \right) \left(\frac{p^\circ}{p} \right); \quad (19)$$

11) *Absorptsiooni koefitsient* [dimensioonitu]

$$\beta = \alpha \left(1 - \frac{p_2}{p^\circ} \right), \quad (20)$$

kus p_2 on solvendi aururõhk;

12) *Henry konstant* [Pa]

$$K_H = \lim_{x_1 \rightarrow 0} \left(\frac{p_1}{x_1} \right), \quad (21)$$

samuti defineeritakse Henry konstanti ka järgnevalt:

$$K_2 = \frac{p_1}{c_1}; \quad K_c = \frac{c_1^g}{c_1}, \quad (22)$$

kus ülaindeks g viitab gaasifaasile. K_2 SI põhiühikuks on [Pa m³ mol⁻¹] ning K_c on dimensioonitu.

Eelpool toodud koefitsiendid (Bunsen'i - α , Kuenen'i - S , Ostwald'i - L ja Absorptsiooni - β) on moolosaga seotud standardtingimustel järgnevalt:

$$x_1 = \frac{1}{1 + \frac{RT^\circ}{p_1^\circ V_{m,2}^* \alpha}} = \frac{1}{1 + \frac{RT^\circ}{p_1^\circ M_2 S}} = \frac{1}{1 + \frac{RT^\circ}{p_1^\circ V_{m,2}^* L}} = \frac{1}{1 + \frac{RT^\circ}{p^\circ V_{m,2}^* \beta}}. \quad (23)$$

Massilahustuvust, väljendatud kontsentratsioonina (molaarsus), on kasutanud lahustuvuse määramisel erinevate arvutuslike meetoditega Jurs [42], Yalkowsky [43] ja Acree [44]. Üldisemalt saab lahustuvust väljendada jaotuskoefitsiendina ning võib käsitleda kui soluudi molaarse kontsentratsiooni suhet vedelas ja gaasifaasi. Abraham [18] ja Katritzky [45] kasutasid gaas/solvent või vesi/solvent logaritmilisi jaotuskoefitsiente, vasatvalt $\log L$ ja $\log P$, lahustuvuse määramisel. L on siin Ostwaldi lahustuvuse koefitsient ning tema logaritmilise võrrandi läbi kontsentratsioonide võib kirjutada kujul:

$$\log L = \log \left(\frac{C_l}{C_g} \right), \quad (24)$$

kus C_l ja C_g on vastavalt soluudi kontsentratsioon vedel- ja gaasifaasis [34].

Antud töös on kasutatud Ostwaldi lahustuvuskoefitsienti eksperimentaalse suurusena. Ostwaldi lahustukoeffitsient kirjeldab gaas-vedelik tasakaalu, ja on väljendatav energia skaalas kui lahustuvuse vabaenergia ΔG_S . Lahustuvuse vabaenergia kirjeldab tööd, mis tuleb teha, et viia 1 mool soluuti gaasifaasist lahusesse (võrrand 25) [34].

$$\Delta G_S = -2.3RT \log L = -2.3RT \log \left(\frac{C_l}{C_g} \right) \quad (25)$$

MODELLERITAVAD ANDMED JA METOODIKA

Eksperimendi andmed

Kokku sisaldab andmemaatriks 154 solventi ja 397 soluuti eksponeerides 4540 eksperimentaalset määratud lahustuvust. Eksperimentaalväärtuste arv varieerub solvendi seeriates ühest punktist kuni 233 punktini (n-heksadekaan) ja soluudi seeriates kuni 94 punktini (n-oktaan).

Kasutatud andmepunktidest on kas eksperimentaalselt määratud 25 °C juures või kogutud erinevatest kirjandusallikatest [34, vt. ka lisa]. Oluliseks allikaks on olnud ka Abraham'i publikatsioonid, eriti heksadekaani ja alkoholide lahustuvusest [46, 47]. Lahustuvuse molaarsed väärtused arvutati ümber Ostwaldi lahustuvuskoeffitsiendile ($\log L$) teisendades võrrandi 24 põhjal ning kasutades vastavaid aururõhu ja tiheduse väärtusi [34, 48].

QSPR mudelite tuletamine

QSPR mudel on MLR võrrand molekulide erinevate füüsikaliste või termodünaamiliste omaduste ennustamiseks. QSPR mudel avaldub kujul:

$$P = a + bD1 + cD2 + dD3 + \dots, \quad (26)$$

kus P on uuritav füüsikaline omadus ja $a, b, c, d, \dots$ on regressiooni koefitsiendid ning $D1, D2, D3, \dots$ on molekuli struktuurist saadud parameetrid, ehk nn. deskriptorid.

Parima QSPR mudeli leidmisel kasutati BMLR-i [49, 50, 51, 52] protseduuri. BMLR-i käigus kõigepealt elimineeritakse tähtsusetud ($r^2 < 0,1$) ja puuduva väärtusega deskriptorid. Sellele järgneb parima kahe-, parima kolme- jne. parameetrilise regressiooni leidmine lähtudes deskriptorite statistilisest olulisusest ja mittekolineaarsuse kriteeriumist ($r^2 < 0,6$). BMLR-is deskriptorite skaala normaliseeritakse ja tsentreeritakse automaatselt ning lõpptulemus antakse algskaalas. Lõplik mudel on parim omaduse kirjeldus etteantud parameetrite arvuga antud deskriptorite kogumi juures.

Põhikomponentide analüüs

Põhikomponentide analüüs (PCA) on tuntud multivariatiivne ekstrapoleerimistehnika, mida on palju kasutatud ka keemias [53, 54, 55]. PCA-d on

mõistlik kasutada analüüsimeks üht andmematriksi (Z), mis sisaldab keemilisi või bioloogilisi muutujaid.

$$Z = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1m} \\ z_{21} & z_{22} & \dots & z_{2m} \\ \dots & \dots & \dots & \dots \\ z_{n1} & z_{n2} & \dots & z_{nm} \end{pmatrix}, \quad (27)$$

kus iga element z_{ji} tähistab j -nda objekti i -nda tunnuse väärtust. PCA leiab limiteeritud hulga uusi muutujaid s , mida kutsutakse põhikomponendi skooriks, mis summeerib originaalmuutujate informatsiooni. Matemaatiliselt PCA vastab Z matriksi faktoriseerimisele, mis on matriksite S ja L produkt, mis on vastavalt komponendi skoorid ja lugemid. Matemaatiliselt väljendatuna on see:

$$Z = l \times \bar{x} + S \times L + E. \quad (28)$$

Antud võrrandis l vastab veeru vektorile elemendi l kõikides positsioonides, $\bar{x}$ on rea vektor mis sisaldab muutujate (väljade) keskvaartust, ja E on residuaal matriks mis pole kirjeldatud antud mudeliga (viga). Põhikomponendi skoorid S näitavad sõltuvust ühendite (kirjete) vahel ja muutuja (välja) lugemid näitavad sõltuvust muutujate vahel [56].

PCA on eriti kasulik andmete klassifitseerimisel. Skoorvektori kandmine kahedimensionaalse graafikule teise skoorvektori suhtes näitab graafik sarnaste ühendite grupeerumist klastritesse. Sama tehakse lugemivektoritega näitab vastavate skoorklastrite statistilisi omadusi. Antud töös teostati PCA SIMCA-P programmi paketiga [57].

Kvantkeemilised arvutused

Ühendite struktuurid joonistati ja eeloptimeeriti programmiga HyperChem [58] kasutades MM+ jõuvälja molekulaarmehaanika arvutuses. Saadud geometriad lõppoptimeeriti samuti programmiga HyperChem kasutades poolempiirilist parametrissatsiooni AM1 [59] defineerides gradiendi limiidiks 0,01 kcal/Å. Edasise deskriptorite arvutuseks kasutati CodessaPro-sse [49, 60] sisseehitatud MOPAC 7.0 [60] programmi.

Molekulaardeskriptorid

Esimese teoreetilise molekulaardeskriptori töötas välja 1947 aastal H. Wiener [61]. Alates sellest ajast on välja toodud mõnituhat teoreetilist deskriptorit. Molekulaardeskriptorid jagatakse kahte klassi: eksperimentaalsed ja teoreetilised [20].

Esimesse kuuluvad struktuurideskriptorid (steerilisuse, resonantsi, induktsiooni, jne konstandid) ja solvatatsiooni deskriptorid (polaarsuse, polariseeritavuse, happelisuse, aluselisuse, jne skaalad). Teoreetilised deskriptorid (arvutatud otse molekuli struktuurist) jagatakse enamasti gruppidesse sõltuvalt nende arvutamise füüsikalisest päritolust või meetodist: konstitutsioonilised, elektrostaatilised, geomeetrilised, kvantkeemilised, ehk MO deskriptorid, termodünaamilised ja topoloogilised [20, 62].

CodessaPro [60] programmi poolt genereeritakse 107 erinevat teoreetilist põhideskriptorit, mida edaspidi laiendatakse lähtudes molekuli geomeetriast, aatomitest ja/või sidemetest aatomite vahel [63]. Sõltuvalt molekuli atomaarsest koostisest võib deskriptorite arv ühe molekuli kohta ulatuda keskimiselt 400-500 deskriptorini.

ARVUTUSLIK EKSPERIMENT

Arvutusliku eksperimendi võib jagada viide etappi järgnevalt (vt. artikli koopia, Joonis 1):

- I. QSPR mudelite tuletamine solventide ja soluutide seeriatele.
- II. Väikese maatriksi täitmine.
- III. PCA tagasiarvutus ja mudelid PCA komponentidele.
- IV. Suure maatriksi täitmine.
- V. Põhikomponentide analüüs.

Etapp I. Kõigepeal sorteeriti algses lahustuvuse maatriksis solvendid ja soluudid nii, et maksimaalne eksperimentaalsete väärtuste tihedus koonduks vasakusse ülemisse nurka. Seejärel eraldati solventide ja soluutide lahustuvuse read, kus oleks vähemalt 15 eksperimentaalset väärtust. Saadud 87 solventi $\times$ 91 soluuti moodustasid väikese maatriksi (SM0), kus on kõige rohkem eksperimentaalseid väärtusi.

Väikeses maatriksis olevatele solventidele ja soluutidele modelleeriti QSPR võrrandid, mida kasutati puuduvate Ostwaldi lahustuvuskoefitsientide ($\log L$) ennustamiseks kõigepealt soluutide mudelite järgi saades vastava väikese maatriksi SM0₁ ja seejärel solventide mudelite järgi saades maatriksi SM0₂.

Etapp II. Soluudi ja solvendi mudelitega ennustatud 2 väikest maatriksit, SM0₁ ja SM0₂, dimensioonidega 87 $\times$ 91 (87 solventi reas ja 91 soluuti veerus) liideti seejärel üheks maatriksiks SM1 kasutades järgnevaid reegleid:

1. kõigepealt kasutati väärtusi, mis olid ennustatud vastava mudel ennustuspiiridesse. Ennustuspiir defineeriti kui $\pm 15\%$ eksperimentaalsete väärtuste vahemikust. Igale mudelile defineeriti vastavalt erinevad piirid (vt. Lisad 4 ja 5),
2. juhul kui ennustatud väärtus oli ennustuspiirides mõlema maatriksi puhul, siis võeti kahest väärtusest mudel-kaaluga keskmine, mis defineeriti järgnevalt:

$$\overline{\log L} = \frac{k_X X + k_Y Y}{k_X + k_Y}, \quad (29)$$

kus X on ennustatud väärtus solvendi QSPR mudelist (vertikaalselt) ja Y on ennustatud väärtus soluudi mudelist (horisontaalselt) ning k on vastavate mudelite statistiline koefitsient ja arvutatud kasutades võrrandit 30

$$k = \frac{n}{(1 - R^2)^2 N^2}, \quad (30)$$

kus n on kasutatud eksperimentaalväärtuste arv mudeli tuletamisel, R^2 vastava mudeli korrelatsiooni koefitsiendi ruut ja N on deskriptorite arv mudelis.

3. juhul kui solvendi mudelist ennustatud lahustuvuse väärtus oli ennustuspiiridest väljas, siis kasutati vastavat soluudi mudeliga ennustatud väärtust, ja vastupidi,
4. kui ennustatud väärtus oli ennustuspiiridest väljas mõlema mudeli puhul, siis solvendid reastati enim andmepunkte sisladava polaarsuse skaala, E_{T30} , järgi ning lahustuvuse keskväärtus arvutati nn. „paremal-vasakul kolme naabri kaalutud keskmise“ järgi, mis on defineeritud järgnevalt:

$$\overline{\log L} = \frac{3(X_{1l} + X_{1r}) + 2(X_{2l} + X_{2r}) + X_{3l} + X_{3r}}{12}, \quad (31)$$

kus X_{1l} on vasaku naabri väärtus ja X_{1r} on parema naabri väärtus. Esimese naabri väärtus puuduva väärtuse suhtes kaalutud 3-ga, järgmine 2-ga ja viimane, kaugeim neist, 1-ga. Kui E_{T30} eksperimentaalne väärtus puudus, kasutati vastavat ennustatud väärtust Katritzky *et al.* poolt [64].

Etapp III. Järgnevalt teostati süstemaatiliselt täidetud maatriksile SM1 ja tema pöördmaatriksile SM2 põhikomponent analüüsi (PCA), kus andmed eelnevalt normaliseeriti ja tsentraliseeriti andmaks neile võrdväärset kaalu.

PCA tulemusel saadud skooridele (S) ja lugemitele (L), standardhälbele (SD) ja maatriksi iga tulba keskväärtusele ($\overline{M}$) koostati QSPR mudelid kasutades sama deskriptorite andmekomplekti, mida kasutati etapis I. PCA-st saadud skoorid, lugemid, standardhälve ja keskväärtus defineeriti kui omadused ja neile vastavad regressiooni võrrandid tuletati kasutades BMLR [50, 51, 52, 63] algoritmi. Võrrandit (32) kasutati järgnevalt lahustuvuste tagasiarvutuseks:

$$\log L = \sum_i (S_i \times L_i) SD + \overline{M}. \quad (32)$$

Etapp IV. Kui solvent/soluut QSPR mudelitest ennustatud väärtused langesid mudeli ennustuspiiridesse siis kasutati neid suure lahustuvuse maatriksi HM0 puuduvate väärtuste ennustamiseks. Kui QSPR-de ennustatud väärtused olid ennustuspiiridest väljaspool, siis kasutati võrrandit 32.

Lahustuvuse maatriksi HM0 võib jagada nelja virtuaalsesse sektorisse: sektor I, maatriksi ülemine vasak nurk (87 solventi $\times$ 91 soluuti), mis sisuliselt on väike maatriks

SM0; sektor II, ülemine parem nurk (67 solventi $\times$ 91 soluuti); sektor III, alumine vasak nurk (87 solventi $\times$ 306 soluuti); ning sektor IV, alumine parem nurk (67 solventi $\times$ 306 soluuti). Sektor I (SM0) täideti juba etapis II. Sektor II täideti kas soluutide QSPR mudelite ennustustest või PCA tagasiarvutuse skeemiga (Etapp III), sektor III vastavalt kas solventide mudelitest või samuti PCA tagasiarvutuse skeemiga (Etapp III) ning viimaks sektor IV täideti täielikult PCA tagasiarvutuse skeemiga (Etapp III).

Seejärel teoststi PCA analüüs täidetud maatriksil HM1 ja tema pöördmaatriksil HM2.

Etapp V. Saadud põhikomponentide interpreteerimine, skooride graafikute ja lugemite graafikute diskussioon, solventide/soluutide klasterite analüüs ning põhikomponentidele füsiko-keemilise tähenduse andmine on plaanis järgmise tööna.

TULEMUSED JA DISKUSSIOON

QSPR mudelite tuletamine (etapp I)

Kõigepealt arvutati CodessaPro-ga 1101 molekulaar deskriptorit. Kuigi BMLR meetod elimineerib ebaolulised deskriptorid automaatselt, kujunes probleemiks lahustuvuste seeriade struktuurne erinevus ja vajadus ennustada struktuure erineva atomaarse koostisega. Seetõttu oli vajalik koostada dekriptorite andmekomplekt, mis on kõigi lahustuvuste seeriade jaoks ühine ja kus kõik deskriptorid on alati arvutatavad kõigi võimalike ühendite jaoks. Sellise deskriptorite andmekomplekti kokkupanemiseks koostati eeskiri, mille abil elimineeriti mitteolulised ja mitesobivad descriptrid järgnevalt:

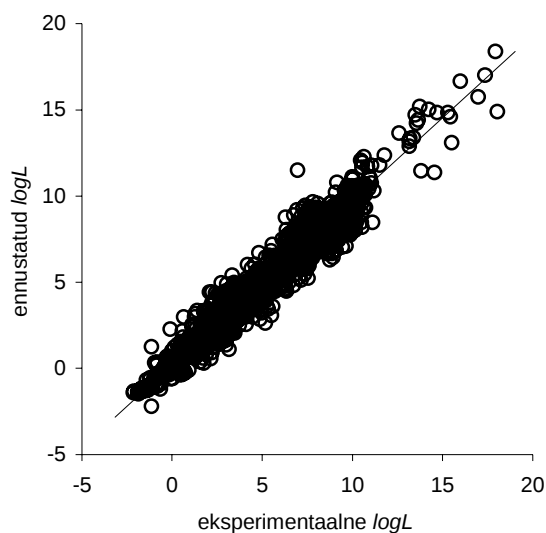
1. kõigepealt jäeti välja 662 deskriptorit mis on seotud kindla aatomiga kuna kõik ained ei pruugi alati neid sisaldada (tüüpiline näide on aatomite loetelud). Erandlikult jäeti alles deskriptorid süsiniku aatomite arv ja suhteline süsiniku aatomite arv, kuna nad on rakendatavad pea kõigile seeriatele. Vaid 7 ühendit 434-st antud maatriksis ei sisalda süsinikku.
2. järgnesid kvantkeemilistest arvutustest 38 laengujaotusega seotud deskriptorit, kuna Mulliken'i laengujaotuse skeem koos AM1 parametriseerimisega ei anna kooskõlalisi väärtusi halogeniididele. Seetõttu kasutati nende asemel Zefirov'i lähendusest (elektronegatiivsuste baasil) arvutatud laengujaotuse deskriptoreid.
3. samuti 81 vesiniksidemeid kirjeldavat deskriptorit, mis baseeruvad kvantkeemilistel arvutustel ja millel on Zefirov'i analoogid.
4. 81 reaktsiooni indeksit, mis on samuti seotud kindla aatomiga.
5. 3 erinevat järku Kier'i kuju indeksit, kuna neid ei saa arvutada molekulidele nagu H_2S , H_2O , NH_3 ja CH_4 .
6. 9 inertsimomendi deskriptorit, kuna neil on ebanormaalselt kõrged väärtused väikestele kolme aatomiga struktuuridel.
7. ja lõpuks 12 konstitutsioonilist deskriptorit (mitmekordsete sidemete arv, tsüklite arv, jne) ja 3 normaalvõnke deskriptorit eemaldati samuti kuna neid ei peetud piisavalt olulisteks.

Kokku eemaldati 889 deskriptorit. Lõplik deskriptorite valim sisaldab 212 kogumolekuli deskriptorit ja koosneb: 8 konstitutsioonilisest, 91 elektrostaatilisest, 12 geomeetrisest, 29 kvantkeemilisest, 35 termodünaamilisest ja 37 topoloogilisest deskriptorist. Detailne QSPR-e tuletamisks kaustatud deskriptorite nimekiri on toodud Lisas 1.

Mudelite tuletamise käigus ilmnes lahustuvuste seeriates 52 normaaljaotusest tugevalt kõrvalekalduvat eksperimentaalset punkti, mis eemaldati seeriast, et vältida mittenormaalset jaotust andmeseeriates (vt. artikli koopa).

Väikeses maatriksis olevatele eksperimentaalsetele andmetele (4540 andmepunkti) tuletati 87 QSPR-i mudelit solventidele (vertikaalselt) ja 91 QSPR-i soluutidele (horisontaalselt). Saadud solventide ja soluutide mudelid on toodud Lisades 2 ja 3 ning mudelitele vastavad statistilised karakteristikud on toodud Lisades 4 ja 5.

Ennustatud Ostwaldi lahustuvuskoefitsiendi väärtuseid võrreldi üle kõigi QSPR-ide eksperimendi väärtustega. Joonistel 1 ja 2 on võrdlus toodud vastavalt solvendi ja soluudi QSPR mudelite kohta. Korelatsiooni koefitsiendi ruut ($R^2 = 0,996$) näitab soluutide puhul kõrgemat ennustuskvaliteeti kui solvendi mudelitega ($R^2 = 0,957$). Võrreldes neid R^2 -e QSPR mudelite keskmiste R^2 -dega (vt. Lisa 4 ja 5) näeme, et solventide puhul on nad küllalt sarnased ($R^2 = 0,957$ ja $0,961$), soluutide puhul pisut erinevad ($R^2 = 0,996$ ja $0,920$). Põhjuseks võib olla, et soluutide mudelite puhul on keskmine eksperimentaalsete punktide väärtuste vahemik tunduvalt kitsam võrreldes solventide omadega, vastavalt 2,1 ja 8,1 lahustuvusühikut (vt. Lisa 4 ja 5). Kokkuvõtvalt on mõlemal juhul kooskõla eksperimentaalsete ja ennustatud väärtuste vahel väga hea.



Joonis 1. 87 solvendi seeria ennustatud *versus* eksperimentaalse lahustuvuse väärtused: $y = 0,9559x + 0,1878$; $R^2 = 0,9574$; (4167 punkti).

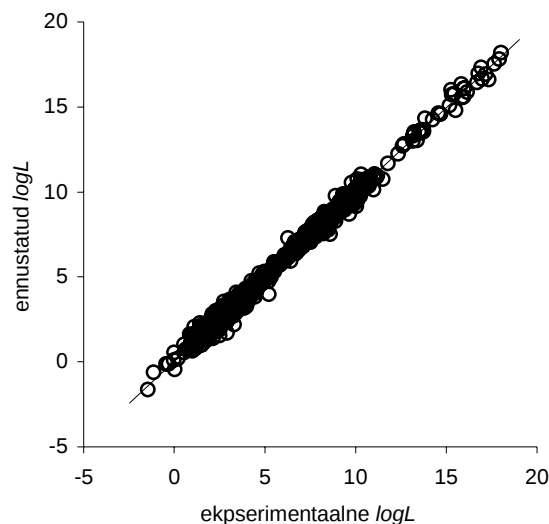
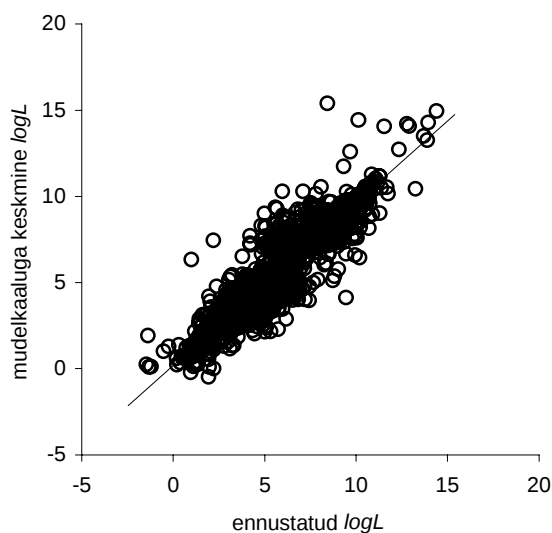


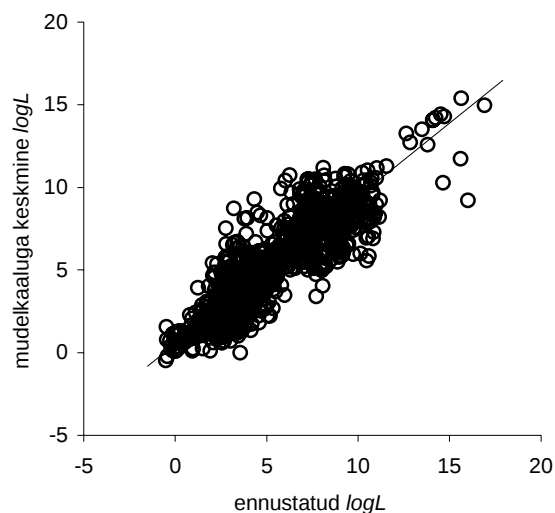
Figure 2. 91 soluudi seeria ennustatud *versus* eksperimentaalse lahustuvuse väärtused: $y = 0,9958x + 0,0203$; $R^2 = 0,9958$; (3394 punkti).

Väikese maatriksi täitmine (etapp II)

Maatriksid $SM0_1$ ja $SM0_2$ liideti vastavalt arvutusliku eksperimendi etappis II kirjeldatud reeglite järgi. Täidetud $SM1$ maatriks sisaldas lõpuks 3074 (38,8%) eksperimentaalväärtust, 3011 (38,0%) mudel-kaalutud keskvväärtust (võrrand 29), 1134 (14,3%) väärtust ennustatud soluudi QSPR mudelitest (ekstrapoleeritud horisontaalselt), 482 (6,1%) väärtust solvendi mudelitest (ekstrapoleeritud vertikaalselt), ja 216 (2,7%) „paremal-vasakul kolme naabri kaalutud keskmise“ keskvväärtust (võrrand 31).

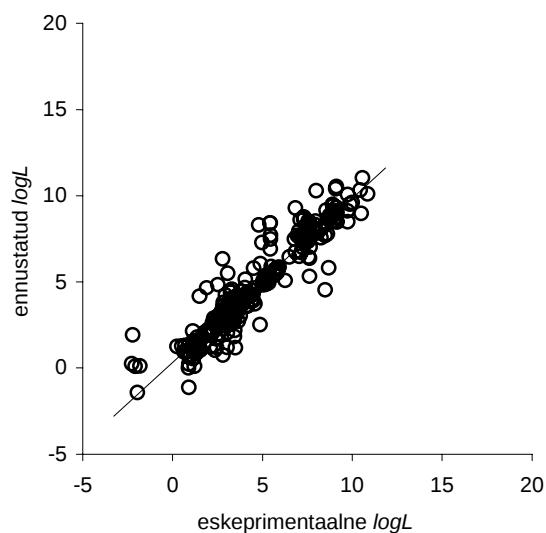


Joonis 3. 87 solvendi seeria mudel-keskmistatud keskvväärtus *versus* ennustatud väärtus: $y = 0,9456x + 0,1832$; $R^2 = 0,8865$; (3011 punkti).



Joonis 4. 91 soluudi seeria model-keskmistatud keskvärtus *versus* ennustatud väärtus: $y = 0,8899x + 0,5237$; $R^2 = 0,8558$; (3011 punkti).

Joonistele 3 ja 4 on kantud modelkaalutud keskvärtused võrreldes vastavalt solventide ja soluutide ennustatud väärtustega QSPR mudelitest. Nagu joonistelt näha, on korelatsiooni koefitsientide ruudud sarnased, $R^2 = 0,886$ solventide ja $R^2 = 0,856$ soluutide puhul. Ühtlasi näitavad mõlemad sõltuvused, et model-keskmistatud väärtuste kasutamine on õigustatud.



Joonis 5. Väline valideerimine: ennustatud *versus* eksperimentaalsed lahustuvuse väärtused: $y = 0,9532x + 0,3159$; $R^2 = 0,8815$; (289 punkti).

Uurimustöö ajal koguti veel lisaks 289 eksperimentaalset $\log L$ väärtust mida ei ole kasutaud QSPR-de tuletamisel. Neid kasutati maatriksis SM1 olevate ennustuste väliseks

valideerimiseks. Saadud korrelatsiooni graafik on toodud joonisel 5, kus $R^2 = 0.882$, mis näitab, et ennustusväärtused on aktsepteeritavad.

PCA tagasiarvutus ja mudelid PCA komponentidele (etapp III)

Põhikomponentide analüüsi teostati nii normaalmaatriksile SM1 (87 solventi $\times$ 91 soluuti), kui ka pöördmaatriksile SM2 (91 soluuti $\times$ 87 solventi). SM1 puhul kirjeldasid esimesed kolm komponenti 96,03% andmete varieeruvusest. Järgmise komponendi lisamine oleks ebaoluline kuna ennustuse kvaliteedi mõõde, Q^2 , ei näita edasist kasvu uute komponentide lisamisel. Esimese kümne komponendi karakteristikud on toodud Tabelis 1.

SM2 PCA andis SM1-ga võrreldes keskpärase tulemuse. Kaks esimest komponenti kirjeldavad vaid 50,33% andmete varieeruvusest. Kolmas ja neljas komponent annavad juurde vastavalt 10,02% ja 5,60%. Nagu Tabelist 2 näha, viienda jne. komponendi panus muutub väga väikeseks ja ka Q^2 väärtus ei kasva enam peale neljandat komponenti.

Tabel 1. Esimese 10 põhikomponenti ja nende protsentuaalne varieeruvuse katmine normaalse maatriksi puhul (solvent $\times$ soluut).

PC	Omapäärtus	%	Summaarne omapäärtus	Summaarne %	Summaarne Q^2
SM1					
1	80,679	92,735	80,679	92,735	0,924
2	2,172	2,496	82,851	95,231	0,946
3	0,692	0,795	83,543	96,026	0,952
4	0,440	0,505	83,982	96,532	0,952
5	0,330	0,380	84,313	96,911	0,950
6	0,297	0,341	84,609	97,252	0,951
7	0,238	0,274	84,848	97,526	0,949
8	0,215	0,247	85,062	97,773	0,948
9	0,202	0,232	85,264	98,005	0,948
10	0,149	0,171	85,413	98,176	0,945
HM1					
1	134,513	87,346	134,513	87,346	0,871
2	4,884	3,171	139,397	90,517	0,899
3	3,125	2,030	142,522	92,547	0,917
4	2,521	1,637	145,044	94,184	0,932
5	1,139	0,739	146,182	94,923	0,937
6	0,789	0,513	146,972	95,436	0,940
7	0,658	0,427	147,629	95,863	0,943
8	0,520	0,338	148,149	96,201	0,943
9	0,465	0,302	148,614	96,502	0,943
10	0,411	0,267	149,026	96,770	0,944

Tabel 2. Esimese 10 põhikomponenti ja nende protsentuaalne varieeruvuse katmine pööratud maatriksi puhul (soluut × solvent).

PC	Omaväärtus	%	Summaarne omaväärtus	Summaarne %	Summaarne Q^2
SM2					
1	23,432	26,933	23,432	26,933	0,233
2	20,358	23,400	43,789	50,332	0,390
3	8,716	10,019	52,505	60,351	0,486
4	4,870	5,598	57,376	65,949	0,526
5	3,101	3,565	60,477	69,514	0,526
6	2,403	2,763	62,880	72,276	0,517
7	2,175	2,499	65,055	74,776	0,507
8	1,870	2,149	66,925	76,925	0,503
9	1,706	1,961	68,631	78,886	0,494
10	1,512	1,738	70,142	80,624	0,478
HM2					
1	41,744	27,106	41,744	27,106	0,205
2	29,333	19,047	71,077	46,154	0,396
3	13,848	8,992	84,924	55,146	0,470
4	12,040	7,818	96,964	62,964	0,526
5	7,108	4,616	104,072	67,580	0,557
6	6,150	3,993	110,222	71,573	0,596
7	4,063	2,638	114,285	74,211	0,607
8	3,702	2,404	117,987	76,615	0,606
9	3,307	2,147	121,294	78,762	0,621
10	2,506	1,628	123,800	80,390	0,632

Edasises andmetötluses kasutati normaalmaatriksit (SM1) kuna PCA andis SM1-l paremad tulemused. SM1 PCA mudelis kirjeldati varieeritavus ära kolme esimese põhikomponendiga, siis tuletati kokku kaheksa QSPR mudelit skooridele (S_i) ja lugemitele (L_i), standardhälbele (SD) ja keskväärtusele (M), kus i on põhikomponendi number. Tabelis 3 on toodud vastavad mudelid ja Tabelis 4 on reastatud mudelites esinevad deskriptorid.

Tabel 3. Kolme esimese komponendi skooride ja lugemite ning standardhälbe ja keskväärtuse QSPR mudelid.

Vör,	Põhikomponentide QSPR mudelid	N	n	R ²	R ² _{cv}	s ²	F	Δ
1	$S_1=1,73(\pm 0,0479)-0,00194(\pm 0,0000502)D17-0,163(\pm 0,0171)D7$	91	2	0,952	0,948	0,0492	871	-3,86-1,55
2	$S_2=2,06(\pm 0,143)-1,70(\pm 0,126)D30+0,0229(\pm 0,00209)D32-115(\pm 10,9)D9-0,629(\pm 0,118)D23$	91	4	0,885	0,864	0,1201	166	-2,85-1,91
3	$S_3=-2,19(\pm 0,152)+0,198(\pm 0,0122)D1+139(\pm 13,9)D9-0,123(\pm 0,0135)D3-0,432(\pm 0,0556)D33-0,0590(\pm 0,0133)D4+0,000661(\pm 0,000269)D2$	91	6	0,794	0,761	0,2209	53,9	-1,47-3,02
4	$L_1=-0,605(\pm 0,0356)-0,209(\pm 0,0194)D20+0,000806(\pm 0,000125)D11-0,107(\pm 0,0175)D15+0,102(\pm 0,0251)D35-0,0155(\pm 0,00399)D24-0,207(\pm 0,0679)D6$	87	6	0,691	0,655	0,0004	29,9	-0,99-0,75
5	$L_2=-1,23(\pm 0,247)-0,134(\pm 0,0150)D23-0,0106(\pm 0,00147)D5-0,144(\pm 0,0203)D30+0,0405(\pm 0,00603)D28-6,92(\pm 1,35)D9$	87	5	0,902	0,883	0,0026	150	-0,54-0,29
6	$L_3=-3,37(\pm 0,542)+3,55(\pm 0,564)D21+0,408(\pm 0,0691)D12-0,0421(\pm 0,00902)D34+0,0101(\pm 0,00306)D31+0,00350(\pm 0,00119)D18-0,0656(\pm 0,0270)D30$	87	6	0,674	0,592	0,0028	27,6	-0,14-0,28
7	$SD=3,14(\pm 0,142)-0,211(\pm 0,0216)D16+33,2(\pm 4,00)D13+0,0388(\pm 0,00516)D27-0,148(\pm 0,0231)D29+0,0648(\pm 0,0136)D22-0,00778(\pm 0,00164)D14-0,490(\pm 0,113)D8+0,0411(\pm 0,0109)D25-0,00914(\pm 0,00338)D3$	87	9	0,763	0,701	0,0148	27,6	2,23-3,35
8	$M=8,75(\pm 1,00)-61,0(\pm 6,36)D9+2,94(\pm 0,356)D12+0,00164(\pm 0,000367)D26-0,0539(\pm 0,0130)D19+0,00921(\pm 0,00226)D10-0,0836(\pm 0,0240)D28$	87	6	0,746	0,646	0,0419	39,2	3,12-5,66

kus N on andmepunktide arv, n on parameetrite arv mudelis, R² and R²_{cv} on vastavalt korelatsiooni koefitsiendi ruut ja rist-valideeritud korelatsiooni koefitsiendi ruut, s² esitab standardhälvet, ja F on Fisher'i kriteerium.

Tabel 4. Deskriptorid ja nende esinemistihedus Tabelis 3 toodud QSPR mudelites.

ID	Deskriptori nimi	esinemine
konstitutsiooniline		
D1	Number of single bonds	1
elektrostaatiline		
D2	1X BETA polarizability (DIP)	1
D3	count of H-donors sites (Zefirov PC) (all)	2
D4	Difference (Pos - Neg) in Charged Part of Charged Surface Area (Zefirov's PC)	1
D5	DPSA3 Difference in CPSAs (PPSA3-PNSA3) (Zefirov PC)	1
D6	FPSA2 Fractional PPSA (PPSA-2/TMSA) (Zefirov PC)	1
D7	HA dependent HDCA-1 (Zefirov PC) (all)	1
D8	HA dependent HDCA-2 (Zefirov PC) (all)	1
D9	H-donors FCPSA (version 2)	4
D10	H-donors PSA (version 2)	1
D11	PNSA2 Total charge weighted PNSA (Zefirov PC)	1
D12	Polarity parameter (Zefirov)	2
D13	Positively Charged Part of Partial Charged Surface Area (Zefirov's PC)	1
D14	RNCS Relative negative charged SA (SAMNEG*RNCG) (Zefirov PC)	1
D15	RPCG Relative positive charge (QMPOS/QTPLUS) (Zefirov PC)	1
D16	WNSA3 Weighted PNSA (PNSA3*TMSA/1000) (Zefirov PC)	1
geomeetiline		
D17	Gravitation index (all bonds)	1
D18	Shadow plane YZ	1
kvantkeemiline		
D19	HOMO - LUMO energy gap	1
D20	Max bonding contribution of one MO	1
D21	Max SIGMA-SIGMA bond order	1
D22	Tot dipole of the molecule	1
D23	Tot hybridization comp. of the molecular dipole	2
D24	Tot molecular 2-center exchange energy	1
D25	Tot molecular 2-center resonance energy	1
termodünaamiline		
D26	Thermodynamic heat of formation of the molecule at 300K	1
D27	Thermodynamic heat of formation of the molecule at 300K /natoms	1
D28	Translational entropy (300K)	2
topoloogiline		
D29	Average Complementary Information content (order 0)	1
D30	Average Information content (order 0)	3
D31	Bonding Information content (order 2)	1
D32	Information content (order 1)	1
D33	Kier&Hall index (order 3)	1
D34	Structural Information content (order 0)	1
D35	Topographic electronic index (all bonds)	1

Parim saadud mudel on esimese skoori (S_1) kohta, sisaldades vaid kaks deskriptorit ja andes $R^2 = 0,95$. Vastupidiselt kehveim mudel saadi esimese lugemi (L_1) kohta, kuue deskriptoriga ja $R^2 = 0,69$. Selle põhjuseks on lugemi väärtuste väga väike varieeruvus, -

0,99 kuni -0,87, vaid vesi omab -0,75, mis näitab, et antud lugemi mõju on peaaegu olematu.

Kuna esimene peakomponent katab 92,7% lahustuvuse varieeruvust, omab esimene skoor suurt tähtsust. Tähtsaim deskriptor (suurima t-test väärtusega) kaheparameetrisel S_1 mudelis on gravitatsiooni indeks üle kõikide sidemete (D17), defineeritud järgneva võrrandiga:

$$D17 = \sum_{i < j}^{N_b} \frac{m_i m_j}{r_{ij}^2}, \quad (32)$$

kus m_i ja m_j on vastavalt i ja j aatomi aatommassid, r_{ij} on aatomite vaheline kaugus ja N_b on keemiliste sidemete arv molekulis. Gravitatsiooni indeks peegeldab massi jaotust molekulis ja näitab intermolekulaarseid dispersiooni jõude vedelas keskkonnas. Teine deskriptor on vesiniksideme doonori laetud pindala, *HDCA-1* (D7), defineeritud kui:

$$D7 = \sum_D S_D, \quad (33)$$

kus S_D on solvendile kättesaadav H aatomi pindala mis on võimeline käituma doonorina vesiniksideme moodustamisel [60, 65]. Kahe deskriptori (D17 ja D7) kombinatsioon kirjeldab selgelt intermolekulaarseid jõude, mis mõjutavad lahustuvuse protsessi. Gravitatsiooni indeks (D17) on seotud dispersiooni ja augu tekke efektidega. *HDCA-1* (D7) on seotud ühendite võimega moodustada vesiniksidemeid.

Esimese skoori (S_1) QSPR mudel on sarnane Katritzky *et al.* poolt tuletatud kaheparameetrilise keemistäpi (T_b) mudeliga ($R^2 = 0,95$), kus gravitatsiooni indeks üle kõikide aatomite (G_P) oli kuupjuures. Teine deskriptor oli seotud vesiniksidemega (*HDCA-2*) [52]. Samuti täheldati sarnasust kaheparameetrilise aururõhu mudeliga, kus gravitatsiooni indeks üle kõikide seotud aatomite (G_I) ja vesiniksideme doonori laetud pindala (*HDCA-2*) annab lineaarse korelatsiooni $R^2 = 0,88$ [66]. Kaheparameetiline mudel vedelike viskoossusele ($\log \eta$) sisaldab samuti deskriptoreid G_I ja *HDCA-2*, andes 337 erineva orgaanilise molekuli korral mudeli korelatsioonikoeffitsiendiga $R^2 = 0,79$ [67] ja 361 ühendi korral $R^2 = 0,81$ [65, 68] eri autoirite töödes.

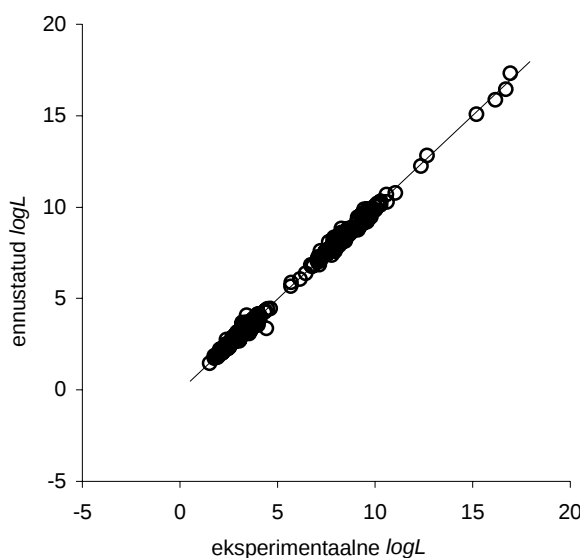
Suure maatriksi täitmine (etapp IV)

Sektorid algses lahustuvuse maatriksis täideti järgnevalt:

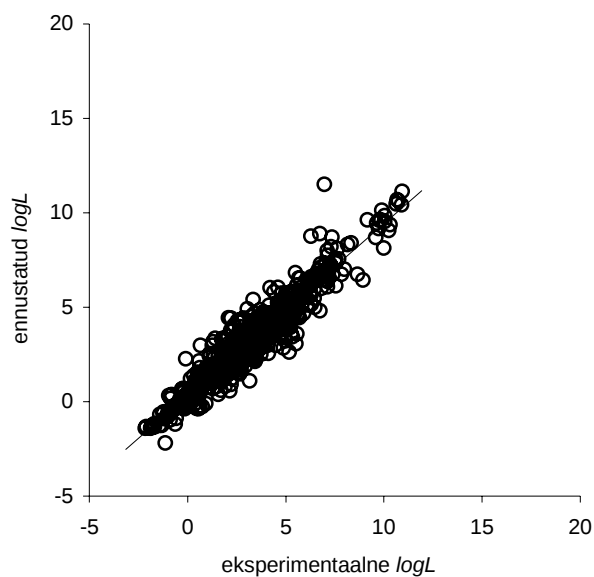
- 1 sektor I (SM0) täideti nagu kirjeldati eespool etapis II;

- 2 sektoris II puuduvad väärtused ennustati 91-st soluudi mudelist. Nendest ennustatud väärtustest 618 (10,1%) jäid QSPR mudelite ennustuspiiridest välja;
- 3 sektoris III puuduvad väärtused ennustati 87 solvendi mudelist. Käesoleval juhul jäid 4207 (15,8%) väärtustest QSPR mudelite ennustuspiiridest välja;
- 4 sektorites II ja III ennustuspiiridest väljajäänud väärtused saadi PCA tagasiarvutus skeemi abil (vt. Etapp III);
- 5 Sektor IV täideti täielikult PCA-st tagasiarvutatud väärtustega.

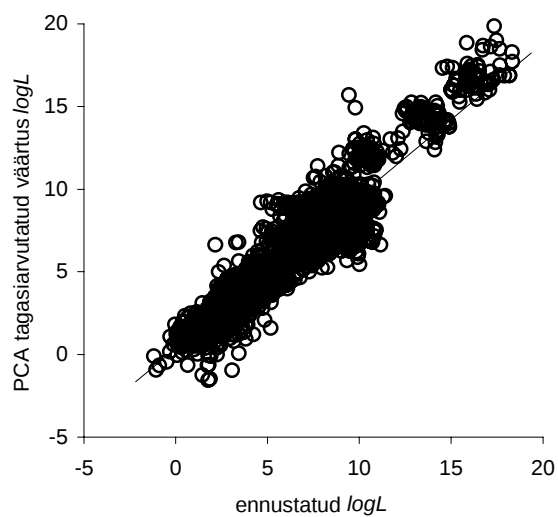
Ennustatud $\log L$ väärtused koreleeriti vastava sektori eksperimentaalsetega, et saada ülevaade ennustuse täpsusest. Sektoris II soluudi mudelite ennustus annab väga hea korelatsiooni ($R^2 = 0,997$) sektoris olevate eksperimentaalsete väärtustega (Joonis 6). Solvendi QSPR mudelite ennustus sektoris III annab aktsepteeritava korelatsiooni ($R^2 = 0,908$) sektoris olevate eksperimentaalsete väärtustega (Joonis 7). Samuti koreleeriti omavahel PCA tagasiarvutatud väärtused soluudi ja solvendi QSPR mudeli piiridesse ennustatud väärtustega, kus sektoris II on korrelatsioonikoefitsiendi ruut $R^2 = 0,900$ (Joonis 8) ja sektoris III $R^2 = 0,641$ (Joonis 9). Kasin tulemus sektoris III on tingitud laiast ning erinevast soluutide nomenklatuurist, mis moodustavad enamuse suurest maatriksist.



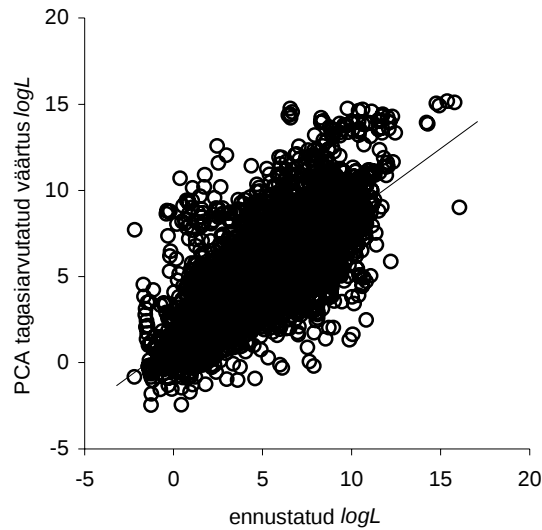
Joonis 6. Sektor II 91 soluudi seeria ennustatud *versus* eksperimentaalsed lahustuvuse väärtused: $y = 1,0041x - 0,0427$; $R^2 = 0,9968$; (357 punkti).



Joonis 7. Sektor III 87 solvendi seeria ennustatud *versus* eksperimentaalsed lahustuvuse väärtused: $y = 0,9092x + 0,3288$; $R^2 = 0,9082$; (1109 punkti).



Joonis 8. Sektor II PCA tagasi arvatud ennustused *versus* 91 soluudi mudeli piiridesse ennustatud väärtused: $y = 0,9235x + 0,3659$; $R^2 = 0,9002$; (5479 punkti).



Joonis 9. Sektor III PCA tagasi arvutatud ennustused *versus* 87 solvendi mudeli piiridesse ennustatud väärtused PCA: $y = 0,7557x + 1,0958$; $R^2 = 0,6414$; (22415 punkti).

Põhikomponentide analüüs (etapp V)

Täidetud maatriksile (HM1) ja tema pöördmaatriksile (HM2) teostati PCA. Kümme esimest põhikomponenti on vastavalt ära toodud Tabelites 1 ja 2. Nagu oli oodata, maatriksi dimensionaalsuse kasv vähendab PCA tulemust. Kumulatiivne protsentuaalne omaväärtus kukub 96,03%-lt 92,55%-le normaalse maatriksi puhul, pöördmaatriksi korral 65,95%-lt 62,96%-le.

KOKKUVÕTE.

Struktuur-omadus sõltuvusi ja põhikomponent analüüs kombineerituna ühtsesse metodoloogiasse kasutati edukalt ennustamaks suurt hulka lahustuvuse väärtusi. Kokku analüüsiti 4540 eksperimentaalset andmepunkti. Tuletati 178 QSPR mudelit 87 solvendi ja 91 soluudi seeriale jaoks, mis vastas andmematriksi tihedamale osale. Saadud mudelitest ennustatud 4843 lahustuvuse väärtusega täideti algmatriksi eksperimentaalväärtuste poolest tihedaimalt esindatud ala (3074 eksperimentaalset punkti). Saadud matriksile, dimensioonidega 87 solvendi ja 91 soluuti, teostati PCA, mida kombinatsioonis 178 QSPR-iga kasutati edukalt edasi kogu lahustuvuste matriksi, dimensioonidega 154 solvendi ja 397 soluuti, puuduolevate väärtuste ennustamiseks. Ennustatud väärtusi valideeriti välise 289 ühendist koosneva test andmekomplektiga .

Kasutatud QSPR-i ja PCA kombineeritud meetod näitab potentsiaali ennustamaks lahustuvuse väärtusi laias vahemikus. Täielikult täidetud lahustuvuste andmematriks on kirjeldatud kolme põhikomponentiga ja katab 92,55% varieeruvusest. Pöördmatriksi PCA annab kehvema tulemuse varieeruvuse kirjeldamisel. Põhikomponendid, mis kirjeldavad lahustuvuse andmeid matriksis, sisaldavad endas dispersioonjõudede, augu tekke jõudede, elektrostaatilisi jõudede ja vesiniksideme tekke panust lahustuvuse vabaenergia kirjeldamisel.

SUMMARY

The quantitative structure property relationships and principal component analysis combined into one methodology have been used successfully to predict a large number of solubility values. A total of 4540 experimental data points was analyzed. The 178 QSPR-s developed for the densest area of the data matrix (87 solvents x 91 solutes) with covering a total of 3074 experimental values were successful in the prediction of the remaining 4843 solubility values. The PCA on the densest area of the data matrix, combined with the 178 QSPR equations, were further used successfully in filling the reminder of the 154 times 397 data matrix. The prediction procedure was validated with an external test set of 289 experimental data points.

The proposed methodology, with its combination of QSPR-s and PCA, shows potential for the prediction of numerous solubility values. The three principal components from the fully filled data matrix where solutes are observations and solvents are variables describe 92.55% of the variability. The PCA on the corresponding transposed matrix results in only a moderate description of the variability. The principal components that describe the variability in the data matrix capture the contributions of the intermolecular dispersion forces, cavity formation forces, electrostatic forces and hydrogen bonding to the solvation free energy.

KASUTATUD KIRJANDUS.

- 1 Reichardt, C. Solvents and solvent effects in organic chemistry. Weinheim: Wiley-VCH, **2003**.
- 2 Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse QSPR Correlations of Technologically Relevant Physical Properties, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1-18.
- 3 Katritzky, A. R.; Fara, D. C.; Petrukhin, R. O.; Tatham, D. B.; Maran, U.; Lomaka, A.; Karelson, M. The Present Utility and Future Potential for Medicinal Chemistry of QSAR/QSPR with Whole Molecule Descriptors. *Current Topics in Medicinal Chemistry*, **2002**, *2*, 1333-1356.
- 4 Franks, F.; Ives, D. J. G. *Quart. Rev. Chem. Soc.* **1966**, *20*, 1.
- 5 Fuchs, O. Fortschr. Chem. Forsch. **1968**, *11*, 74.
- 6 Rouff, R. S.; Tse, D. S.; Malhotra, R.; Lorents, D. C. *J. Phys. Chem.* **1993**, *97*, 3379
- 7 James, K. C. *Education in Chemistry*, **1972**, *9*, 220.
- 8 Karelson, M.; Diercksen, G. H. F. Models for Simulating Molecular properties in Condensed Systems, in Wilson, S.; Diercksen, G. H. F. Problem Solving in Computational Molecular Science, Kluwer Academic Publisher, Netherland, **1997**.
- 9 Erös, D; Kéri, G; Kövesdi, I, Szántai-Kis, C.; Mészáros, G; Örfi, L. Comparison of Predictive Ability of Water Solubility QSPR Models Generated buy MLR, PLS and ANN Methods. *Mini-Rev. Med. Chem.* **2004**, *4*, 167-177.
- 10 Yalkowsky, S. H.; Banerjee, S. Aqueous Solubility. Methods of Estimation for Organic Compounds; Dekker: New York, **1992**.
- 11 Fowler, F. W.; Katritzky, A. R.; Rutherford, R. J. D. The Correlation of Solvent Effects on Physical and Chemical Properties. *J. Chem. Soc., Part B* **1971**, 460-469.
- 12 Koppel, I. A.; Palm, V. A. The Influence of the Solvent Organic Reactivity. In *Advances in Linear Free Energy Relationships*; Chapman, N. B.; Shorter, J. Eds.; Plenum Press: London, New York, **1972**, 203-247.
- 13 Kamlet, M. J.; Abboud, J.L.; Abraham, M. H.; Taft, R. W. Linear Solvation Energy Relationships. 23. A Comprehensive Collection of the Solvatochromic Parameters, π^* , α , and β , and Some Methods for Simplifying the Generalized Solvatochromic Equation. *J. Org. Chem.* **1983**, *48*, 2877-2887.
- 14 Fawcett, W. R.; Krygowsky, T. M. A Characteristic Vector Analysis of Solvent Effects for Thermodynamic Data. *Can. J. Chem.* **1976**, *54*, 3283-3292.
- 15 Swain, C. G.; Swain, M. S.; Powell, A. L.; Alunni, S. Solvent Effects of Chemical Reactivity. Evaluation of Anion and Cation Solvation Components. *J. Am. Chem. Soc.* **1983**, *105*, 502-513.
- 16 Mayer, U. NMR-Spectroscopic Studies on Solute- Solvent and Solute-Solute Interactions. In *Ions and Molecules in Solution*; Tanaka, N.; Ohtaki, H.; Tamamushi, R. Eds.; Elsevier: Amsterdam, **1983**; 219.

- 17 Dougherty, R. C. A Perturbation Molecular Orbital Approach to Ionic Solvation. A New Multiparameter Solvent Correlation Based on Solvent Ionization Potentials and Electron Affinities. *Tetrahedron Lett.* **1975**, 385-388.
- 18 Abraham, M. H. Physicochemical and Biological Processes. *Chem. Soc. Rev.* **1993**, 73-83.
- 19 Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 439-445.
- 20 Karelson, M. Molecular Descriptors in QSAR/QSPR; John Wiley & Sons: New York, **2000**.
- 21 Koppel, I. A.; Palm, V. A. *Reakts. Sposobnost Org. Soedin.* **1971**, 8, 291.
- 22 Abboud, J. L.; Kamlet, M. J.; Taft, R. W. Regarding a Generalized Scale of Solvent Polarities *J. Am. Chem. Soc.* **1977**, 91, 8325-8327.
- 23 Kamlet, M. J.; Abboud, J. L.; Taft, R. W. *J. Am. Chem. Soc.* **1977**, 99, 6027.
- 24 Abboud, J.-L. M.; Kamlet, M. J.; Taft, R. W. *Progr. Phys. Org. Chem.* **1981**, 13, 485.
- 25 Taft, R. W.; Abboud, J.-L. M.; Kamlet, M. J.; Abraham, M. H. *J. Sol. Chem.* **1985**, 14, 153
- 26 Hildebrand, J. H.; Prausnitz, J. M.; Scott, R. L. Regular and Related Solutions, Van Nostrand-Reinhold, Princeton, **1970**.
- 27 Barton, A. F. M. Handbook of Solubility Parameters and other Cohesion Parameters, CRC Press, Boca Raton/Florida, **1983**.
- 28 Hildebrand, J. H.; Scott, R. L. The Solubility of Nonelectrolytes, 3rd ed., Dover, New York, **1964**.
- 29 Scatchard, G. *Chem. Rev.* **1931**, 8, 321; **1932**, 10, 229.
- 30 Dack, M. R. J. The Importance of Solvent Internal Pressure and Cohesion to Solution Phenomena. *Chem. Soc. Rev. (London)* **1975**, 4, 211; Dack, M. R. J. *Aust. J. Chem.* **1975**, 28, 1643.
- 31 Burrell, H. Solubility Parameter Values, in Brandrup, J.; Immergut, E. H. (eds.) Polymer Handbook, 2nd ed., Wiley-Interscience, New York, **1975**.
- 32 Kamlet, M. J.; Taft, R. W. *Acta Chem. Scand. Part B* **1986**, 40, 619.
- 33 Kamlet, M. J.; Abraham, M. H.; Doherty, R. M.; Taft, R. W. *J. Am. Chem. Soc.* **1984**, 106, 464; *Nature (London)* **1985**, 313, 384.
- 34 Katritzky, A. R.; Oliferenko A. A.; Oliferenko P. V.; Petrukhin R.; Tatham D.B.; Maran U.; Lomaka A.; Acree W. E. Jr. A General Treatment of Solubility. 1. The QSPR Correlation of Solvation Free Energies of Single Solutes in Series of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1794-1805.
- 35 Katritzky, A. R.; Oliferenko A. A.; Olifenko P. V.; Petrukhin R.; Tatham D.B.; Maran U.; Lomaka A.; Acree W. E. Jr. A General Treatment of Solubility. 2. QSPR Prediction of Free Energies of Solvation of Specified Solutes in Ranges of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1806-1814.

- 36 Katritzky, A. R.; Tamm, T.; Wang, Y.; Karelson, M. A Unified Treatment of Solvent Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 692-698.
- 37 Cramer, R. D. BC(DEF) Parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state. *J. Amer. Chem. Soc.* **1980**, *102*, 1837-1849.
- 38 Chastrette, M.; Raizmann, M.; Chanon, M.; Purcell, K. F. Approach to a general classification of solvents using a multivariate statistical treatment of quantitative solvent parameters. *J. Amer. Chem. Soc.* **1985**, *107*, 1-11.
- 39 Dunn III, W. J.; Koehler, M. G.; Grigoras, S. The role of solvent-accessible surface area in determining partition coefficients. *J. Med. Chem.* **1987**, *30*, 1121-1126.
- 40 Castells, C. B.; Reta, M. R. Study of gas-liquid partitioning of alkane solutes in several organic solvents by using principal analysis and linear solvation energy relationships. *Analytica Chimica Acta.* **2003**, *488*, 107-122.
- 41 Fogg, P. G. T.; Bligh, S.-w. A.; Derrick, M. E.; Yampol skii, Y. P.; Clever, H. L.; Skrzecz, A.; Young, C. L. IUPAC-NIST Solubility Data Series. 76. Solubility of Ethyne in Liquids. *J. Phys. Chem. Ref. Data.* **2002**, *30*, 1693-1875.
- 42 McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237-1247.
- 43 Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubilities by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354-357.
- 44 Acree, W. E.; Powell, J. R.; McHale, M. E. R.; Pandey, S.; Borders, T. L.; Campbell, S. W. Thermodynamics of Mobile Order Theory. Research Trends in Physical Chemistry, Council of Scientific Research Integration, TriVandrum, India; **1997**; Vol. 6, pp 197-233.
- 45 Katritzky, A. R.; Mu, L.; Karelson, M. A QSPR Study of the Solubility of Gases and Vapors in Water. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1162-1168.
- 46 Abraham, M. H.; Andonian-Hafvtan, J.; Whiting, G. S.; Leo, A.; Taft, R. S. Hydrogen Bonding. Part 34. The Factors that Influence the Solubility of Gases and Vapours in Water at 298 K, and a New Method for its Determination. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1777-1791.
- 47 Abraham, M. H.; Grellier, P. L.; McGill, R. A. Determination of Olive Oil-Gas and Hexadecane-Gas Partition-Coefficients, and Calculation of the Corresponding Olive Oil-Water and Hexadecane-Water Partition-Coefficients. *J. Chem. Soc., Perkin Trans. 2* **1987**, 797-803.
- 48 Handbook of Physical Properties of Organic Chemicals; Howard, P. H., Meylan, W. M., Eds.; Lewis Publishers: Boca Raton, FL, **1997**.
- 49 Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA Reference manual (Version 2.0); Gainesville, FL, **1994**.
- 50 Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, *24*, 279-287.
- 51 Draper, N. R.; Smith, H. Applied Regression Analysis; Wiley: New York, **1981**.

- 52 Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400-10407.
- 53 Malinowski, E.R.; Howery, D.G. *Factor Analysis in Chemistry*, Wiley-Interscience (1980).
- 54 Strouf, O. *Chemical Pattern Recognition*, Wiley, New York (1986).
- 55 Meloun, M.; Militky, M.; Forina, M. *Chemometrics in Analytical Chemistry*, Ellis Horwood, New York (1992).
- 56 Eriksson, L.; Johansson E. Multivariate design and modeling in QSAR. *Chemometrics and Intelligent Laboratory Systems* **1996**, *34*, 1-19.
- 57 www.umetrics.com
- 58 www.hyper.com
- 59 Dewar, M. J. S.; Zebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902-3909.
- 60 www.codessa-pro.com
- 61 Wiener, H. Structural Determination of Paraffin Boiling Points *J. Amer. Chem. Soc.* **1947**, *69*, 17-20.
- 62 Petrukhin, R. Industrial Applications of the Quantitative Structure-Property Relationships, PhD Dissertation, University of Tartu, Estonia, **2001**.
- 63 Codessa Pro User's Manual, University of Florida, **2005**.
- 64 Katritzky, A. R.; Fara, D. C.; Kuanar, M.; Hur, M.; Karelson, M. A Three-Dimensional Classification of Solvents Based on QSPR *Unpublished work*.
- 65 Katritzky, A. R.; Chen, K.; Wang, Y.; Karelson, M.; Lučić, B.; Trinajstić, N.; Suzuki, T.; Schüürmann, G. J. Prediction of liquid viscosity for organic compounds by a quantitative structure-property relationship. *J. Phys. Org. Chem.* **2000**, *13*, 80-86.
- 66 Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water-Air Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720-725.
- 67 Ivanciuc, O.; Ivanciuc, T.; Filip, P. A.; Cabrol-Bass, D. Estimation of the Liquid Viscosity of Organic Compounds with a Quantitative Structure-Property Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 515-524.
- 68 Lučić, B.; Bašić, I.; Nadramija, D.; Miličević, A.; Trinajstić, N.; Suzuki, T.; Petrukhin, R.; Karelson, M.; Katritzky, A. R. Correlation of liquid viscosity with molecular structure for organic compounds using different variable selection methods. *ARKIVOC*, **2002**, *4*, 45-59.