

HELSINGIN YLIOPISTO

Kemian asiantuntijoiden hyödyntämä keminformatiikka: laadullinen tutkimus kemian opetuksen näkökulmasta

Matematiikan, fysiikan ja kemian opettajan maisteriohjelma, kemian opintosuunta
Maisterintutkielma

Laatija:
Aleksi Takala

Ohjaaja(t):
Yliopistolehtori ja kemian tiedekasvatuksen dosentti, Johannes Pernaa
Tiedekasvatuksen professori, Maija Aksela

11.9.2023
Helsinki

Tiedekunta: Matemaattis-luonnontieteellinen tiedekunta

Koulutusohjelma: Matematiikan, fysiikan ja kemian opettajan maisteriohjelma

Opintosuunta: Kemian opintosuunta

Tekijä: Aleksi Takala

Työn nimi: Kemian asiantuntijoiden hyödyntämä keminformatiikka: laadullinen tutkimus kemian opetuksen näkökulmasta

Työn laji: Maisterintutkielma

Kuukausi ja vuosi: 9/2023

Sivumäärä: 72

Avainsanat: keminformatiikka, kemian opetus, laadullinen tutkimus, TPASK

Ohjaaja tai ohjaajat: Johannes Pernaa, Maija Aksela

Säilytyspaikka:

Muita tietoja:

Tiivistelmä:

Tutkielman tavoitteena oli selvittää keminformatiikan mahdollisuuksia kemian opetukseen. Aiempaa suomenkielistä tutkimusta aiheesta ei ole julkaistu. Tutkielma toimii alustavana katsauksena keminformatiikan opetuksen mahdollisuuksiin haastatteleamalla asiantuntijoita, joiden osaamisen ja työskentelytavoista pyrittiin laadullisella sisällönanalyysillä ymmärtämään, mitä keminformatiikalla voidaan kemiallisissa tieteissä saavuttaa. Keminformatiikan luonne monitieteellisenä alana vaatii kehittyneitä kemian taitoja, joten opetuksen ja oppimisen mahdollisuuksien selvittäminen rajattiin yliopistokoulutukseen.

Tutkielmassa haastateltiin avoimella haastattelulla keminformatiikan asiantuntijoita (N=3). Seuraavat tutkimuskysymykset ohjasivat tutkielmaa: 1. Miten kemian asiantuntijat hyödyntävät keminformatiikkaa tutkimuksessa? 2. Millaisia taitoja asiantuntijat kokevat tarvitsevänsä keminformatiikan hyödyntämisessä? 3. Miten keminformatiikka voi edistää kemian opetusta ja oppimista TPASK-mallilla analysoituina? Tutkimuksella vastataan yliopisto-opetuksen kontekstissa tutkimuskysymyksiin avoimella haastattelulla, jota perustellaan sillä, ettei aihetta tarvitse rajata, sillä sitä ei ole aiemmin tutkittu kemian opetuksen tutkimuksessa.

Tutkijoita haastateltiin etäyhteydellä, haastattelut nauhoitettiin, käännettiin tarvittaessa suomeksi ja litteroitiin. Haastateltavat pseudonymisoitiin EU:n tietoturvalain ja tietosuoja-asetuksen mukaisesti. Tutkimuskysymyksiin vastattiin laadullisella TPASK-mallin teoriaohjaavalla sisällönanalyysillä aineistosta. Aineistoa luokiteltiin ensin induktiivisesti, ja pelkistyttyään deduktiivisesti yläluokiteltiin sopivilla TPASK-mallin luokilla, SK, TK ja TSK, sillä aineisto ei kuvannut pedagogisia tilanteita tai käytäntöjä.

Tuloksissa havaittiin, että keminformatiikkaa hyödynnettiin erilaisiin tutkimusongelmiin, jotka ovat luonteeltaan monitieteellisiä. Erityisesti datatieteelliset menetelmät, kemiallisen tiedon käsittely standardoidusti, mallintamisen objektiivisuus sekä tavoite täydentää tai korvata kokeellista kemiaa

olivat keskeisiä keminformatiikan tavoitteita. Keminformatiikan menetelmien kehitykseen vaikutti erityisesti empiirisen datan niukkuus, säätely menetelmien kuvailusta ja toisaalta automatisoitujen laboratorioden tuottamat suuret datamäärät. Keminformatiikkaa hyödynnettiin kohtaamaan tiedon tarpeita etsimällä uusia ratkaisuja paremmilla malleilla ja datan simuloinnilla, jonka tarkoitus on säästää resursseja ja korvata osa kemian kokeellisuudesta.

Keminformatiikan mahdollisuudet oppimiseen ja opetukseen ovat riippuvaisia taidoista, erityisesti ohjelmointitaidosta. Erityisesti Python-koodin merkitys korostui, mutta samalla erilaisten algoritmien ja koneoppimisen mallien soveltaminen. Toisaalta osa haastateltavista näki, että kemistit hyötyvät enemmän keminformatiikasta, kun taas toiset ajattelivat, että menetelmällisyys lähestyy datatieteellistä asiantuntijuutta. Keminformatiikan tutkimusongelmat ja taitovaatimukset osoittautuivat moninaisiksi, mikä vaikeutti selkeiden osaamistavoitteiden laadintaa tutkimuksessa.

TPASK-mallilla tuloksista analysoitiin, että keminformatiikan hyödyntäminen voi kehittää metakognitiivisia taitoja ja mahdollisesti tuottaa uusia ideoita. Tätä luonnehdittiin luovuuden käsitteellä, ja todettiin tieteellisten ongelmien ratkaisussa teknologias-tieteellisen tiedon, TSK:n, jossa keminformatiikka teknologiana voi muuttaa aiempia käsityksiä kemiasta ja täten edistää oppimista ja taitoja.

Kemian opetuksen ja oppimisen suhteen, uuden teknologian hyödyntäminen voi edistää aiempien tietorakenteiden arviointia metakognitiivisesti, ja ohjelmointitaidot ovat niin jatkuvan oppimisen kuin korkeakoulutuksen kannalta merkittäviä osaamistavoitteita. Keminformatiikan myötä opetus voi hyötyä uusimmista kemiallisen datatieteellisistä menetelmistä, mutta kynnys niiden hyödyntämiseen vaatii lisätutkimuksia. TPASK-malli toimi onnistuneesti kuvaamaan, kuinka keminformatiikan hyödyntäminen voi muuttaa käyttäjien tapaa ymmärtää kemiaa, jonka vaikutukset ovat niin tutkimuksessa ja koulutuspoliittisesti ulkomailla tunnustettu.

Sisällysluettelo

1	Johdanto	1
2	Teoreettinen viitekehys	4
2.1	Keminformatiikan perusteet	4
2.1.1	Kemiallisen tiedon käsittely	5
2.1.2	Kemiallisen rakennedatan tiedostotyypit ja linjanotaatiot	8
2.1.3	Tietokannat	11
2.2	Keminformatiikan käytänteet	13
2.2.1	Tiedonhallinta	13
2.2.2	Avoin lähdekoodi ja avoin data keminformatiikassa	14
2.2.3	Molekulaariset deskriptorit	15
2.3	Keminformatiikan menetelmät	16
2.3.1	Laskennallinen kemia ja molekyylihallinnus	16
2.3.2	Yleisiä keminformatiikan ohjelmistoratkaisuja	17
2.3.3	Ominaisuuksien ennustaminen	19
2.3.4	Koneoppimisen ja tekoälyn rooli keminformatiikassa	20
2.4	Keminformatiikan opetuksen tutkimus	22
2.4.1	Kemian opetuksen tutkimuksia keminformatiikasta yliopisto-opetuksessa	23
2.4.2	Keminformatiikan ja TPASK-mallin yhteydet	25
3	Tutkimus	30
3.1	Tutkimuskysymykset	30
3.2	Tutkimusote	30
3.3	Tutkimuskohde	31
3.4	Tutkimuksen luotettavuus	33
4	Tulokset	35
4.1	Keminformatiikan hyödyntäminen (TK 1)	35
4.2	Keminformatiikan vaikutus tutkimukseen (TK 1)	37
4.2.1	Keminformatiikka ja ongelmanratkaisu	38
4.3	Keminformatiikan osaaminen (TK 2)	40
4.3.1	Ohjelmointitaidot	40
4.3.2	Keminformatiikan datakeskeinen tieteellinen luonne	41
4.3.3	Keminformatiikan monitieteellinen tutkimusote	42

4.4	Keminformatiikan arvo opetuksessa ja oppimisessa TPASK-mallilla analysoituna (TK 3)	44
4.4.1	Tieteellinen tieto (SK) voi laajentaa kemian ajattelua	44
4.4.2	Teknologinen tieto (TK) täydentää tieteellistä menetelmää	45
4.4.3	Teknologis-tieteellinen tieto (TSK)	47
5	Keskustelu ja johtopäätökset	50
5.1	Keminformatiikka kemian tutkimuksessa (TK 1)	50
5.2	Keminformatiikka on ohjelmointia ja kemian mallintamista (TK 2)	52
5.3	Keminformatiikka ja TPASK: mahdollisuudet kemian opetukseen (TK 3)	54
5.4	Johtopäätökset ja jatkotutkimustarve	56
	Lähteet	58
	Liitteet	71

1 Johdanto

Keminformatiikka on poikkitieteellinen ala, jossa tietojenkäsittelytieteen, datatieteen ja laskennallisen kemian osaamista sovelletaan kemialliseen dataan. Käytännössä tietokoneilla käsiteltävät yhdisteiden rakennekaavat, yhdisteiden ja spektrien tietokannat, laskennallisen kemian pilvipalvelut ja molekyylihallinnusohjelmistot perustuvat keminformatiikkaan (Ebejer ym., 2013; Jeliaskova & Kochev, 2011; Wild, 2021, s. 32.) Keminformatiikka on kuitenkin laajempi käsite, kuin edellä mainitut sovellukset. Se on oma välttämätön osa nykyistä kemian soveltavaa tutkimusta kasvavasta ja suuresta määrästä kemiallista dataa laskennallisilla ja koko ajan kehittyvillä menetelmillä. (Lo ym., 2018.)

Keminformatiikka, joka tunnetaan myös kemoinformatiikkana, on tästä huolimatta vähän tunnettu ala, eikä esimerkiksi tutkimusalueen nimestä ole yhtenäistä nimitystä. Yliopisto-opetuksessa on usein puhuttu siitä, mitä taitoja tulevaisuuden ammattilaiset tarvitsevat ammatillisen pätevyyden ja elinikäisen oppimisen suhteen (Bachelier, 2015; Louko & Blomqvist, 2018). Usein tästä puheesta kuitenkin unohtuu kemiallisten tieteiden ja datatieteen menetelmien, kuten koneoppimisen ja tekoälyn läpimurto tieteissä, kuten kemiassa (Zdrasil & Guha, 2022), mutta tästä huolimatta keminformatiikka ymmärretään vielä heikosti hyödynnettynä tieteenhaarana kemian opetuksessa. Usein keminformatiikan sovellukset tunnetaan nimeltä tieto- ja viestintäteknologiana, välineenä jotakin muuta työtä varten, jolloin keminformatiikan perusteet ja sen menetelmällisyyden mahdollisuudet jäävät hyödyntämättä. Opetuksessa keminformatiikka on kuitenkin aina läsnä piirto-ohjelmistoissa linjanotaatioina, molekyylihallinnoissa ja spektriennusteina esimerkiksi Avogadro-ohjelmistossa ja monissa muissa ohjelmistoissa (Hanwell ym., 2012; Kaushik, 2014).

Keminformatiikan roolin ennustetaan kasvavan osana luonnontieteellistä tutkimusta ja teollisuuden tuotesuunnittelua koneoppimisen, data-analytiikan ja tekoälyn rinnalla (Takeda ym., 2020) – täten myös opetuksen tutkimuksen tulee pyrkiä selvittämään keminformatiikan mahdollisuuksia ja kysyntää tulevaisuuden ammatillisten taitojen suhteen yliopisto-opetuksessa ja -opinnoissa. Erityisesti taidot hyödyntää keminformatiikkaa ja tunnistaa sen mahdollisuudet opetukselle ovat etusijalla. Datatieteellisten menetelmien tärkeyttä yleensä perustellaan sillä, että

niiden avulla voidaan laajentaa tieteen tekemisen näkökulmia enemmän kuin nojaten vain ihmisten väliseen vuorovaikutukseen, ja samalla täydentää pluralistisesti tieteen ymmärtämistä (Clauset ym., 2017; Fortunato ym., 2018). Tätä tukee kemian ja muitten luonnontieteiden muuttuva luonne teknologisten innovaatioiden ja tieteellisten muutosten rajapinnassa. Erityisesti tutkimusdatan määrän kohtaamisesta puhutaan tieteessä neljäntenä paradigmana. Tieteilijöihin ja työnkuvaan liitetään usein termit e-Scientist ja e-Research. (Goble & Roure, 2009; Meng ym., 2016.) Tällaisten tieteen kehityksen teorioiden mahdollisuuksia voidaan tutkia eri tavoin, mutta laadullista kemian opetuksen näkökulmaa aiheeseen ei ole aiemmin tutkittu. Tästä huolimatta keminformatiikan yliopisto-opetusta on tutkittu jo ensimmäisen vuoden opiskelijoille suunnatuista opinnoista alkaen (Esim. Kadtsyna ym., 2022), mutta valitettavasti keminformatiikan osaamiselle ei ole laajaa opetustarjontaa, vaikka useille keminformatiikan opetusta koskeville julkaisuille on jo vuosia ollut selkeä tarve (Guha & Willihagen, 2020; Wild, 2013).

Tutkielma keskittyy tarkastelemaan keminformatiikkaa yliopisto-opetuksen kontekstissa, koska ala on hyvin monitieteellinen, johon lukeutuu monia tieteitä ja käytäntöjä (Kim ym., 2021). Täten myös kemian opetuksen mahdollisuuksia keminformatiikan asiantuntijoiden tietoon ja kokemukseen pohjautuen etsitään alan tutkijoilta. Tähän tutkielma pyrkii syventymällä keminformatiikan perusteisiin, luonteenpiirteisiin, käytänteisiin ja menetelmiin teoriaohjaavalla sisällönanalyysillä, eli laadullisella tutkimuksella, haastatellen kolmea erilaista keminformatiikan asiantuntijaa. Tällaista laadullista tutkimusta aiheesta ei ole aiemmin tehty ja se voi tuoda metodina esiin tietoa kemian opetuksen uusista mahdollisuuksista. Asiantuntijoiden työnkuvan ja haastattelun tarkoitus on kuvata luotettavasti, miten keminformatiikkaa voidaan hyödyntää ja miten tieteenala näyttäytyy kemian työssä. Toisaalta on tehty joitain haastattelututkimuksia keminformatiikan asiantuntijan työnkuvasta ja siitä, mitä ajankohtaista keminformatiikan tutkimuksessa tapahtuu tai miten kemiallista tietoa käsitellään osana kirjastotiedettä (Baykoucheva, 2015; McEwen & Li, 2014), mutta aihetta ei olla aiemmin tarkasteltu keminformatiikan roolia kemian opetuksen näkökulmasta. Täten aineistonkeruun metodiksi valittiin avoin haastattelu (Hirsjärvi & Hurme 2008, s. 45–46), jolla pyritään ymmärtämään vapaammin ja epästrukturoidusti, miten keminformatiikan merkitystä ja hyödyntämistä voidaan paremmin ymmärtää kemian opetuksen kontekstissa.

Kemian opetuksen näkökulmaan kuuluu oleellisesti kemian opettajakoulutus, ja täten tutkielmassa hyödynnetään pedagogista TPASK-mallia (Jimoyiannis, 2011) siitä, miten opettajat voivat vaikuttavan oppimisen lähestymistapaan perustuen hyödyntää tieteellistä tietoa ja teknologiaa opetuksessaan. Se miten keminformatiikkaa opitaan ja voi hyödyntää opetuksessa, peilataan sisällönanalyysin tuloksista kirjallisuuteen TPASK-mallia hyödyntäen.

Tutkielma tuottaa alati teollistuvan ja yhteiskunnallisesti merkityksellisen, mutta kemian opetuksen tutkimuksessa vähän tutkitun alan, keminformatiikan parempaa ymmärrystä uudesta näkökulmasta. Opetuksen näkökulmaa tarvitaan kemian tieteen uusien menetelmien parempaan ymmärtämiseen ja tarkasteluun, jolla voidaan taata oppimisen ja ammattitaidon laatu yliopistokoulutusjärjestelmissä kemian tiedon suhteen ja vastata kemian muuttuvaan luonteeseen. Teoreettisessa viitekehyksestä syvennyttään keminformatiikkaan, aiempaan opetuksen tutkimukseen ja TPASK-malliin (Luku 2). Menetelmissä (Luku 3) esitetään tutkimuskysymykset, tarkastellaan haastattelumenetelmää ja sisällönanalyysia osana laadullista tutkimusotetta ja sen luotettavuutta. Tuloksista (Luku 4) raportoidaan tutkimuksen sisällönanalyysin teemat ja luokittelut. Keskustelu ja johtopäätökset (Luku 5), käydään keskustelua tuloksista, merkityksestä ja luotettavuudesta. Lopuksi esitetään tutkimuksen johtopäätökset ja jatkotutkimustarpeet.

2 Teoreettinen viitekehys

Tässä luvussa kuvataan tutkimukselle keskeisiä aihealueita. Keminformatiikan perusteet, käytänteet ja yleisimmät sovellukset sekä tietojenkäsittely- ja datatieteen menetelmällisyys osana keminformatiikkaa. Lopuksi esitellään miten keminformatiikkaa on hyödynnetty kemian opetuksen tutkimuksessa sekä esitetään TPASK-malli, jota hyödynnettiin aineiston luokittelussa.

2.1 Keminformatiikan perusteet

Keminformatiikassa kemiallista informaatiota, kuten aineen rakennetta ja ominaisuuksia, kuvataan, tallennetaan, arvioidaan tai muunnetaan tietoteknisillä menetelmillä. Kemiallisia ilmiötä, kuten aineiden välistä vuorovaikutusta tai reaktiomekanismeja, arvioidaan hyödyntäen erilaisia ohjelmistoja, tilastollisia menetelmiä, algoritmeja kuten koneoppimista tai rakennehakualgoritmeja sekä kemian tietokantoja ja visualisointiohjelmistoja. (Wishart, 2016.)

Tietotekniikan hyödyntämisestä kemiassa on tullut yleistä tieteellisen tiedon määrän kasvaessa, ja luonnontieteissä tarvitaan yhä enemmän tietojenkäsittelytieteen menetelmin suoritettavia kvantitatiivisia mallinnuksia sekä analyyseja. Niiden tarkoitus on nopeuttaa hypoteesien testaamista laskennallisilla menetelmillä suuresta määrästä dataa ennen kokeellisia menetelmiä ja suuremmassa mittakaavassa, jossa olennaista on myös keminformatiikan mallien arviointiin, kuten yhdisteiden seulonnassa (Ash & Hughes-Oliver, 2022). Täten keminformatiikka on kemian informatiikkatiedettä eli soveltavaa tietojenkäsittelytiedettä erilaisin datatieteellisin menetelmin, kuten datalouhinnalla ja eri tilastollisilla menetelmillä, joilla pyritään ratkaisemaan kemian tutkimuksen ongelmia (Gressling, 2021). Keminformatiikan etuna empiiriseen aineiston hankkimiseen on mallintaa ilmiöitä ennen laboratoriotyöskentelyä luotettavilla matemaattisiin malleihin perustuvilla simulaatioilla säästään aikaa ja materiaalisia resursseja sekä alentaen taloudellisten kustannusten riskiä erityisesti lääketeollisuudessa (esim. Jarada ym., 2020; Lee ym., 2017).

Uudet tietojenkäsittelyteknologiat ovat osa tätä prosessia, sillä uusien algoritmien käyttöönotto koneoppimisen ja tekoälyn ohella tilastotieteellisen datan tulkintaan on arkea monilla aloilla (Mitchell, 2014). Keminformatiikan kehitys on nopeatahtista,

sillä tekniikan edistyksestä huolimatta keminformatiikan asiantuntijat ovat havainneet selkeitä ongelmia ja haasteita laskennallisten menetelmien sekä empiirisen tutkimuksen lähestymistavan eli de novo tutkimusotteen soveltamisesta uusiin haasteisiin, kuten COVID-19 hoitojen löytämiseen (Guha ym., 2021). Tätä lähestymistapaa lääkeaineiden tutkimuksessa kutsutaan tietokoneavusteiseksi lääketutkimukseksi (CADD, engl. Computer-aided Drug Discovery), jossa yhdisteiden rakenteesta riippuvat biologiset ja kemialliset ominaisuudet toimivat bioaktiivisuuden mallinnuksen lähtöarvoina (Zin ym., 2020).

2.1.1 Kemiallisen tiedon käsittely

Kemian tietoa voidaan käsitellä tietoteknisesti, kun tietojenkäsittelytieteelliset menetelmät, kuten aineiston kuvaukset, selitteet, määritelmät ja hakutoiminnot muodostavat perustan, millä tiedeyhteisö hyödyntää tietojärjestelmiin varastoitua tietoa eli informatiikkaa (McEwen & Li, 2014). Kemian tiedon järjestelyä voidaan myös ymmärtää englanninkielisellä termillä data curation, jossa standardoiduilla menetelmällisyydellä tiedosta, kuten kemiallisesta tiedosta, saadaan luotettavuudeltaan korkealaatuista (Fourches ym., 2010). Data voi olla puutteellista, virheellistä, epämääräistä, epäolennaista ja/tai pelkistymätöntä, jota voidaan ehkäistä tilastotieteellisillä ja laskennallisilla menetelmillä (Fourches ym., 2015). Mittausvirheet ja virhemarginaali vaikuttavat kokeelliseen dataan, mutta virheitä voi myös syntyä, jos käytetään erilaisia menetelmiä kuvaamaan yhdisteiden rakennetta tietokoneelle (Young ym., 2008), mikä korostaa tiedon luotettavuutta (Fourches ym., 2010).

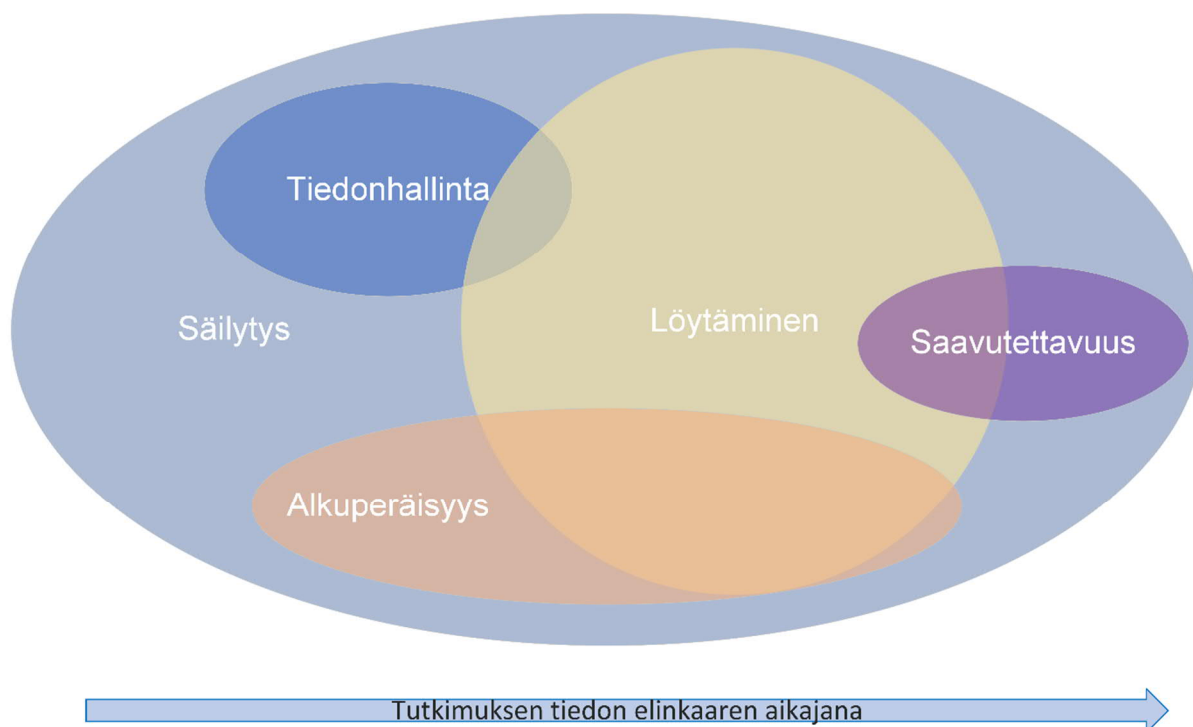
Käsitys tieteellisen tiedon käsittelystä tietojenkäsittelytieteessä on linjassa esimerkiksi Reisigin (2020) käsityksen kanssa mallintaa tietojenkäsittelytiedettä ja datatiedettä informatiikkana, jota ohjaavat samanlaiset periaatteet, joita luonnontieteellisessä mallintamisessa on. Tämä ei ole varsinaisesti uusi idea, koska tietojenkäsittelytieteiden soveltamiselle luonnontieteissä on jo laadittu periaatteita, jotka koskevat tutkimuksen tekemisen piirteitä sekä tarkoitusta, kuten ongelmien ja tavoitteiden rakentamista, arviointia, teoriointia ja oikeuttamista mallien, konstruktioiden, metodien ja installaatioiden avulla (March & Smith, 1995).

Keskeistä keminformatiikalle ja tätä myötä tiedonkäsittelylle on tietojenkäsittelytieteiden tapa mallintaa maailmaa, koska ne ovat keskeisesti

mallintamista ja simuloimista (Tolk, 2018). Esimerkiksi molekyylien enumerointi, eli listaus perustuu systemaattisesti luotuun listaan molekyyliä aiemmasta datasta ja teoriasta. Esimerkiksi orgaanisten molekyylien joukon voi laajentaa miljoonista miljardeihin molekyyliin enumeroinnilla. (Ruddigkeit ym., 2012.) Tutkimukseen tarvitaan menetelmiä, joilla tiedonkäsittelyä voidaan myös kemiassa ja keminformatiikassa standardoidusti hyödyntää osana luonnontieteellistä tutkimusta. Luonnontieteellisten mallien ollessa matemaattisia ovat myös keminformatiikan menetelmät tiedonkäsittelyssä matemaattisia, jolla voidaan pyrkiä mallintamaan tietojenkäsittelytieteen luonnetta algoritmeilla. (Reisig, 2020.)

Avoin tiede ja täten julkinen tieto ja tutkimuksen jakaminen mahdollistavat kemiallisen datan käyttöönoton laajemmin moni- ja poikkitieteellisestä näkökulmasta. (Bird & Frey, 2013) Tämän vain keminformatiikka voi mahdollistaa alati kiihtyvässä tiedon digitalisoinnin ajassa, kun tietoa tuotetaan koko ajan enemmän erilaisiin tietokantoihin ja julkaisuina esimerkiksi kombinatoriaalisessa kemiassa (Prakash & Gareja, 2010). Tutkimukselle ja datalle on oma elinkaarensa, ja tämän elinkaaren tehokas ja merkityksellinen tutkimuskäyttö vaatii tiedonkäsittelyn kulttuuria, jonka Bird ja Frey (2013) jakavat seuraaviin luokkiin ja prosesseihin (Kuva 1):

- Säilytys on periaate, jossa data säilyy löydettävänä ja hyödynnettävänä sen koko elinkaaren ajan myös silloin kuin tutkija ei sitä hyödynnä.
- Tiedon hallinta tutkimuksen elinkaareissa tarkoittaa tiedon jakamisen ja tallentamisen suunnitelmia sekä mahdollista täydentämistä.
- Alkuperäisyys tarkoittaa tutkimustiedon kannalta sitä, että data on tallennettu osana tutkimusprosessia, joka on tutkimusmenetelmältään kuvattu niin, että tutkimusprosesseilla se on toistettavissa. Data on siis autenttista ja osa metadataa.
- Löytäminen on tieteellisen tiedon datatieteellistä löytämistä semanttisesti ja aggregoiden trendejä ja syy-seuraussuhteita erilaisista datajoukoista.
- Saavutettavuus tarkoittaa datan käyttöoikeutta kuin sen antamista erilaisilla välineillä, kuten embargo menetelmällä tai lisensseillä, jonka päämääränä on avoin tiede, data ja tiedon vaikuttavuuden maksimointi.



Kuva 1. Tutkimuksen elinkaaren aikana tietoa käsitellään monipuolisesti. Mukailtu ja suomennettu Birdin ja Freyn (2013) julkaisusta.

Tiedon elinkaaren ajattelun edistämisestä ja tiedon määrän kasvamisesta puhutaan myös paradigmaattisena muutoksena, jota luonnehditaan tieteen neljänneksi muutokseksi, jota on edeltänyt empiirisen todisteen, tieteellisen menetelmän ja laskennallisten menetelmien paradigmat. Paradigmojen sarjassa, tieteen neljäs paradigma tarkoittaa monitieteellistä datatieteen muutosta tieteen kehityksessä, jossa tietoteknisillä teknologioilla muokataan aineistohallintaa ja -käsittelyä. Se perustuu avainteknologioihin hyödyntäen tietokantoja ja data-analysiohjelmistoja sekä tietoputki-, visualisointi- ja pilvipalveluja. Tähän liittyy käsitys siitä, kuinka Big Data ja tiedonhallinta tietopankeissa ratkaistaan tieteen keinoin elinkeinoelämän sijaan, ja osana tätä kehitystä syntyy datapainotteisia tieteenaloja johtaen muutokseen tieteen luonteessa. (Hey ym., 2009, s. 8–9, 210.) Big Data käsitteellä viitataan tieteissä käsitykseen suurista ja alati kasvavasta tieteellisen tiedon määrästä, jonka piirteinä voidaan pitää sen määrää, uuden tiedon tuottamisen nopeutta ja laadun vaihtelevuutta. Keminformatiikan voidaan nähdä täyttävän tämän kuvauksen piirteet, sillä sen keskeiset sovellukset ovat samat kuin edellä mainitun datapainotteisen tieteenalan teknologiat. Nämä teknologiat ja keminformatiikka ovat

yhdessä Big Datan ja tekoälyn kanssa erityisenä kiinnostuksen kohteena tutkijoille erityisesti lääketieteellisessä kemiassa, jossa otetaan keskeisesti huomioon kemiallisten yhdisteiden monimutkainen ja laaja joukko. (Griffen ym., 2018; Hu & Bajorath, 2017.)

2.1.2 Kemiallisen rakennedatan tiedostotyypit ja linjanotaatiot

Keminformatiikan ensimmäisenä haasteena oli kuvata paperille luodut rakenteet tietoteknisesti saataville osana tietokantateknologiaa. Ensimmäisissä sovelluksissa Yhdysvalloissa 1960-luvulla, alettiin hyödyntämään linjanotaatioita, jotka merkkijonoina rakenneinformaatiota tietokoneelle. Yhtenä etuna oli, että tämä tietoformaatti vei hyvin vähän muistitilaa, joka oli hyvin kallista tietokoneille 1900-luvun puolivälissä. (Gasteiger & Engel, 2006, s. 26.) Huomattiin kuitenkin jo varhain, että tietokoneella saman merkkijonon voi tulkita tai ymmärtää eri tavoin. Ongelmista merkityksellisimpiä ovat graafien ja alagraafien isomorfian ongelmat, jotka johtavat rakenteiden erottamattomuuteen ilman aliverkon isomorfiaan erikoistuneita algoritmeja ja tietoteknisiä nimeämismenetelmiä. (Wild, 2021, s. 23.) Tietokantoihin tallennetulle kemialliselle tiedolle ja sen rakenteiden oikeelliselle esittämiselle onkin annettu paljon painoarvoa, koska pienetkin virheet voivat vaikuttaa tutkimusten virhemarginaaliin merkittävästi erityisesti mallinnuksessa (Fourches ym., 2010).

Kemiallisten yhdisteiden rakenteiden esittämiseen erilaisissa formaateissa on kehitetty erilaisia järjestelmiä, joista yksinkertaisiin linjanotaatiohin lukeutuu SMILES ja InChI. Tiedostopohjaiset ratkaisut, kuten CML (engl. Chemical Markup Language), joka perustuu XML-tiedostotyyppiin, ja jota voidaan käyttää myös spektrien ja reaktioiden esittämiseen (Kuhn ym., 2007). MDL Molfile on suosittu taulukkolaskentaan perustuva tiedostotyyppi siitä johdetun SDF- tiedoston (engl. Structure-Data File) rinnalla, koska sitä tukee moni ohjelmisto ja sillä voidaan esittää yhdisteen rakenne XYZ-koordinaatiossa (Kufák ym., 2022). Taulukossa 1. kuvataan tiedostotyyppejä ja linjanotaatioita, joilla tietokoneella yhdisteiden rakenteita voidaan selittää.

Taulukko 1. Yleisimpiä keminformatiikan tiedostotyyppejä ja linjanotaatioita.

Nimi	Teknologia	2D-rakenne	3D-rakenne
MDL Molfile	Tekstitalukko	Kyllä	Kyllä
CML	XML-merkintäkieli	Kyllä	Kyllä
SDF	XML-merkintäkieli	Kyllä	Ei
SMILES	Linjanotaatio	Kyllä	Ei
InChI	Linjanotaatio	Kyllä	Ei
InChIKey	Linjanotaation hajautusjakso	Kyllä	Ei

SMILES on yksinkertaisin linjanotaatio, joka perustuu ASCII-merkistön lineaariseen esitystapaan yhdisteen atomeista, sidoksista, sidospituuksista ja sidoskulmista. Tämän linjanotaation tulkinta perustuu graafiteoriaan eli verkkoteoriaan. Yhdisteitä mallinnetaan verkkoteorian olioilla, jossa atomit ovat solmuja (engl. Node) ja sidokset viivoja (engl. Edge), jotka yhdistävät atomit kemiallisilla sidoksilla kaksiulotteiseksi verkkomaiseksi esitykseksi. Verkkojen lukemiselle on standardimenetelmät, joissa ilmenee atomien kemiallinen merkki, mahdollinen hybridisaatio, sidokset muihin atomeihin sekä atomien numerointi IUPAC-nimeämissääntöjen mukaan. (Wild, 2021, s. 22–23.)

Verkkoteorian linjanotaatiojärjestelmää luetaan liitostaulukolla, jossa jokaisella atomilla on oma sarake ja rivi, mistä niille tunnusomainen tieto voidaan lukea. Liitostaulukkoa hyödyntämällä tietokone voi päätellä linjanotaation tarkoittavan tiettyä rakennekaavaa. Tähän tehtävään tarvitaan kahta algoritmia, joista ensimmäinen, CANON antaa jokaiselle atomille sen kemiallisen merkin eli symbolin ja määrittää graafin, jossa atomit ovat solmuja ja sidokset viivoja, ja jokaiselle atomille määritetään numerot perustuen IUPAC nimeämissääntöihin. GENES on toinen algoritmeista ja ottaa huomioon symmetriat ja erilaisia sääntöjä, joilla voidaan erotella samankaltaiset, mutta erilaiset yhdisteet ja niiden verkkoteorialla muodostetut linjanotaatiot toisistaan. Yhdistelmänä molemmista algoritmeista

käytetään nimitystä CANGEN, joka pyrkii siis noudattamaan IUPAC-nomenklatuuria ja kuvaa atomit solmujen ja kulmien välisillä riippuvuuksilla hyödyntäen muuttumattomia solmuarvoja: atomin sidoslukua, muitten kuin atomiin sitoutuneiden vetyjen lukumäärää, järjestyslukuja, varauksen merkkiä, absoluuttisella varausta sekä sitoutuneiden vetyjen lukumäärää. (D., Weininger, A., Weininger, & J.L., Weininger, 1989.)

Tietokone ja ohjelmisto muodostavat linjanotaation muodossa puumaisen tietorakenteen, jossa hyödynnetään vain symbolista kieltä atomien kemiallisista merkeistä sekä merkkejä =, (,), [,], #, @, /, \ ja c. Merkit kuvaavat sidosten luonteita ja sidoskulmia eli atomien suhteellista avaruudellista asemaa toisiinsa nähden. Esimerkiksi CCCC luetaan butaanina, koska hiiliä on neljä, ne ovat sitoutuneet toisiinsa, ja tulkitaan implisiittisesti, että niissä on vetyjä, jotka lisätään hiiliin perustuen kemiallisen sitoutumisen sääntöihin. CCC(=O)C tarkoittaisi 2-butanonia, jonka ketoniryhmä luetaan kaksoissidoksena toisesta hiilestä happeen. Rakenteen tarkka kuvaaminen ja lukeminen on edellytys sen virheettömälle esittämiselle ja hyödyntämiselle tietokannoissa, jossa hakutulosten nimikkeiden täytyy muodostaa järkeviä ja virheettömiä tuloksia. Rakenteiden kuvaamista voidaan itsessään hyödyntää haussa, josta hyvänä esimerkkinä on SMILES-kaavan sukulainen, linjanotaatio SMARTS, jolla voidaan etsi molekyylirakenteita tietokannoista. SMARTS on luotu SMILES:n ylijoukoksi, eli ne ovat SMILES linjanotaatioita, joita on laajennettu lisäsäännöillä samankaltaisten rakenteiden tai täysin samojen rakenteiden etsimiseen SMILES-kaavoilla esitetyistä nimikkeistä tietokannassa. (Jeliazkova & Kochev, 2011.)

InChI, eli englanniksi International Chemical Identifier on kaava, jonka tavoitteena on kehittää kansainvälinen standardi kemiallisten aineiden tietoteknisesti luettavalle nomenklatuurille (Yerin ym., 2013). Se on linjanotaatio, joka vanhoihin kaavoihin, kuten SMILES-kaavaan verrattuna, huomioi tarkemmin molekyyliden stereokemialla ja tautomeriaa sekä ehkäisee saman linjanotaation moninaista tulkintaa. InChI luetaan systemaattisesti linjanotaatioilla, jotka on rivitetty omille tasoilleen: päätasolla ilmoitetaan linjanotaationa kemiallinen kaava, sidokset ja vedyt. Varaustasolla positiiviset ja negatiiviset varaukset, ja stereokemiallisella tasolla kaksois- ja kolmoissidokset, tetraedrinen stereokemia ja stereokemiallinen tyyppi. (Heller ym., 2013.)

Silti InChI on liian pitkä merkkijono kirjoittaa manuaalisesti kokonaisuudessa hakukoneille. Ratkaisuksi on kehitetty lyhyt InChI-linjanotaatiosta johdettu linjanotaatio nimeltä InChIKey, jolla voidaan tehdä luotettavampia hakuja tietojärjestelmissä ja hakukoneilla. InChIKey-kaavaa ei voida tulkita täysin eksplisiittisesti ja monitulkinnaisuus sen morfologiasta säilyy ongelmana, mutta hakukoneissa se toimii melko hyvin käyttäjän tarkoitukseen löytää haluttu yhdiste, sillä se rajaa haetun tiedon määrää voimakkaasti eikä siitä johdettu kaava katkea yhtä helposti kuin pidemmät InChI linjanotaatiot, mikä johtaisi lauseopillisiin virheisiin kaavaa luotaessa. InChI:n ja InChIKey:n merkitys korostuu kansainvälisellä tasolla auttaen tiedeyhteisöä ylittämään kielimuureja ja samalla se parantaa kansainvälistä tiedeviestintää ymmärrettävien yhdisteiden tietoteknisellä kielellä IUPAC nomenklatuurin rinnalla. (Heller ym., 2015.) InChIKey tunnetaan myös InChI-avaimena (Wild, 2021, s. 22).

2.1.3 Tietokannat

Kemiallisten aineistojen ylläpito, talletus, haku ja vienti tapahtuu pääosin tietokantateknologiaan perustuvilla ohjelmistoilla ja infrastruktuurilla niin tutkimustarkoituksiin kuin elinkeinoelämän harjoittamiseen. Kemiallisten aineistojen tietokantoja on perustettu kemian tutkimusta edistämään tallentamalla yhdisteiden rakenteita tai mittausdataa, kuten erilaista spektroskooppista dataa yhdisteistä, eri alan tutkijoiden hyödynnettäväksi (Fletcher ym., 1996). Yritykset ovat perustaneet kemiallisia tietokantoja erityisesti lupaavien orgaanisten lääkeaineiden löytämiseksi, mutta avoimeen dataan perustuvia tietokantoja löytyy myös (Sander ym., 2015). Lääkeainetutkimukseen tarkoitettuja tietokantoja kutsutaan virtuaalisiksi kemian tietokannoiksi (engl. Virtual Chemical Libraries), jotka on seulottu niiden reaktiivisuuteen ja mallinnettuihin biokemiallisiin ja fysiologisiin ominaisuuksiin perustuen keminformaattiseen analyysiin suuremmasta joukosta yhdisteitä (Walters, 2019).

PubChem (Kim ym., 2019), ChemSpider (Pence & Williams, 2010) ja ChEMBL (Mendez ym., 2019) ovat yleisesti hyödynnettyjä ilmaisia tietokantoja, joista voidaan hakea tietoa kemiallisista aineista, kuten yhdisteistä. Maksullisista tietokannoista CAS:n palvelut, kuten SciFinder, Elsevierin Reaxys ja Web of Science ovat tunnetuimpia ja laajimpia (Bharti ym., 2016). Tietokantojen nimikkeissä on tietoa

niiden rakenteesta niin kaksiulotteisesti kuin kolmiulotteisesti. Yhdisteiden fysikaaliset, kemialliset ja biokemialliset ominaisuudet on usein kuvattu tarkasti perustuen kokeellisiin tuloksiin (Kim ym., 2019). Lisäksi spektrejä, patentteja, tutkimustuloksia ja monia muita faktoja voidaan säilöä tietokantaan joko empiiriseen dataan tai laskennalliseen kemiaan ja kemometriaan perustuen (Wolfender ym., 2019). Taulukkoon 2 on koottu tiedot julkisista tietokannoista.

Taulukko 2. Julkistien kemian tietokantojen keskeinen sisältö ja laajuus.

Tietokanta	Laajuus	Datan lähde	Datan sisältö	Lähde
PubChem	• 293 milj. yhdisteen kuvausta • 111 milj. eri yhdistettä • 1,2 milj. bioaktiivisuustutkimusta (8.2020)	• Tieteelliset julkaisut • Patentit • Muut tietokannat • Teollisuus	Pienmolekyylien data: rakenne, bioaktiivisuus ja spektrit	Kim ym. (2019)
ChemSpider	• 128 milj. yhdistettä • 277 eri lähdettä	• Tieteellisistä julkaisut • Tutkimusryhmät • Muut tietokannat • Teollisuus	Yhdisteiden rakennekaavat ja spektrit (esim. NMR ja MS)	Pence & Williams, (2010)
ChEMBL	• 67 000+ julkaisua ja patenttia • 1,8 milj. yhdisteen bioaktiivisuuden tutkimusta	• Tieteelliset julkaisut • Patentit	Pienmolekyylien rakenne ja bioaktiivisuuden data sekä metatietomenetelmistä	Mendez ym. (2019)

PubChem on maailman suurin julkinen kemiallisten yhdisteiden tietokanta kemiallisen tiedon jakamiseen ja tiedonhallintaan. Se sisältää tietoa erilaisten

bioaktiivisten yhdisteiden kemiallisista ja fysikaalisista ominaisuuksista. (Kim, ym., 2019.) PubChemissa nimikkeitä kuten yhdisteitä, aineita ja biologista tietoa myös indeksoidaan, näitä tietoja voidaan hakea hyödyntäen tunnisteita, yhdisteiden nimiä tai linjanotaatioita tai rakennehakua hyödyntäen, ja oikean nimikkeen löydettyä sen tiedot voidaan tulostaa eri formaateissa, kuten kuvina, XML- tai CSV-tiedostona. Tällaista ominaisuutta voi hyödyntää datankeruussa ja datatieteellisissä lähestymistavoissa kemian ongelmiin, kun yhdisteitä voi etsiä ja seuloa julkisista tietopankeista tutkimustarkoitukseen käytettävää ohjelmistoympäristöä varten. (Kim ym., 2015.)

ChemSpider tarjoaa hyvin samanlaisia palveluita kuin PubChem, mutta ChEMBL selkeästi tarjoaa tietoa liittyen biokemiaan ja farmasiaan. Kaikki tietokannat tarjoavat ohjelmointirajapintoja julkisten aineistojensa hyödyntämiseen, mutta nimikkeiden kohdalla poikkeuksena muihin tietokantoihin, PubChem-tietokannassa on oma molekyylihallinnusohjelmisto (PubChem3D). Kaikkia tietokantoja ja niiden erilaisia hakukoneita ja indeksointijärjestelmiä halutaan kehittää palvelemaan koko tiedeyhteisöä ja tämän takia niiden nimikkeissä on usein viitteitä patentteihin ja julkaisuihin. Tämä tavoite on osa laajempaa kokonaisuutta, jossa keminformatiikkaa kehitetään entistä paremmaksi myös kirjastotieteellisestä näkökulmasta. (McEwen & Li, 2014.)

2.2 Keminformatiikan käytänteet

2.2.1 Tiedonhallinta

Tiedonhallinnalla voidaan viitata moneen erilaiseen kokonaisuuteen tiedon hallinnallisista menetelmistä (McEwen & Li, 2014; Nevalainen & Toivonen, 2007). Tiedonhallinnan termiä sivuttiin jo aiemmin osana Birdin ja Freyn (2013) tutkimusta luvussa 2.1.1., ja suomenkielisessä muissa tutkimuksissa tiedonhallinta usein katsotaan osana termiä Informatics (Kuusisto-Niemi & Saranto, 2009). Toisaalta suomenkielinen termi, aineistonhallinta, sisältää tiedonhallinnan piirteitä, mutta laajentaa käsitystä datan, metadatan ja menetelmien hallintaan osana aineiston tulkintaa ja käsittelyä sen elinkaaren ja tutkimuksen aikana. Aineistonhallinta on osa suomalaisten yliopistojen strategiaa (Tietoarkisto, ei pvm.). Tässä tutkielmassa hyödynnetään termiä tiedonhallinta kuvaamaan kyvykkyydeksi hyödyntää keminformatiikan dataa, menetelmiä ja käytänteitä. Konkreettisesti tiedonhallinnalla

voidaan tarkoittaa tietokantojen, ohjelmistojen ja erilaisten tietotyyppien hyödyntämistä sekä organisointia tutkimustehtäviin (Gasteiger, 2006).

Keminformatiikan tietokannat ja työkaluohjelmistot, jotka sisältävät kemiallisesti merkittävän tiedon yhdisteen nimeämisestä, ominaisuuksien esityksestä ja rakenteen esityksestä muodostavat keminformatiikan perustan (Gasteiger, 2006). Faulon ja Bender (2010, s. 343) kuvailevat keminformatiikan työkaluohjelmistojen koostuvat pääsääntöisesti algoritmeista, jotka muodostavat tietoteknisen perustan kemiallisen tiedon tulkinneille. Keminformatiikan tavoite on käsitellä kemiallista tieteellistä tietoa ihmiselle ja tietokoneelle ymmärrettävässä muodossa. Tiedon kuvaaminen, käsittely ja hallinta ovat avainasemassa. Aluksi tietokoneelle on kommunikoitava kemiallinen tieto, jota sen toivotaan käsittelevän, mikä on lähes aina kemiallista myös bioinformatiikassa. Tämän tiedon tulee perustua loogiseen ja systemaattiseen käsittelyyn, jotta tietokoneet voivat tulkita kemialla oikealla tavalla. (Wegner, ym., 2012.)

2.2.2 Avoin lähdekoodi ja avoin data keminformatiikassa

Ohjelmistokehityksessä keminformatiikan tarpeisiin korostetaan akateemisena periaatteena avointa lähdekoodia ja avoimen datan merkitystä. Avoimuuden merkitys korostuu tieteellisten julkaisujen jakelussa, ohjelmistojen jakelussa sekä resurssien, kuten tietokantojen vapaassa käytössä. Avoimella lähdekoodilla (engl. Open Source) tarkoitetaan ohjelmistokehitystä ja -jakelua, jossa ohjelmiston, ohjelman tai ohjelmien lähdekoodi on saavutettavissa ja hyödynnettävissä lisenssillä, joka mahdollistaa sen hyödyntämisen kaikille käyttäjille ilman rajoituksia, kuten maksuja (Open Source Initiative, 2023).

Keminformatiikan ohjelmistokehityksen periaatteita ja hyviä käytänteitä, jotka perustuvat avoimeen koodiin, dataan, ohjelmistokehitykseen ja julkaisuun on muotoiltu jo pitkään johtavien akateemisten tutkijoiden ryhmässä nimeltä Blue Obelisk (O'Boyle ym., 2011). Keminformatiikan julkaisut vetoavat avoimeen lähdekoodiin ja dataan julkaisuissaan ja tiedon luotettavuudessa, koska sen uskotaan edistävän tutkimusta parhaalla mahdollisella tavalla. Tällä pyritään työn toistettavuuteen korostaen menetelmällisyyden pätevyyden arviointia. (Clark, 2019). Ohjelmistokehitystä on tutkittu myös sen tarkoituksellista katsauksessa, jossa selvisi kuusi merkittävää kategoriaa: uuden ohjelmiston kehitys, aiempien ohjelmistojen tai

ohjelmien tai menetelmien edistäminen, käyttäjäkokemuksen edistäminen, avoimen tieteen ja lähdekoodin tukeminen, informaation tarpeen täyttäminen sekä oppimisen ja opetuksen tukeminen (Pernaa ym., 2023).

Hoyt ym. (2023) mainitsee *Journal of Cheminformatics* -lehden julkaisuissa erilaisia haasteita avoimelle tieteele. Ohjelmistojen puutteellinen lähdekoodi ja asennus, tutkimusvaiheiden ohjelmoinnin epäselvä kuvaus, ohjelmistolisenssien lähdekoodin rajoitukset, ja dokumentaation puutteet ovat silti keskeisiä ongelmia julkaisutoiminnassa. Tämän lisäksi aineistohallinta julkisiin talletuspalveluihin tieteellisten julkaisujen kuvauskantojen (esim. Zenodo) talletuspalvelujen sijaan voi vähentää datasta saatavaa hyötyä. Tähän ratkaisuna lehti pyrkii kysymään arviointiprosessissa seitsemän kysymystä, jotka parantavat parempia tiedonhallinnan menetelmiä ja standardeja kemiallisen tiedon käsittelylle sekä ohjelmointiprosessille. (Hoyt ym., 2023.)

2.2.3 Molekulaariset deskriptorit

Keminformatiikassa yhdisteen rakennetta ja ominaisuuksia voidaan kuvata eksplisiittisesti eli yksiselitteisesti ja implisiittisesti eli epäsuorasti (O’Boyle, 2012). Rakenteet esitetään matemaattisella tiedolla, joka perustuu verkkoteoriaan ja niiden tulkintaan algoritmeilla (Faulon & Bender, 2010, s. 1–27). Rakenteiden esittämisen lisäksi yhdisteiden rakenteesta voidaan mallintaa algoritmeilla ja logiikalla luotettavaa tietoa kyseisten yhdisteiden ominaisuuksista (Randić, 1993).

Yhdisteiden rakenteista ja laskennallisella kemian ennusteista koostetaan molekyylikemian kuvauksia, eli molekulaarisia deskriptoreita (engl. Molecular descriptors) niiden kemiasta. Molekulaariset deskriptorit kuvaavat tietoteknisesti jonkin molekyylin rakennetta tai ominaisuuksia numeerisesti ja ohjelmistolle ymmärrettävässä muodossa (Faulon & Bender, 2010, s. 90). Ne voivat perustua kokeellisiin tuloksiin tai teoreettisiin malleihin. Yleensä lääkekehityksessä molekulaariset deskriptorit ovat teoreettisia, sillä niillä halutaan kuvata yhdistekirjaston bioaktiivisuudeltaan tuntemattomien nimikkeiden ominaisuuksia. Näitä yhdisteitä mallintamalla niistä voidaan valita ominaisuuksiltaan lupaavimmat yhdisteet. Tämä antaa paremmat valmiudet rajata aineistoa, josta tutkia yhdisteitä, joilla on lääketieteellisesti parhaat ominaisuudet jatkotutkimusta varten, jossa tavoitteena on löytää kokeellisesti parhaat lääkeaineet. (Brown, 1998, s. 375).

Molekulaarisia deskriptoreita rakennetaan myös erilaisilla algoritmeilla ja kuvaamisen tasoilla, jotka vaihtelevat erityisesti sen mukaan, mitä halutaan tarkemmin kuvata ja miten eri yhdisteet ovat rakenteeltaan tai ominaisuuksiltaan tietokannassa vertailtavissa (Cereto-Massagué ym., 2015). Erityisesti rakenteiden samankaltainen kuvaaminen bittijonoina linjanotaatioiden sijaan mahdollistaa myös samankaltaisten yhdisteiden löytämisen tietokannasta (Muegge & Mukherjee, 2016).

Lääkeyhdisteiden kokeellinen de novo -tutkimus sekä lupaavien lääkeaine-ehdokkaiden jatkotutkimukset ovat usein kalliita ja työläitä. Keminformatiikan in silico -menetelmiin perustuvalla yhdisteiden virtuaaliseulonnalla (engl. Virtual Screening) pyritään säästämään rahaa ja aikaa seulomalla yhdisteitä tietokannoista teoreettisiin malleihin perustuen ennen kokeellisuutta (Ponder ym., 2014).

Molekulaariset deskriptorit soveltuvat hyvin seulontaan, koska menetelmällisyys on usein samanlainen, laskennallisesti helppoa ja hoituu avoimen lähdekoodin ohjelmistoilla. Tämän lisäksi seulontaa harvoin suoritetaan ilman muita menetelmiä, kuten koneoppimista. (Muegge & Mukherjee, 2016.)

2.3 Keminformatiikan menetelmät

Tässä luvussa syvennyttään keminformatiikan menetelmällisyyteen esittämällä laskennallisen kemian ja keminformatiikan yhteyden, alan vaikuttavimpia ohjelmistoratkaisuja, kemiallisten ominaisuuksien ennustamisen ja lopuksi datatieteellisiä menetelmiä koneoppimisen ja tekoälyn malleilla.

2.3.1 Laskennallinen kemia ja molekyylihallinnus

Laskennallisella kemialla on keskeinen rooli keminformatiikan mallinnuksessa, koska sitä pidetään laajan soveltuvuuden kannalta modernina kemiana, mikä mahdollistaa mm. biokemiallisten reaktioiden ja siirtymämetallikatalyyttien mallintamisen (McArdle ym., 2020). Molekyylien kvantittuneiden energioiden laskeminen teoreettisesti approksimoiduilla malleilla voidaan eri tavoin: empiirisesti, semi-empiirisesti, tiheysfunktioaaleilla tai ab initio -menetelmillä (Chen ym., 2002; Stewart, 2007).

Laskennallista kemiaa voidaan sisällyttää erityisesti molekyylihallinnusohjelmistojen algoritmeilla kuvaamaan molekyylin avaruudellisen rakenteen suotuisinta energiatilaa. Ohjelmistoissa hyödynnetään laskennallisen kemian mallinnustyökaluja

ennustamaan niiden fysikaalisia ja kemiallisia ominaisuuksia. (Wild, 2021, s. 32.) Koska laskennallinen kemia on usein liian raskasta yhdelle laitteelle, nojaa sen toteutus yleensä pilvipalveluihin. (Ebejer ym., 2013.) Erityisesti tutkimuskäyttöön pilvipalveluiden soveltaminen vaatii jatkotutkimuksia, mutta sen suosio ja vaikuttavuus on kasvanut ja moninaistunut viime vuosina laskennallisissa kemiassa (Senyo ym., 2018). Myös keminformatiikassa on luotu pilvipalveluun perustuvia älypuhelinsovelluksia eri tutkimusympäristöihin (Clark ym., 2013).

2.3.2 Yleisiä keminformatiikan ohjelmistoratkaisuja

Seuraavat keminformatiikan ohjelmistoratkaisut esitellään siksi, että ne ovat alan siteeratuimpia ohjelmistoja, jotka sisältävät keminformatiikan datatieteellisen menetelmällisyyden merkittävimmät piirteet.

CDK eli Chemistry Development Kit on vuonna 2003 BSD-lisenssillä julkaistu avoimeen lähdekoodiin perustuva ohjelmisto. CDK-ohjelmistolla voidaan käsitellä biologista ja kemiallista dataa erityisesti yhdisteiden rakenteen mallintamista varten. CDK-ohjelmisto käyttää Java-ohjelmointikieltä, jolla voidaan hakea, rakentaa ja esittää yhdisteiden rakennetta joko kaksiulotteisesti tai kolmiulotteisesti sekä mallintaa niiden ominaisuuksia eri tavoin. Ominaisuuksien kuvaamiseen ja ennustamiseen liittyy läheisesti rakenteiden esittäminen, hakeminen ja mallinnus perustuen molekulaaristen deskriptorien hyödyntämiseen. (Steinbeck ym., 2003.)

CDK-ohjelmistoa on kehitetty avoimen datan ja avoimen lähdekoodin periaatteilla akateemisessa Blue Obelisk -järjestössä, ja se on onnistunut ohjelmistojakelussa saavuttamaan suuren käyttäjäkunnan sekä laajan hyödynnettävyyden muissa keminformatiikan projekteissa uudistaen itseään arkkitehtuurin, ominaisuuksien, käyttökokemuksen, vertaisarvioinnin ja jakelun suhteen (Willighagen ym., 2017). CDK-ohjelmistoa yhdistetään useisiin muihin avoimen lähdekoodin ohjelmistoihin, joista yksi vaikuttavimmista on KNIME-CDK, joka yhdistää KNIME-ohjelmiston helppokäyttöisyyttä CDK-ohjelmiston keskeisten osien ympärille tutkimusprojektien tietoteknisissä asiankäsittelytehtävissä hyödynnettäväksi kokonaisuudeksi (Beisen ym., 2013).

RDKit on yksi käytetyimmistä keminformatiikan C++ -ohjelmistoista, jota voidaan hyödyntää joko C#, C++, Java tai Python ohjelmointikielillä. RDKit-ohjelmistolla on

kaupallisen käytön mahdollistavalla BSD-lisenssillä julkaistu avoimen lähdekoodin keminformatiikan ohjelmisto, joka julkaistiin vuonna 2006 ja jota ylläpidetään yhteisöllisellä ja avoimella ohjelmistokehityksellä (Landrum, 2019). RDKit sisältää erilaisia työkaluja, joilla voidaan rakentaa, muokata, analysoida ja arvioida molekyyliin liittyvää fysikaalista, biologista ja kemiallista tietoa (Landrum, 2021). RDKit on erityisen suosittu sen saavutettavuuden, ohjelmointikielen joustavuuden ja moninaisten ominaisuuksien puolesta. Useita erilaisia kemiallisia tietokantoja jaetaan Anaconda cloud -pilvipalvelussa RDKit-ohjelmistoa varten (Anaconda cloud, 2023). Erilaisia työkaluja, jotka hyödyntävät RDKit-ohjelmistoa tai sen osia on kehitetty lähivuosina aktiivisesti (Esim. Bento ym., 2020; Moriwaki ym., 2018; Riniker & Landrum, 2015; Tosco ym., 2014).

KNIME on avoimen lähdekoodin datatieteellinen työkaluun (engl. Workflow) tarkoitettu ohjelmisto, jota hyödynnetään monitieteellisesti eri tutkimusaloilla. KNIME mahdollistaa erityisesti helposti ymmärrettävän ja visuaalisen tavan mallintaa tuloksia erilaisesta datasta, ja erityisesti dataputken datatuotteiden luontiin erilaisin menetelmin, kuten koneoppimisen mallien avulla. (Knime, 2023). Dataputki (engl. Data pipeline) ja erilaiset datatieteelliset menetelmät mahdollistavat keminformatiikassa ja biologisissa tieteissä analyyseja suuresta määrästä dataa, mutta merkittävää on myös se, että tietoon voidaan luottaa. Erilaisia luotettavuuden standardeja dataputkimenetelmille ja niiden tuottamille malleille on ehdotettu. (Heil ym., 2021.) KNIME-ohjelmistoa on hyödynnetty laajalti niin monitieteellisissä kuin keminformatiikan konteksteissa, ja hyödynnetään usein CDK- ja RDKit-ohjelmistojen yhteydessä (Esim. Beisken ym., 2013; Dong ym., 2007; Mazanetz ym., 2012, McGuire ym., 2017). Puolet KNIME-ohjelmistoa käyttävistä hyödyntävät sitä biologisissa tieteissä missä keminformatiikka on usein keskeisessä roolissa (Warr, 2012).

Open Babel on kehitetty kemiallisen tiedon varastointia, tulkintaa, analyysia ja hallintaa varten eri tiedostotyypeillä ja ohjelmistokieliä. Ohjelmisto antaa tuen yli sadalle tiedostotyyppille ja sisältää työkaluja molekyylihallinnuksen, keminformatiikan ja bioinformatiikan tarkoituksiin. Sitä kutsutaankin tästä syystä ohjelmistojen työkalupakiksi, ja se on alun perin tuotettu kaupallisesta akateemikoille tarkoitettusta OEChem -ohjelmistosta (Wild, 2021 s. 106). Sitä voidaan myös hyödyntää C/C++ -työkaluna, jota voi hyödyntää muilla

ohjelmistokielillä kääreiden avulla (engl. Wrapper), mikä on mahdollista ohjelmointikielellä ja avoimella GPL-lisenssillä. Open Babelin toimintoihin kuuluu työkaluja konformeerien samankaltaisuuden ja alarakenteiden etsimiseen sekä rakennekaavojen luomiseen jostakin tuodusta tietokannasta. Tietokannoista voidaan tehdä liitännäisiä, jotka voidaan ladata ja hyödyntää tiedon käsittelyssä; käyttäjä voi seuloa dataa ja muuntaa tiedostotyyppejä muutamasta tiedostosta kokonaiseen kirjastoihin tai ajaa simulaatioita. (O'Boyle ym., 2011)

2.3.3 Ominaisuuksien ennustaminen

Yhdisteiden fysikaalis-kemiallisia ja biologisia ominaisuuksia voidaan mallintaa niiden kemiallisen rakenteen perusteella. Tämä nojaa empiiriseen käsitykseen, että avaruudellinen malli molekyylistä ja sen fysiologisista ominaisuuksista ovat läheisesti yhteydessä sen kokeellisiin kemiallisiin ja biologisiin ominaisuuksiin, joiden suhdetta voidaan mallintaa tilastomatematiikalla. (Hansch ym., 1962.)

Rakenne-aktiivisuussuhde SAR (engl. Structure-activity-relationship) on laskennallisen kemian malli, joka kuvaa yhdisteen kemiallisia, fysikaalisia ja biologisia ominaisuuksia kvalitatiivisesti. QSAR tarkoittaa kvantitatiivista rakenne-aktiivisuussuhdetta (engl. Quantitative Structure-activity-relationship), joka on matemaattinen malli kuvaamaan aineen kemiallisia ja fysikaalisia ominaisuuksia biologisissa systeemeissä. SAR- ja QSAR-mallin erona on eritellä, mallinnetaanko biologista aktiivisuutta vai biologista potenssia, ja usein jälkimmäinen perustuu enemmän matemaattisiin ennustuksiin, kun edellinen taas in vitro ja in vivo tulosten analyysiin (McKinney ym., 2000).

Keminformatiikassa käytetään usein QSAR-malleja, koska halutaan ennustaa yhdisteille ominaisuuksia, joita ei vielä ole syntetisoitu ja halutaan eritellä, onko tämä taloudellisesti, terveydellisesti ja kemiallisesti kannattavaa, turvallista tai mahdollista. QSAR-malleja testataan jatkuvasti, ja niiden on oltava tarpeeksi tarkkoja, jotta niitä voidaan hyödyntää, kun taas implisiittisillä tietoteknisillä esityksillä rakenne-aktiivisuussuhteella voidaan esittää mahdollisia ominaisuuksia, mutta kemistin rooli on merkittävämmässä osassa tulkita ovatko rakenteet käytännössä mahdollisia tai järkeviä kokeellisesta näkökulmasta. QSAR-mallit perustuvat molekulaarisiin deskriptoreihin. (Fourches ym., 2010; Tropsha, 2010.)

Tärkeitä sovelluksia (Q)SAR-mallinnukselle on etenkin toksisuuden ja lääkeominaisuuksien ennustamisessa, mutta nykyään uskotaan, että erityisesti QSAR-mallinnus voi olla hyödyllinen biokemian konteksteissa. Big Datan myötä yhdisteiden bioaktiivisuuden tieteellistä dataa on paljon ja vapaasti saatavilla erilaisille tutkijoille. (Muratov ym., 2020.) Tämän lisäksi esimerkiksi kemianteollisuuden säätelyssä hyödynnetään QSAR-mallinnusta terveysriskien hallinnassa. ECHA käyttää OECD:n periaatteiden mukaisia QSAR-malleja ympäristönsuojeluun osana EU:n kemikaalilainsäädännön täytäntöönpanoa, johon on myös laadittu ohjelmisto avoimella lähdekoodilla erilaisten kemianteollisuuden toimijoiden jakeluun. (Diderichs, 2010; ECHA, 2008.) Lääketeollisuudessa etenkin eläinkokeet ovat olleet historiallisesti keskeinen tapa testata niin biologisia kuin fysikaalis-kemiallisia ominaisuuksia ja toksisuutta, mutta QSAR-malleihin perustuvat in silico -menetelmät voivat tehdä joistain eläinkokeista tarpeettomia ja säästää resursseja. Eläinkokeiden ja esimerkiksi syöpäriskin arvioimista voidaan lähestyä QSAR-mallien vertailulla, mikä voidaan toteuttaa koneoppimisen menetelmiä hyödyntäen. (Pradeep ym., 2016.)

2.3.4 Koneoppimisen ja tekoälyn rooli keminformatiikassa

Koneoppiminen on tieteenhaara, jossa pyritään luomaan tietokoneille kyky oppia jotakin datasta ilman, että niitä on ohjelmoitu tarkasti oppimaan. Mitchell (1997, s. 3) määrittelee koneoppimisen seuraavanlaisesti: koneoppimisen ohjelmisto voi oppia, O tehtävää, T varten, jota mitataan suoriutumisen arvolla, S, jos T:n suorittaminen, joka mitataan arvolla S, paranee kun kokemus, K kasvaa. Koneoppiminen nähdään hyödylliseksi, kun halutaan selvittää joukosta mahdollisia hypoteeseja, matemaattisesti paras malli, joka vastaa myös jotakin havaittua dataa tai tietoa, joka tutkijalla jo on. Kyseessä on datan, mallintamisen ja oppimisen, sykli, joka tuottaa paremman oppimisen malleja. Tutkijan hyödyntämä tilastotieteellinen ja datatieteellinen menettely ja mallien arviointi on keskeisessä asemassa onnistuneen mallin saavuttamisessa. (Mitchell, 1997, s. 14.) Koneoppimisen algoritmit jaetaan kolmeen päätyyppiin: ohjattu oppiminen (engl. Supervised learning), ohjaamaton oppiminen (engl. Unsupervised learning) ja vahvistusoppiminen (engl. Reinforcement learning) (Banerjee & Das, 2021). Keminformatiikassa hyödynnetään lukuisia koneoppimisen algoritmeja, joilla voidaan ennustaa tai seuloa aineistosta haluttuja yhdisteitä jatkotutkimuksia varten. (Mitchell, 2014.)

Koneoppimista ja tekoälyä, joka perustuu koneoppimisen malleihin, hyödynnetään keminformatiikassa jo laajalti lääkekehityksessä, materiaalitieteissä ja ympäristötieteissä toksisuuden mallinnukseen (Esim. Luechtefeld ym., 2018; Moshawih ym., 2022; Xiouras ym., 2022). Koneoppimisen algoritmien soveltamista keminformatiikkaan kuvataan usein kaksivaiheiseksi. Ensimmäisessä vaiheessa muodostetaan molekulaarisia deskriptoreita yhdisteistä, joista ollaan kiinnostuneita muuttamalla niiden sisältämä tieto ominaisuusvektoreiksi. Tämän jälkeen mallinnus tai kartoitus (engl. mapping) seuloo koneoppimisen algoritmeilla haluttuja yhdisteitä aiemmasta pääjoukosta mallintaen niiden fysikaalisia, kemiallisia ja biologisia ominaisuuksia, jotka perustuvat ominaisuusvektoreihin, hakukriteereihin, algoritmeihin ja molekulaaristen deskriptorien tiedon eheyteen. Koneoppimisen algoritmeja on eri tyyppejä, joista yleisimpinä pidetään neuroverkkoja, syväoppimista, joukkoälyä, kemiallisten ominaisuuksia ennustavien tekoälyjen kokoonpanoja, satunnaismetsää, tukivektorikonetta ja lähimmän naapurin k luokitin. Myös naiivia Bayesin luokitinta, jota hyödynnetään luokittelemaan yhdisteitä kahteen tai useampaan luokkaan. Niiden luotettavuus ja käyttöarvo on erityisen korkea muitten koneoppimisen algoritmien rinnalla yhdisteiden monimutkaisten biologisten ominaisuuksien ennustamisessa. (Mitchell, 2014.)

Lääkeaineiden tutkimuksessa koneoppimisen algoritmit ovat laajalti käytettyjä, koska ne seulovat suuria määriä yhdisteitä tietokannoista, jotka sisältävät monia lääketieteellisesti lupaavia yhdisteitä. Tätä joukkoa, josta lääkeaineita tai muita yhdisteitä etsitään, kutsutaan kemialliseksi avaruudeksi (engl. Chemical Space). Se on keskeinen tutkimuskohde koneoppimisen algoritmien sovelluksille keminformatiikassa muitten tutkimusmenetelmien ohella, jotka voivat olla myös empiirisiä (Gromski ym., 2019). Esimerkiksi koneoppimisen malleja voidaan kehittää keminformatiikan avulla löytämään haluttuja yhdisteitä, suunnittelemaan parhaita synteesireittejä ja arvioimaan monimutkaisia ominaisuuksia, kuten lääkeaineen vastetta ja toksisuutta. Tällaisten mallien onnistumiseen tarvitaan laadukkaita ja tiedonhallinnaltaan standardoituja tietopankkeja ja datajoukkoja sekä biologisen vasteen koneoppimisen mallien hyödyntämistä. (Saifi ym., 2023.)

Koneoppimiseen perustuva tekoäly on laitteistojen ja ohjelmistojen kokoon, menetelmiin ja tutkimukseen perustuva tietotekninen älykkyyden malli, joka itsessään simuloi ihmisaivojen neuroverkkojen eli ihmisen aivojen ja ajattelun

ominaisuuksia ja toimintaa. Tämä johtaa siihen, että tekoälyn omaava tietokone vaikuttaa ihmiseltä ajattelumalleiltaan, mutta automatisoidusti ja tämä ajattelu perustuu tehtäviin ilman ohjeistusta eli tekoäly nojaa itseensä. Tämän lisäksi voidaan sanoa, että vuorovaikutus tekoälyn ja ihmisen välillä pyrkii usein tilanteeseen, jossa tekoäly pystyy simuloimaan ihmisperäisiä kokemuksia, ajatuksia ja tuntoaisteja. Edellä mainittua vuorovaikutusta pyritään luomaan erityisesti chattiboteilla. (Nagy, 2018) Nykyään kuitenkin tekoälymalleja käytetään laajalti rajoitetumpina versioina (engl. Narrow artificial intelligence), jonkin tietyn ongelman ratkaisuun, vaikka ihmisen ajattelua kokonaisemmin mallintavia tekoälyjä (engl. General artificial intelligence) on olemassa (Zohuri & Sadeh, 2020, s. 5–10).

Tekoälyn sovellukset keminformaattiseen lääkeaineiden seulontaan perustuu koneoppimiseen. Niitä sovelletaan erityisesti kemiallisen avaruuden seulonnassa sekä suurten empiiristen aineistojen tulosten (engl. High-throughput screening, HTS) ennustamisessa. HTS, jonka voi suomentaa automatisoiduksi seulontamenetelmäksi, on monimutkainen kokeellinen automatisoitu prosessi ja menetelmä, jossa tutkittavan kohteen ja erilaisten yhdisteiden kemiallista tai biokemiallista dataa voidaan seuloa empiirillä ja tilastotieteellisillä menetelmillä (Attene-Ramos ym., 2014). HTS:llä iteroidaan kohti joukkoa yhdisteitä, jotka täyttävät esimerkiksi lääketutkimukselle suotuisat ehdot eli joukosta yritetään löytää osumia tai haluttua dataa. HTS:llä voidaan tutkia niin biologista dataa, biokemiallista dataa kuin puhtaasti kemiallista dataa. Erilaisten HTS-tutkimusten tulosten tulkinnassa ja tulevien suunnittelun paremmassa ymmärtämisessä koneoppiminen on keskeisessä roolissa (Chen ym., 2018), ja koneoppimiseen perustuvia tekoälyjä on kehitetty samaan tehtävään (Dreiman ym., 2021).

2.4 Keminformatiikan opetuksen tutkimus

Tässä luvussa selostetaan viime vuosien kemian opetuksen tutkimustietoa keminformatiikan opetuksesta yliopistoissa. Lopuksi esitellään TPASK-malli kemian opettajankoulutuksesta, joka korostaa tieteellisen tiedon ja teknologisen tiedon yhteyttä muuttaa opiskelijoiden käsityksiä tieteestä ja teknologiasta.

2.4.1 Kemian opetuksen tutkimuksia keminformatiikasta yliopisto-opetuksessa

Kemian opetuksen tutkimuksesta keminformatiikan mahdollisuuksiin on tehty tutkimusta melko vähän, mutta kursseja kemian korkeakouluopiskelijoille on suunniteltu (esim. Kadtsyna ym., 2022). Tässä alaluvussa tarkastellaan julkaisuja keminformatiikasta, jotka käsittelevät myös kemian opetuksen tavoitteita.

Wild (2009) oli huolissaan Journal of Cheminformatics -lehdessä siitä, että keminformatiikka on ala, jossa tehdään paljon erilaisia asioita, mutta asiantuntijoiden muodostamat yhteisöt ja tutkimuksen fokus, näkyvyys ja luonteenpiirteiden puutteellinen keskustelu osana julkaisuja vaikuttaa merkittävästi alan rahoitukseen ja vetovoimaisuuteen. Kuitenkin rahoitusta on onnistuttu järjestämään esimerkiksi Euroopassa vuodesta 2013 alkaen on ollut keminformatiikan oma opinto-ohjelma kandidaatti- ja maisteritutkintona, jossa tutustutaan keminformatiikan ja bioinformatiikan kirjallisuuteen, ohjelmointiin ja tehdään keminformatiikkaan läheisesti pohjautuva tutkielma ja projekti (Jirát ym., 2013). Koulutusta perustellaan taidoilla, joita keminformaatikot työssään tarvitsevat, kuten Python-ohjelmointitaitoja, vahvan kemiallisen osaamisen, matemaattisen mallintamisen taidot erityisesti tilastomatematiikan soveltamisen muodossa tehtäviin joko ohjelmistokehityksessä tai tutkimuksessa (ACS, ei pvm.).

Korkeakouluopiskelijoiden keminformatiikan oppimiseen ja alan näkyvyyteen on kohdattu erityisesti verkkokursseilla. Kim ym. (2021) ovat kehittäneet verkkopohjaisen koulutusmateriaalin, jossa voi opiskella keminformatiikan perusteita ilmaiseksi jo vuodesta 2015 alkaen. Kurssilla painotetaan erityisesti tiedonhallintaa ja kykyä hyödyntää kemian aineistoja, kuten yhdisteiden tietokantoja osana kemistin työtä. Kurssilla avustettiin niin lähiopetuksen kuin etäopetuksen avulla oppimaan käytänteitä, joita asiantuntijat hyödyntävät, ja täten myös keminformaatikot suunnittelivat kurssia. (Kim ym., 2021.)

Kadtsyna ym. (2022) järjestivät keminformatiikan peruskurssin, joka kesti lukuvuoden. Kurssi järjestettiin etänä. Siinä tutustuttiin kemiallisiin tietokantoihin, rakenteiden kuvauksiin, reaktioiden ja kristallografian tietokantoihin, datatieteellisiin ja tiedonhallinnisiin menetelmiin. Näistä kirjoitettiin itsenäisesti lyhyt kirjoitus ja keskusteltiin kirjoitusten tuloksista. Tutkimuksessa selvitettiin kyselyllä, miten muitten kurssien opettajat ja kurssin suorittaneet opiskelijat hyötyivät

kurssista. Kaikista vastanneista enemmistö näki kurssin hyvin tärkeänä osaamistavoitteiltaan erityisesti analyyttisessä ja fysikaalisessa kemiassa. Opiskelijat kokivat hyötyä erityisesti kirjoittamisesta ja argumentoinnissa epäorgaanisen ja orgaanisen kemian opintojen aikana. Kurssin toteutus ja saatavuus nähtiin positiivisena ja resursseista erityisesti tietokannat nähtiin mielekkäinä, ja ainoana ongelmana nähtiin englanninkielinen oppimista tukeva kirjallisuus. Taidot nähtiin opinnoissa mielekkäinä, ja kurssin ainoana haasteena nähtiin sen järjestämisessä aiheen laajuus sekä opettajan tieto- ja taitovaatimukset. (Kadtsyna ym., 2022.)

Hardy ym. (2021) ovat kartoittaneet keminformatiikan mahdollisuuksia ja haasteita monitieteisestä, poikkitieteisestä ja tieteidenvälisestä näkökulmasta korkeakouluopetuksessa STEMM-viitekehyksessä (engl. Science, Technology, Engineering, Mathematics and Medicine). Keminformatiikan roolia korostetaan osana laskennallista kemiaa, matematiikkaa ja ohjelmointitaitoa, jossa otettaisiin huomioon oikea kohde ryhmä ja opiskelijan tutkinto-ohjelma (esim. farmaseuttinen kemia), jonka tarkoituksena on laajentaa keminformatiikan monitieteellistä hyödyntämistä eri aloilla (Hardy ym., 2021). Fowler ym. (2017) suunnittelivat Big Dataan perustuvan uuden kemian materiaalitieteiden koulutusohjelman jatko-opiskelijoille, jonka tarvetta perusteltiin keminformatiikan opetuksen puutteella, joka nähtiin kuitenkin keskeiseksi alalla.

Rodríguez-Becerra ym. (2020) ovat tutkineet ensimmäisinä opettajaopiskelijoille laskennallisen kemian oppimismoduulissa uusien laskennallisen kemian ja keminformatiikan työvälineiden käyttömahdollisuuksia TPASK-mallin avulla. Tutkimuksessa saatiin selville, että opiskelijat kokivat haasteelliseksi hyödyntää tieto- ja viestintäteknologian (TVT) työvälineitä, jotka olivat heille uusia tai kun heiltä puuttui jokin taito hyödyntää sitä. Tästä huolimatta suurin osa näki tieteellisen tutkimuksen ohjelmistojen auttaneen heitä ymmärtämään paremmin kuin opettajalähtöiset opetusmenetelmät. Autonomia ja ohjelmistojen sovittaminen reaali maailman ongelmiin nähtiin hyödyllisinä. Moduuli nähtiin keinona muuttaa pedagogisia käytäntöjä, koska se saa opettajaopiskelijat miettimään laskennallisen kemian ohjelmistojen kehitystä ja rooleja tulevaisuudessa. (Rodríguez-Becerra ym., 2020.)

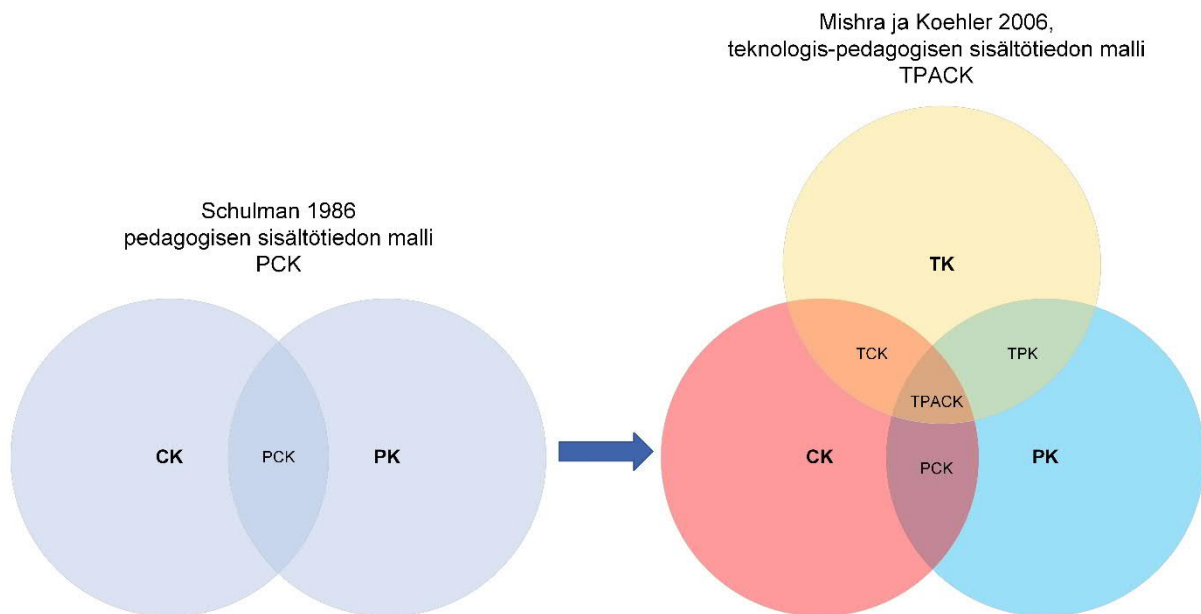
2.4.2 Keminformatiikan ja TPASK-mallin yhteys

Tässä tutkielmassa hyödynnetään Jimoyianniksen (2010) TPASK-mallia, joka perustuu TPACK-malliin (Mishra & Koehler, 2006). Keminformatiikka voidaan lukea osaksi TPASK-mallia, koska se on teknologiaa, jonka arviointiin opetuksessa TPASK-malli soveltuu (esim. Rodríguez-Becerra ym., 2020). Molempien mallien perustana on Shulmanin (1986; 1987) kirjoitukset pedagogisen sisältötiedonmallista laadukkaaseen opetukseen.

TPACK on malli teoreettisesta viitekehyksestä, joka tukee opetusta kuvailemalla opettajan tietämystä oppiaineen sisältötiedon, pedagogiikan ja teknologian muodostamasta kokonaisuudesta opetuksessa. TPACK koostuu kolmesta osa-alueesta: sisältötiedosta (CK, Content Knowledge), pedagogisesta tiedosta (PK, Pedagogical Knowledge) ja teknologisesta tiedosta (TK, Technological Knowledge), joiden leikkaus muodostaa TPACK-mallin kokonaisuuden eli Teknologis-pedagogis-ainesisällöllinen osaamisen, jota Mishra ja Koehler (2006) kuvailevat opettajan kyvyksi järjestää ”merkityksellistä ja taitavaa opetusta teknologian avulla, joka muodostaa erillisen kokonaisuuden muihin mallin osaamiskokonaisuuksiin verrattuna”. Kyse on siis erilaisen tiedon ulottuvuuksien yhdistämisestä osaamiseksi, joka toteutuu laadukkaana, jäsennehtynä ja pidettynä opetuksena. Pedagogisen tiedon, sisältötiedon ja teknologisen tiedon PK, CK ja TK osa-alueet tai joukot muodostavat leikkauksia eli yhteisiä osaamisalueita, jotka kuvaavat tarkemmin osa-alueiden välistä vuorovaikutusta. Nämä muut leikkaukset ovat: pedagoginen sisältötieto, PCK, teknologis-pedagoginen tieto, TPK ja teknologinen sisältötieto, TCK. joilla tarkoitetaan joukkojen leikkausten synteysiä, eli uutta joukkoa. Edellä mainittujen leikkausten leikkaus eli teknologis-pedagogisen sisältötiedon TPACK osaamisalue on mallin pääväite tehokkaasta, uudistumiskykyisestä ja merkityksellisestä opetuksen kehittämisestä teknologialla. TPACK on kokonaisuus, jonka keskeisin tavoite on arvioida opettajan tarvitsemaa tietokokonaisuutta, menetelmiä ja tavoitteita. (Mishra & Koehler, 2006.)

TPACK-malli on johdettu Shulmanin (1986; 1987) pedagogisen sisältötiedon mallista, joka kuvaa opettajan kykyä toteuttaa laadukasta ja tehokasta opetusta perustuen kahden ulottuvuuden tai tietokokonaisuuden synteysiin: ymmärrykseen pedagogiikasta ja opetettavista sisällöistä uutena käsitteenä, pedagogisena

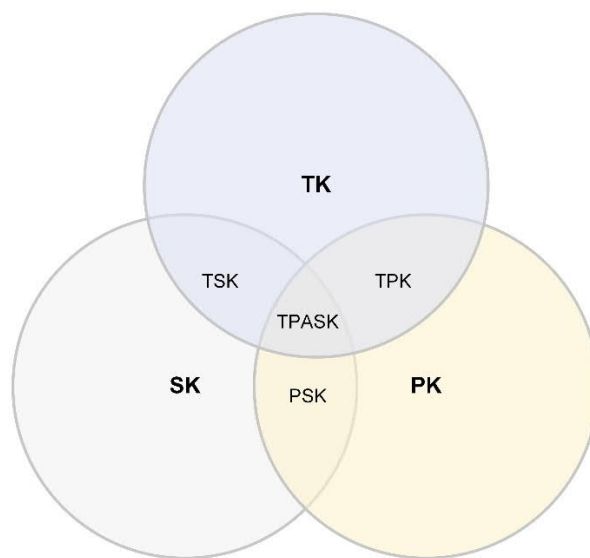
sisältötietona. Shulman (1986) oli ensimmäinen, joka jakoi opettajien tietämyksen pedagogiseen tietoon (PK) ja oppiaineen sisältötietoon (CK) ja näiden synteysiin, pedagogiseen sisältötietoon (PCK, Pedagogical Content Knowledge). PCK:lla voidaan kuvailla opettajan hyödyntämää tietoa opetuksen suunnittelussa, toteutuksessa ja arvioinnissa, mikä heijastuu kyvyssä opettaa. Mishra & Koehler (2006) kehittivät Shulmanin (1986) mallia ottaen huomioon ajankohtaisen opetuksellisen ulottuvuuden: teknologian opetusvälineenä ja teknologiaan liittyvän osaamisen. Opettajat hyödyntävät opetuksessa ja erityisesti kemian opetuksessa teknologisia sovelluksia, kuten mittausautomaatiikkaa, molekyylimalleja ja simulaatioita; esimerkiksi virtuaaliset laboratoriot ovat nykyaikana erityisenä tutkimuskohteena etäopetuksessa (Chan ym., 2021). Teknologian ominaisuudet muodostavat yhteyden pedagogiikkaan ja sisältötietoon, mitä leikkaukset TPK (Technological Pedagogical Knowledge) ja TCK (Technological Content Knowledge) havainnollistavat kuvassa 2.



Kuva 2. Shulmannin (1986) malli pedagogisesta sisältötiedosta on vaikuttanut keskeisesti siihen, miten Mishra ja Koehler (2006) muodostivat teknologian vaikutuksen pedagogiseen sisältötietoon teknologis-pedagoginen mallin muodossa.

TPASK on Jimoyiannisin (2010) teorettinen malli opetuksesta, teknologiasta ja opettajan taidoista, joilla pyritään kehittämään TPACK-malliin perustuen opettajankoulutusta erityisellä opintosuunnitelmalla kohti autenttista eli tehokkaampaa ja kontekstuaalisesti hyödyllistä oppimista ja parempaan teknologian integraatiota opetukseen. TPASK-malli sisältö on jaettu samantapaisesti kuin

TPACK-mallissa, mutta sisältötieto on korvattu tieteellisellä tiedolla (SK, Science Knowledge), koska TPASK-mallin avulla opetuksessa sisällöllisesti pitää käsitellä tietoa, jota asiantuntijat kuten kemistit, fyysikot ja biologit käyttävät. TPASK-mallilla ei kutienkaan pyritä asiantuntijoiden tietorakenteisiin, vaan pyrkii selittämään, millainen tieto on tarpeen opettajille, jotka opettavat luonnontieteitä. Sama pätee tämän tiedon hyödyntämiseen. Tällä pyritään paremmin integroimaan tieto- ja viestintäteknologiaa (TVT) opetukseen sekä opettajien koulutukseen ja tulevaan ammatilliseen osaamiseen. (Jimoyiannis, 2010.) Kuvassa 3 on visualisoitu TPASK-malli.



Kuva 3. Jimoyiannisin (2010) esitys TPASK-mallista, jossa 'Science' vastaa SK:ta, 'Pedagogy' PK:ta ja 'Technology' TK:ta. Edellä mainittujen tiedon luokkien leikkaukset luovat osaamisalueita, jotka ovat enemmän kuin aiemmat luokat erikseen. Nämä leikkaukset vuorostaan muodostavat kokonaisuuden, TPASK:n.

Jimoyianniksen (2010) TPASK-mallin osaamisalueet voidaan esittää seuraavanlaisesti:

1. Teknologinen tieto (TK) kuvaa kykyä ymmärtää ja hyödyntää teknologiaa opetuksessa.
2. Pedagoginen tieto (PK) on yleistä tietoa oppimisesta ja opetuksesta, joka nojaa johonkin oppimisen filosofiaan.
3. Tieteellinen tieto (SK) kuvaa käsitekohtaista tietoa, kuten alakäsitteitä ja relaatioita orbitaaleista.

4. Teknologis-tieteellinen tieto (TSK) hyödynnetään opetuksessa ymmärtämään, mikä teknologian rooli on luonnontieteellisen tiedon tuottamisessa tieteellisellä menetelmällä.
5. Pedagoginen tieteellinen tieto (PSK) on tietoa siitä miten tieteellistä ja pedagogisesti vaikuttavaa ja tehokasta tiedeopetusta ja oppimista voi järjestää.
6. Teknologis-pedagoginen tieto (TPK) on tietoa teknologian hyödyntämisen tuomista ongelmista ja hyödyistä opetukseen.
7. Teknologis-pedagoginen tieteellinen tieto (TPASK) on opettajan osaamista vaikuttavan opetuksen järjestämiseksi edellä mainittujen osaamisalueitten synteessin muodossa.

Kriteerit ja esimerkit kuvailevista luokista TSK:lle, TPK:lle ja TPASK:lle voidaan esittää seuraavanlaisesti TPASK-mallille taulukossa 3. TPK nähtiin yleisenä teknologian soveltuvuutena ilman kriteerejä, miten sitä hyödynnetään aiheen suhteen, vaan se sisältää tietämyksen siitä, mikä on mahdollista teknologialla saavuttaa erilaisissa pedagogisissa strategioissa. (Jimoyiannis, 2010.)

Taulukko 3. Jimoyianniksen (2010) TPASK-mallin kriteerit osaamisalueille TSK, TPK ja TPASK.

TPASK-mallin osaamisalue	Kriteerit
TSK	• Resurssit ja työvälineet • Tieteelliseen tiedon hyödyntämisen käytännön tekniset taidot • Tieteellisen osaamisen muutos • Tieteellisten prosessien muutos
TPK	• Tieteellinen osaamisen • Tieteellisen osaamisen muuttuminen • Opetussuunnitelma • Oppimisen haasteet • Oppimisstrategiat • Pedagogia • Opetuksen konteksti
TPASK	• TVT-työkalujen käyttömahdollisuudet

	• TVT:n tukemat oppimisstrategiat • Tutkimuksellisuuden tukeminen TVT:llä • Opiskelijoiden tukeminen • Opiskelijoiden tekniset haasteet
--	---

3 Tutkimus

Tässä luvussa esitetään tutkimusta ohjaavat tutkimuskysymykset, perustellaan menetelmällisyys suhteessa tutkittavaan ilmiöön ja todetaan tutkimusote.

3.1 Tutkimuskysymykset

Tutkimuskysymyksillä pyrittiin tutkittavan ilmiön eli keminformatiikan soveltamisen, oppimisen ja luonteen piirteiden ymmärtämiseen. Seuraavat tutkimuskysymykset ohjasivat tutkimusta:

1. Miten kemian asiantuntijat hyödyntävät keminformatiikkaa tutkimuksessa?
2. Millaisia taitoja kemian asiantuntijat kokevat tarvitsevänsä keminformatiikan hyödyntämisessä?
3. Miten keminformatiikka voi edistää kemian opetusta ja oppimista TPASK-mallilla analysoituna?

3.2 Tutkimusote

Tutkimusote on laadullinen haastattelututkimus, joka toteutetaan keminformatiikan asiantuntijoiden (N=3) avoimella haastattelulla. Haastateltavia pyydettiin kertomaan ja kuvailemaan miten ja millä tavalla keminformatiikkaa hyödynnetään, opitaan ja millainen luonne sillä on tieteellisessä ajattelussa avoimella haastattelulla.

Avoimessa haastattelussa haastatteli hyödyntää avoimia kysymyksiä, joiden tehtävänä on syventää haastateltavien vastauksia sosiaalisessa vuorovaikutuksessa. Tätä menetelmää voidaan kuvata avoimeksi, eli strukturoimattomaksi, ja tunnetaan myös syvähaastatteluna. Luonteeltaan avoin haastattelu muistuttaa keskustelua, mutta keskustelussa henkilöillä on selkeät roolit – haastatteli ja haastateltava, ja haastattelu on luonteeltaan syvällistä ja perinpohjaista. Haastateltavat valittiin haetun tiedon hallintaan erikoistuneista henkilöistä eli informanteista. Haastattelun menetelmällä pyritään tietoon, jota kerätään mahdollisimman kielellisesti luontevassa vuorovaikutuksessa, kuten keskustelutilanteessa, jossa haastatteli voi käyttää erilaisia lähestymistapoja tutkittavan aineiston keruuseen. (Hirsjärvi & Hurme 2008, s. 45–46.)

Sisällönanalyysi empiirisestä aineistosta perustuu teoriaohjaavaan sisällönanalyysiin, jossa luokittelu etenee aineiston ehdolla, mutta liitetään teoreettisiin käsitteisiin.

(Tuomi & Sarajärvi, 2018, s. 98.) Teoriaohjaava on kaksiosainen induktiivista päättelyä hyödyntävä prosessi, jossa teoria ohjaa lopputulokseen eli myös deduktiivisuus on läsnä. Aineistosta luokitellaan pääluokat, yläluokat ja alaluokat hyödyntäen TPASK-mallin luokituksia yläluokkina, joiden merkitystä tuodaan esiin nousseiden teemojen avulla. (Tuomi & Sarajärvi, 2018, s. 85, 115, 133.)

Tutkittava ilmiö keminformatiikan luonteen, soveltamisen sekä opetuksen ja oppimisen suhteen rajattiin haastattelututkimusten tulosten myötä korkeakouluopetukseen, mitä perustellaan kirjallisuudella keminformatiikan osaajien työnkuvan ja opinto-ohjelmien painotusten suhteen (Esim. Hardy ym., 2021; Jirát ym., 2013). Aineistonkeruu ja analyysi eteni seuraavanlaisesti:

- Haastattelut suoritettiin, äänitettiin ja litteroitiin hyödyntäen tietoteknisiä apuvälineitä; yhden haastattelun nauhoitti haastateltava omasta aloitteestaan
- Aineisto luokiteltiin ja teemoiteltiin sisällönanalyysissa aluksi induktiivisesti.
- Tulkintoja ja päätelmiä tehtiin lopulta teoriaohjaavasti perustuen deduktiiviseen päättelyyn TPASK-mallilla, sillä tutkimus pyrkii luonnehtimaan parasta mahdollista kuvausta haastateltavien ajatuksista suhteessa aiempaan tutkimukseen ja teoriaan teoriaohjaavassa sisällönanalyysissa.
- TPASK-mallin luokituksia (SK, TK ja TSK) tehtiin perustuen Jimoyianniksen (2010) erittelyyn, joka löytyy luvusta 2.4.2.
- PK:n, PSK:n, TPK:n ja TPASK:n kriteerien ei nähty täyttyneen, koska kaikilla haastateltavilla ei ollut tietoa keminformatiikan pedagogiikasta oppimisen tai opettamisen suhteen. Tätä perustellaan sillä, että TPASK-mallin pedagogisen arvon ja osaamisen luokittelu vaatii aina niin oppilaiden kuin opettajien työn arviointia ja mittareita jossakin opetus- ja oppimistilanteessa.

3.3 Tutkimuskohde

Tutkimus toteutettiin haastattelemalla kolmea asiantuntijaa, joihin otettiin yhteyttä sähköpostitse. Yhteydenotto toteutettiin sähköpostitse ja ennen haastattelua lupaa

kysyttiin uudelleen. Haastattelut toteutettiin Zoom-alustalla ja ne nauhoitettiin ja litteroitiin helmi-maaliskuussa 2022 ja analysoitiin laadullisella sisällönanalyysilla.

Asiantuntijoita keminformatiikasta etsittiin tieteellisten julkaisujen ja raportoitujen keminformatiikan asiantuntijataitojen pohjalta hakukoneista, kuten Google Scholarista ja SciFinder -tietokannasta sekä työelämän sosiaalisen median kanavalta, LinkedIn -sivustolta. Haastateltavat valikoitiin kemian osaamisen perusteella myös taustaltaan seuraavilla kriteereillä: kemian alan korkeakoulututkinnolla, keminformatiikan julkaisuilla ja raportoidulla osaamisella tai työkokemuksella teollisuudessa.

Haastateltavat antoivat luvan haastatteluun ja ymmärsivät pseudoanonymisointi prosessin. Pseudonymisointi oli välttämätöntä, koska tutkimuksen luottavuudelle on olennaista kertoa heidän akateemisia eroavaisuuksia ja yhtäläisyyksiä. Pseudonymisoinnissa kunnioitetaan Euroopan tietosuojalakia ja asetusta tutkittavien yksityisyyttä. Kriteerit eettiselle tutkimukselle perustuvat haastateltavien tunnistamattomuuden takaaviin menetelmiin lainmukaisesti, asianmukaisesti ja rekisteröidyn eli haastateltavan kannalta läpinäkyvästi. (Tietosuoja-asetus, 679/2016; Tietosuojalaki, 1050/2018.)

Haastateltava 1 on suomalaisessa yliopistossa työskennellyt laskennallisen kemian väitöskirjatutkija. Hän on tehnyt laskennallisen kemian lisäksi keminformatiikan tutkimusta ja raportoi tämän osaamisalueekseen. Hänelle lähetettiin kutsu haastatteluun työelämän sosiaalisen median kautta.

Haastateltava 2 on yliopistossa työskentelevä keminformatiikan apulaisprofessori, joka työskentelee eurooppalaisessa tutkimusyksikössä lääkeyhdisteiden tutkimuksen ja suunnittelun parissa. Haastateltava 2 on väitellyt keminformatiikasta ja julkaissut lukuisia keminformatiikan tutkimusartikkeleita sekä työskennellyt yksityisellä sektorilla keminformatiikan parissa erityisesti teollisuudessa. Haastateltavan 2 toimenkuvaan kuuluu myös opetusta. Hänet kutsuttiin haastatteluun lukuisten julkaisujen myötä, joita löydettiin Google Scholar -hakukoneen avulla.

Haastateltava 3 puolestaan on väitellyt keminformatiikasta eurooppalaisesta yliopistosta ja julkaissut useita keminformatiikan ja muitten alojen datatieteellisiä artikkeleja erityisesti koneoppimiseen ja tekoälyyn liittyen erilaisissa tutkimusyksiköissä.

3.4 Tutkimuksen luotettavuus

Luotettavuus ilmaisee, onko tutkimuksen tutkimusmenetelmä valittu mittaamaan tarkoitettua ilmiötä (Tilastokeskus, ei pvm.). Tutkimuksen luotettavuuteen vaikuttaa tarkasti raportoitu aineistonkeruu, eli haastattelun suoritus vaiheineen ja raportointi sekä sisällönanalyysin kattavuus, josta jälkimmäiseen pyrittiin raportoimalla tehdyt eri tasojen luokitukset ja teemat tuloksissa sekä liitteessä 1. Tälle menetelmällisyydelle pääasiallisena kritiikkinä on aineiston laadullinen luonne, mutta toisaalta ennen muita tutkimuksia tai samanlaisen mallin hyödyntämistä validiteettia ei voida suoraan kumota tai kyseenalaistaa valitulla menetelmällä. Toisaalta monimenetelmällisellä analyysillä voidaan trianguloida merkityksellisiä havaintoja. Tähän puutteeseen vastattiin tutkimuksessa kvantifioimalla (Tuomi & Sarajärvi, 2018, s. 122) luvussa 4.4. ja liitteessä 1 luokittelut, joka avaa mahdollistaa havaintojen vertailun, erityisesti jatkotutkimuksia varten. Tutkimusaineisto koostui erilaisilta informanteilta tarkoituksella, jotta tiedon kohteen triangulaation myötä tutkimuksen luotettavuus otettaisiin paremmin huomioon (Tuomi & Sarajärvi, 2018, s. 168). Tällaisella menetelmällisyydellä voidaan saada parempi luotettavuus, jos havainnot ovat samanlaisia tai mahdollisesti eritellä eroavaisuuksien syitä. Myös teoriaohjaavan sisällönanalyysin päättely köyhän ja rikkaan aineiston suhteen voidaan kyseenalaistaa valitsemalla toinen tutkimusmenetelmä, kuten aineistolähtöinen sisällönanalyysi (Tuomi & Sarajärvi, 2018, s.122–123.)

Toisaalta haastattelututkimuksen haastateltavien eli informanttien joukko mahdollisesti olisi voinut olla erilainen. Vaikka moninaisuuteen pyrittiin, niin esimerkiksi kemistit, jotka hyödyntävät keskeisesti keminformatiikan menetelmiä ja ohjelmistoja osana kokeellisuutta, mutta eivät kehittä uusia menetelmiä voivat toimia vertailukelpoisena informanttien joukkona. Avoimen haastattelututkimuksen heikkous on olennaisesti toistettavuus, koska sellaiset ovat aina ainutlaatuisia (Hirsjärvi & Hurme, 2018, s. 35).

Aineistonkeruun ja haastateltavien valinnan tavoitteena oli saavuttaa mahdollisimman monipuolinen joukko asiantuntijoita, joiden laajan kokemuksen avulla keminformatiikkaa voitiin kuvata kattavasti monitieteellisenä alana niin tutkimuksen kuin teollisuuden osalta. Olennaista oli myös haastateltavien työuran pituuden ja osaamisen erilaisten painoalojen valikoiminen. Tämä vaikutti siihen,

miksi esimerkiksi haastateltavat 1 ja 2 edustavat tutkimisen erilaisia näkökulmia kokemuksen ja menetelmien suhteen, kun haastateltava 3 erottui joukosta erilaisten datatieteellisten julkaisujen puolesta.

4 Tulokset

Tutkimuksessa vastattiin kolmeen tutkimuskysymykseen, joista käytetään lyhenteitä TK 1, TK 2 ja TK 3. Alaluokat luokiteltiin induktiivisesti ja yläluokat deduktiivisesti. Pääluokat selittävät aiempia luokituksia, ja aineistosta esiin nousseet teemat esitellään yhteenvedossa (Kts. Taulukko 4).

4.1 Keminformatiikan hyödyntäminen (TK 1)

Keminformatiikka nähtiin lähestymistapana tutkimuksellisuuteen, jossa tietokoneita ja datatieteen menetelmällisyyttä kemiallisesta informaatiosta hyödynnettiin jatkeena ihmisen luovalle ongelmanratkaisulle, hypoteesien testaamiselle ja mallintamiselle.

Haastateltavilta kysyttiin mihin he hyödyntävät keminformatiikkaa ja millaisena he näkevät alan. Vastauksiin vaikutti vahvasti asiantuntijoiden aiempi työkokemus ja tieteenhaarat, joihin he ovat erikoistuneet.

Haastateltavat kuvailivat keminformatiikkaa eri tavoin:

Haastateltava 1: Algoritminen tapa käsitellä molekyylejä ja niiden ominaisuuksia. Tietojenkäsittelytieteen ja kemian leikkaus.

Haastateltava 2: keminformatiikka on käytännössä kemiallisen datan hyödyntämistä.

Haastateltava 3: Kemiallisten prosessien simulointia, kuin ongelmanratkaisua 'paperilla ja kynällä' laboratorion sijaan.

Haastateltava 1 korosti laskennallisen kemian vaikutusta ymmärrykseensä keminformatiikasta ja näki keminformatiikassa mahdollisuuden ymmärtää kemiaa laaja-alaisemmin kuin keskittymällä yhteen tapaukseen, kuten yhteen molekyyliin:

Haastateltava 1: ... kun oma tausta on laskennallisesti kemiassa, missä alkuun aloitin käsittelemällä molekyylejä manuaalisesti yksitellen, että rakensin sen jollain molekyylieditorilla. Sitten laskin molekyylin ominaisuuden ... kemoinformatiikka ensinnäkin antoi minulle keinot manipuloida näitä molekyylejä ohjelmallisesti ja käsitellä tuloksia mitä saan laskuista. Ajan ne laskut, näen sen jossain määrin kanssa kemoinformatiikkana.

Tämän lisäksi hän näki, että kyse on tutkimuksen mallintamisen optimoinnista tai parhaan mahdollisen ratkaisun löytämisestä yhdessä laaja-alaisemman käsityksen rinnalla.

Haastateltava 1: ... minun mielestäni tämä oli tämmöinen optimointiongelman kaltainen juttu, että meillä on joku tavoite mitä halutaan saada. Tämä molekyyli on tietynlainen. [Halutaan] tietty ominaisuus ja se ominaisuus on sen rakenteen funktio. Eli mikä mitä atomeja tässä on ... se tuntui mielekkäämmältä sen sijaan, että käytän ihmisen mielikuvitusta muotoilla tämän jonkinlaisena niin kuin optimointiongelmana.

Haastateltava 2 abstrahoi keminformatiikan osaksi suurempaa kokoelmaa informaatiota kemiasta ja biologiasta, ja korosti tämän informaation tai datan avoimuuden merkityksellisyyttä.

Haastateltava 2: Keminformatiikasta tuli paljon suurempi, koska se on datalähtöistä ... enemmän dataa, enemmän algoritmeja, mutta keminformatiikassa kaikki kaikessa on data ... ja datan ymmärtämiseksi tarvitsemme avointa dataa ainakin jollakin asteella ... keminformatiikan ja bioinformatiikan suhteen, bioinformatiikassa kaikki oli avointa, mutta asiat olivat toisin, kun tein väitöskirjani keminformatiikasta.

Hän nostaa esimerkin farmasiasta.

Haastateltava 2: yritys ei voi julkistaa kaikkea dataansa, mutta osan siitä ja tulokset, mutta bioinformatiikassa kaikki on avointa ja saatavilla ... päinvastoin kuin keminformatiikassa 15 vuotta sitten. Lääketutkimus tarkoittaa rahaa ja suljettuja systeemejä.

Haastateltava 2 näkee kuitenkin ongelman muuttuneen paremmaksi säätelyn myötä:

Haastateltava 2: Kova paine säätelyelimiltä on kohdistanut työskentelyn keminformatiikassa malleihin, joiden pitää olla hyväksyttyjä erityisesti toksisuuden suhteen ... tarvitset tunnustusta ja hyväksyntää malleille ja menetelmillesi, mikä saavutetaan julkaisuilla. Tämä säätely ja avoimuus toimii ajurina alalla.

Tätä tuki osittain haastateltavan 3 käsitykset datan kattavuudesta, jota selittää automatisoitujen laboratorioden (esim. HTS-menetelmä) käyttöönotto.

Haastateltava 3: Dataa on kallista saada, ja vaikka nykyään on automatisoituja laboratorioita ja mittauksia [High-throughput screening], kun taas aiemmin tämä oli hyvin kallista ja dataa ei ollut, niin nyt luulen, että saamme paljon dataa [jaettavaksi] ja että sitä on yhtä paljon kuin kieliprosessoinnissa tai genetiikassa.

Tästä huolimatta hän nosti huolensa datan riittävydestä, avoimuudesta ja kustannuksista.

Haastateltava 3: Sanoisin että dataa ei ole tarpeeksi saatavilla, vaikka tietopankeista tulee suurempia ja monipuolisempia, se ei riitä. Tämä on yksi syy, miksi pyritään luomaan [enumerointi] molekyylejä kemiallisesta avaruudesta.

Mutta toisaalta haastateltava 3 lisää, että tämän tiedon hyödyntäminen tutkijoille voi olla puutteellista, koska se ei ole aina avointa keminformatiikan alalla. Menetelmät ovat niin haastavia, että vain valmiit menetelmät ja ohjelmistot ovat käytettävissä. Alan kehityksestä voi jäädä jälkeen.

Haastateltava 3: Luulen että kehitys on liian nopeaa tavanomaisiin askeliin [ajattelussa] ... nykyään yritetään ottaa kiinni kaikkea mitä tehdään tietokannoissa ja uusissa julkaisuissa ... luulen että siihen menisi koko päivä ymmärtää, mitä uusimpia menetelmiä [nyt] käytetään.

Datan analysoimisesta tulee olemassa olevien aineistojen hyödyntämistä, mikä johtaa omien menetelmien hyödyntämiseen. Haastateltava 3 luo kontrastin lääketieteellisuuden ylläpitämiin tietokantoihin, jotka ovat yksityistä omaisuutta ja avoimen lähdekoodin välille.

Haastateltava 3: Jos ei työskentele lääketieteellisyydessä, vaan olet toisenlainen tutkija, käytössäsi oleva data on enemmän tai vähemmän vakio ... täten kehitetään menetelmiä [datan sijaan].

Haastateltava 3 kuitenkin toteaa keminformatiikan merkityksen tärkeäksi kustannustehokkuuden myötä:

Haastateltava 3: Nykyajan keminformaatikko yrittää ymmärtää datasta, joka on olemassa, kaiken mitä voi vähentääkseen kokeellisuudesta seuraavia kustannuksia tulevaisuudessa.

4.2 Keminformatiikan vaikutus tutkimukseen (TK 1)

Keminformatiikan yhteys laskennalliseen kemiaan, tiedonhallintaan, koneoppimisen malleihin ja datatieteelliseen päättelyyn näyttäytyi selkeästi haastateltavien vastauksista. Tutkimuskonteksteissa korostui erityisesti datan määrä ja laatu tai tarve sekä aiempien mallien testaaminen, kehittäminen tai laajentaminen yleistettävään suuntaan. Keminformatiikan rooli nähtiin erityisesti kokeellisia tuloksia täydentävänä.

4.2.1 Keminformatiikka ja ongelmanratkaisu

Haastateltava 1 näki keminformatiikan muuttavan hänen tapaansa ratkaista ongelmia ja nähdä kemia uudesta näkökulmasta, ikään kuin vapauttavana ja ajattelua edistävänä. Hänen mielestään pelkästään ihmisen luova toiminta ei yksin riitä.

Haastateltava 1: [Keminformatiikka] ottaa ongelman ja pyrkii viemään sitä yleisempään suuntaan. Siitä tulee sitten optimointiongelma, ja mitä olen paljon tehnyt, on koneoppimiseen liittyvät ongelmat ... koneoppimisbuumi alkoi seitsemisen vuotta sitten ja innostuin siitä, koska näin, että molekyylin suunnittelussa ihmisen mielikuvitus on rajoittava tekijä. Tämä muutti hyvin nopeasti minun tapaani nähdä kemian ongelmia.

Mutta hän toteaa kuitenkin keminformatiikan mahdollistavan erilaista ongelmanratkaisua.

Haastateltava 1: Siinä työskentelee metatasolla eikä manuaalisesti rakenna molekyylejä., vaan siinä yhdistyy luova prosessi kysymysten asetteluun ja tilanne elää, kun se ei ole teknistä puurtamista ... se on hyvinkin yleistävä ja luova tapa.

Kaikki haastateltavat korostivat sitä, että keminformatiikan hyödyntämisessä pitää ottaa huomion kemian kokeellisuuden perusta mallintamiselle. Silti haastateltavien mielestä keminformatiikalla vastataan täydentävästi kokeellisiin tuloksiin tai muodostamalla valideja hypoteeseja aiempiin kokeellisiin tuloksiin perustuen.

Haastateltava 1: Kokeellisuus on aina se [perusta] mitä tavoitellaan.

Haastateltava 2: [keminformatiikan in silico menetelmistä] Ne eivät ikinä yksin ole tarpeeksi. Teemme yhteistyötä kokeellisten kemistien kanssa, jotka validoivat mallin ennen sen lähempää tarkastelua. Tämä on tärkeää ennen tuotantoa [lääketeollisuudessa]. Keminformatiikalla saamme tietää miten hyvin mallit tulevat toimimaan, saadaan mustavalkoisia sävyjä hyvin testatuilla menetelmillä, jotka perustuvat tilastotieteeseen. Ongelmana on vain datan puute.

Haastateltava 3: Tarvitaan malleja, jotka ovat yleistettäviä ja sovellettavissa eri ongelmiin, ja pelkistyy kokeellisuuteen, mitä pitää tehdä [tiedon varmistamiseksi].

Haastateltava 2 jatkaa selittämällä, kuinka keminformatiikalla saadaan ennen kokeellisuutta oleellista tietoa, mikä on kiinnostavaa tutkijoille, vaikka mallin ennuste ja aiempi data olisikin puutteellisia; mallinnuksilla on mahdollisuus luoda suuntaa antavia hypoteeseja kokeellisuudelle.

Haastateltava 2: [Keminformatiikalla] voidaan vastata selkeästi kysymyksiin, jotka vastaavat jo tiedettyä tietoa molekyylistä. Jos tutkitaan molekyylin kemiallista avaruutta malleilla, saadaan joko validia tai suuntaa antavaa tietoa yhdisteen ominaisuuksista. Joka tapauksessa saamme johtopäätöksen siitä, onko malli totta vai tarvitseeko kokeellisuutta tehdä.

Haastateltavat antoivat esimerkkejä, joissa keminformatiikkaa käytettiin ongelmanratkaisuun. Haastateltava 1 kertoi, kuinka automatisoi ongelmanratkaisuaan ja hyödyntää erityisesti kemiallisen avaruuden tutkimista.

Haastateltava 1: Olen ennen ... katsonut yksittäisiä molekyyliä. Nyt olen kiinnostunut molekyylien enumeroinnista eli olen ottanut miljoonia molekyyliä [kemiallinen avaruus] ja pystyn laskemaan ja kehittämään näiden kaikkien molekyyliarakenteet. Pystyn automatisoimaan osan, mikä oli aiemmin hidastamassa ... mulla on molekyyli, jossa on laskettu virrantiheys, jotta näen sidoksesta yli menevän virran voimakkuuden. Pystyn tekemään rakenteellisia hypoteeseja, jotka huomaan datasta ... voin käydä kaikki potentiaaliset yhdisteet läpi ja laskea niiden ominaisuudet ... Voin jossain määrin laskea alustavasti mitä kokeellisten [kemistien] kanssa kannattaisi tutkia.

Haastateltava 2 kertoi ongelmasta, jossa tutkimusryhmä pohti voiko toksisuutta ennustaa ottamalla huomioon yhdisteen metabolia ja metaboliitit.

Haastateltava 2: Halusimme mallintaa toksisuutta ja sen ennustamista metabolialla. Kysyimme kysymyksen: jos yhdiste ei ole toksinen, onko sen metaboliitit toksisia vai onko yhdiste sittenkin toksinen? Halusimme selvittää, saammeko selkeämpiä tuloksia vai lisäämmekö kohinaa koneoppimisen malleilla [mittausvirhettä] ... Toimimme yhteistyössä kokeellisten kemistien kanssa, jotka validoivat malliimme tarkkuutta ... meillä on suuri mielenkiinto hyödyntää näitä malleja [lääkeaineiden] tuotannossa. Tämän jälkeen tiedetään, kuinka hyvin ne todellisuudessa toimivat [kokeellisesti] ... ongelmana vain aina on sama asia, datan niukkuus ja suuri kemiallinen avaruus, mutta menetelmät toimivat hyvin.

Haastateltava 3 kertoi, kuinka käytti osana laboratoriotöitään keminformatiikan ohjelmistoja kemiallisten ominaisuuksien ennustamiseen.

Haastateltava 3: Ongelmia, joita yritin ratkaista, liittyivät rakenteen tulkintaan. Hyödynsin [keminformatiikkaa] päivittäin osana kokeellisuutta. Esimerkiksi liukoisuuden ennustaminen, LogP-arvo, ja myöhemmin retentioaika. Pidin ideasta, että voin tehdä tutkimuksia tietokoneella ja saada tuloksia aiemmin kuin kokeellisuudella.

4.3 Keminformatiikan osaaminen (TK 2)

Kemian osaaminen näyttäytyi asiantuntijoista keskeisenä, mutta keminformatiikan hyödyntäminen ja rooli monitieteellisissä ryhmissä nähtiin erilaisista näkökulmista joko osana kemistin taitoja tai oman alan asiantuntijan, keminformaattikon taitoja. Tämän lisäksi oppimisen nähtiin erityisesti alkavan ohjelmoinnin alkeista, erityisesti Python-ohjelmoinnin osaamisesta, mutta samalla oltiin kriittisiä siitä, onko kemistien hankala aloittaa työskentelyä keminformatiikassa. Monitieteellisyys asiantuntijuudessa ja keminformatiikan rooli osana kemian menetelmiä ja ohjelmointia näyttäytyvät keskeisinä piirteinä siitä, miten keminformatiikkaa aletaan hyödyntää ja, mitä käyttäjän tulee osata ja oppia. Samalla kemian rooli näytti säilyvän joidenkin haastateltavien mielestä keskeisenä tarpeena, jota perusteltiin kemisteille kehittyneiden metakognitiivisten taitojen vaatimuksena. Käytännön haasteina soveltamiseen pidettiin ohjelmointiosaamisen lisäksi ongelmia menetelmällisyydessä, joka lähestyy enemmän datatieteiden osaamista kemian osaamisen jäädessä kontekstuaaliseksi.

4.3.1 Ohjelmointitaidot

Ohjelmoinnin osaaminen oli haastateltavien mielestä olennaista. Tämä näkyy myös työnkuvassa.

Haastateltava 1 puhui siitä, kuinka ohjelmointitaidot ovat perusteellisia luovalle ongelmanratkaisulle ja korosti tarvetta löytää muualta ohjelmoinnista algoritmeja ja ajatuksia optimoimaan ongelmanratkaisua molekyyleillä:

Haastateltava 1: Kyllä se niin kuin ohjelmointitaidot ja niin kuin yksinkertaisimmillaan sitä, että saisi tehdä Python ohjelmointikielellä asioita... on [iso] puute, jos ei osaa ohjelmoida. On rajoittunut siihen, että käyttää olemassa olevia tuotteita ... Löysin paljon algoritmeja, joita on käytetty arkkitehtuurissa optimoimaan huoneiden järjestystä eri parametreilla. Osaisin ohjelmoida niin, että pystyin tuoda sen kemiaan optimoimaan molekyyleillä ... joo, se on tosi suuri puute [ohjelmointitaidot], ja tekee siitä mahdotonta.

Haastateltavasta 2 ohjelmointi on perustaito, jolla keminformatiikkaa hyödynnetään.

Haastateltava 2: Siis, kemian osuus on erittäin tärkeä, sitä varmasti tarvitaan, mutta kuten monen muun luonnontieteellisen tieteen kohdalla, tarvitset Python-koodaustaitoja vai mitä? Tätä se on nykyään, kaikki toimii Pythonilla. Python on kaikista tärkeintä osata.

Haastateltavasta 3 tämä näkyy myös työpaikkailmoituksissa.

Haastateltava 3: Mielestäni moni työryhmä on monitieteellinen, jossa tarvitaan kemistejä ... mielestäni sen näkyy työpaikka ilmoituksissa, kun tarvitaan keminformatikkaa, ja profiileja on monia, johon sisältyy kemistit, jotka osaavat koodata.

4.3.2 Keminformatiikan datakeskeinen tieteellinen luonne

Keminformatiikan luonne informaatiota tilastotieteellisesti kemian kontekstissa käsiteltävänä monitieteellisenä kokonaisuutena oli samanlainen haastateltavien välillä. Käsitys empiirisen ja teoreettisen aineiston vuorovaikutuksesta toisaalta vaihteli. Erityisesti menetelmät ja lähtökohdat tutkimukselle olivat erosivat toisistaan. Osa haastateltavista ei kokenut aina koneoppimiselle tarvetta, kun taas toiset pitivät sitä toimenkuvalleen välttämättömänä tai perustana. Haastateltavat selkeästi painottavat datan merkitystä mallien testaamisessa ja empiirisen aineiston täydentämisessä. Menetelmällisyydessä molekyylihallinnus ja simulaatiot nähtiin välttämättömänä uutena ”laboratoriona” kokeellisen tutkimuksen rinnalla:

Haastateltava 1: ... me pystymme koneella mallintamaan juttuja menemättä laboratorioon. Se on paljon tehokkaampaa, ei tarvitse käyttää viikkoa labrassa syntetisoiden asioita, vaan [tulokset] voi laskea puolessa tunnissa.

Hypoteesien testaamisesta haastateltava 2 näki keminformatiikassa dataluonteisen perustan sekä yksinkertaisten mallien tärkeyden:

Haastateltava 2: Keminformatiikka on kemiallisen datan talletusta, testausta, jäsentelyä, tiedonhallintaa [data curation], avoimuutta, louhintaa – oikeanlaista analyysia kohti ennusteita ... se on paljolti tiedonhallintaa [data curation] eli datan pitää olla yhtenevää ja standardoidusti vertailukelpoista ... yritetään tuottaa malleja, jotka noudattavat Occamin partaveistä ... mallit pidetään niin yksinkertaisina kuin mahdollista, mikä pitää huolta niitten käyttöönoton takaamisesta.

Mielenkiintoisesti haastateltavat 1 ja 3 olivat taas sitä mieltä, että tarvitaan uusia teknologioita kohtaamaan datan puutetta eli empiirisiä aineistoja, ja kuinka keminformatiikalla voidaan löytää ratkaisuja kokeellisten töiden vähentämiseksi ja seuloa mallinnuksella molekyyliä haluttujen löytämiseksi:

Haastateltava 1: Keminformatiikalla pystytään karsimaan turhia molekyyliä tällaisesta kemiallisesta avaruudesta pystymällä laskemaan se ... käytetään laskennallisia menetelmiä kuten ab initioa, kvanttimekaniikkaa ja tiheysfunktionaaliteoriaa.

Haastateltava 3: Keminformatiikan ongelma on datan niukkuus, joka johtuu siitä, miten dataa kerätään – pienissä laboratorioissa [kokeellisesti]. Keminformatiikalla voidaan vähentää kokeita luomalla simuloituja data settejä, jotka auttavat muun kokeellisen työn ohella. Jos dataa ei ole, sitä voidaan luoda ... voit tehdä kokeita noin viikon, mutta nyt [keminformatiikalla] voidaan ennustaa joitakin tuloksia, mikä vähentää kokeiden määrää ... keminformatiikan tarkoitus on luoda parempia ennusteita.

Haastateltava 2 korosti taas empiiristä dataa tarpeeksi tarkkana alustana mallinnukselle, jossa koneoppimista ei tarvita. Hän korosti tämän sijaan kontekstia mallinnukselle ja niistä saaduille tuloksille.

Haastateltava 2: Tiedonhallinta on tärkeää standardoiduilla säännöillä, mutta jos tietoa on jo riittävästi, 100–200 datasetti, ei sillä kannata tehdä syväoppimista ... ongelmat keminformatiikassa ovat aina samanlaisia. Sinulla on rajallinen datasetti ja laaja kemiallinen avaruus, mitä tutkia. Jos luotettavuus on korkea tai tarpeeksi hyvä, ilmoita se. Ilmoita se kaikille ja kerro miten sitä hyödynnät ... se voi vähentää kokeellisuutta, koska voit omilla dataseiteillä simuloida datasettejä tulevaa kokeellisesta työtä varten.

Haastattelija 3 näki koneoppimisen ja datasettien simulaation välttämättömänä taloudellisena etuna kalliille kemian kokeellisille tutkimuksille. Hän kyseenalaisti kemian alan käytännön hyödyntää pieniä tutkimuslaboratorioita datasettien hankkimiseen. Muutosta perusteltiin vetoamalla paradigmaattiseen muutokseen luonnontieteellisen datan käsittelyssä.

Haastateltava 3: Pyrkimyksenä on kehittää menetelmiä, jotka ovat yleistettäviä ja myöhemmin hyödynnettäviä samaan tai toiseen ongelmaan, mikä vähentää kokeellisuutta. Tämä vie keminformatiikkaa menetelmällisyyteen, kuinka mallintaa laskennallisesti ilmiöt ... mielestäni keminformatiikasta tulee enemmän ja enemmän soveltavaa koneoppimisinsinöörin tai soveltavan datatieteilijän työtä.

Haastateltava 3 toteaa keminformatiikan merkityksen tärkeäksi kustannustehokkuuden myötä.

Haastateltava 3: Nykyajan keminformatiikko yrittää ymmärtää datasta, joka on olemassa, kaiken mitä voi vähentääkseen kokeellisuudesta seuraavia kustannuksia tulevaisuudessa.

4.3.3 Keminformatiikan monitieteellinen tutkimusote

Haastateltavilta kysyttiin monitieteellisyydestä keminformatiikassa. Moni haastateltava piti monitieteellisyyttä yhteistoimintana, jossa eri alan asiantuntijat ratkaisevat samoja ongelmia. Ongelmanratkaisussa keskeistä ovat erilaiset taidot ja

asiantuntijuus. Molempia yhdistää datatieteellisten menetelmien hallinta.

Haastateltava 3 ajatteli, että keminformatiikkaan mahtuu kuitenkin puhtaasti niin kemian työryhmiä ja opetusta kuin monitieteellisiä työryhmiä.

Haastateltava 3: Luulen että Euroopassa on 1–2 maisteriohjelmaa keminformatiikasta, mutta luulen että muut ryhmät ovat monitieteellisiä. Luulen että tarvitaan kemistejä, datatieteilijöitä ja matemaatikkoja. Luulen että tämä näkyy myös avoimista työpaikoista. Välillä tarvitaan kemistiä, joka osaa koodata ja välillä koodaria, jolla on kokemusta biologisista tieteistä ... tässä näkyy se poikkitieteellisyys. Keminformatiikka on laaja ala.

Haastateltavan 2 mielestä, vaikka keminformatiikkaa voi tietojenkäsittelytieteilijänä tehdä, on kemian tietämys ja metakognitiivisiin taitoihin sisältyvä intuitio korvaamatonta.

Haastateltava 2: se [keminformatiikan data] voisi olla melkein mitä tahansa dataa, mutta tarvitset vain herkkyyttä ymmärtää, että datassa on paljon vaihtelua ja ongelmia, kuten virheitä mukaan lukien systemaattisia virheitä ... epämääräisyydet opitaan. Se vaatii kemian tietämystä.

Haastateltava 2: Hyvä asia keminformatiikassa on, että sinun täytyy pääsääntöisesti ymmärtää kemiaa, ja kaikki muu on periaatteessa datan kanssa työskentelyä eli arvioit dataa ... se tarkoittaa sitä, että harjoitat ja hyödynnät kemiaa. Jos joku tulee tietojenkäsittelytieteestä, hänen täytyy olla hyvin varovainen [työskentelyssä], koska hänen on vaikea tehdä samanlaista tutkimusta kemian osaamisellaan.

Haastateltava 3 näkee taas alan väistämättä moninaistuvan: aiemmat mallit ja teknologiat eivät enää riitä kuvaamaan kemiallisia yhdisteitä ja menetelmällisyys uusille yhdisteiden esityksille on monimutkaista ja haastavaa, mikä luo tarpeen monitieteellisyydelle ja tekoälylle.

Haastateltava 3: Ollessamme rajoittuneita kemiallisen datan suhteen, tavoitteena on parantaa menetelmiä, jolla kuvata [paremmin] kemiaa esitysmuodoilla, jotka eivät ole yksinkertaisia. Ne syventyvät matematiikkaan ja tietojenkäsittelytieteeseen, jotka esittävät kemialliset rakenteet monimutkaisempina olioina kuin aiemmin. Edes molekulaarinen deskriptori ei enää riitä. Siksi neuroverkot, joita harjoitetaan ymmärtämään kemiallista dataa metatasolla ovat mielestäni se monimutkainen, matemaattista ja teknistä osaamista vaativa asia, ja sama koskee mallien tulkintaa.

4.4 Keminformatiikan arvo opetuksessa ja oppimisessa TPASK-mallilla analysoituna (TK 3)

Haastatteluaineiston sisältöä luokiteltiin yläluokkiin SK, TK ja TSK. Rajausta tehtiin edellä mainittuihin luokkiin, koska TPASK-mallin osaamisalueiden kriteerien ei nähty täyttyneen luokille PK, PSK, PTK ja TPASK (Kts. Luku 2.4.2 ja 3.1).

Yläluokkien lisäksi sisällönanalyysissä oli myös alaluokkia, joiksi aineistosta pelkistyi saavutettavuus, monitieteellisyys, kemian simulointi sekä datan kartoitus ja mallintaminen. Pääluokkia oli kaksi: keminformatiikan muuttuva luonne sekä keminformatiikan menetelmällisyys. Pääluokista johdettiin neljä teemaa, jotka selittävät aiempien luokkien yhteyttä toisiinsa. Luokitusten lukumäärällä haluttiin ilmaista analyysissä osaamisalueiden jakautuneisuutta (Esim. TK ja SK) ja vuorovaikutusta (Esim. TSK). Taulukkoon 4. eriteltiin sisällönanalyysin luokat ja teemat. Liitteessä 1 on kaikkien luokitusten pelkistetyt ilmaisut.

Taulukko 4. Yhteenveto alaluokista, yläluokista, pääluokista sekä teemoista sisällönanalyysissä.

Alaluokka	Yläluokka (TPASK-malli)		Pääluokka	Teema
Monitieteellisyys (N=4)	SK (N=16) TK (N=9)	TSK (N=9)	Keminformatiikan muuttuva luonne	Tieteellisen tiedon monitieteellisyys
Datan kartoitus ja mallintaminen (N=4)				Keminformatiikan datakeskeinen luonne
Kemian simulointi (N=9)			Keminformatiikan menetelmällisyys	Ohjelmointitaidot
Tiedon saavutettavuus (N=2)				Keminformatiikan hyödyntäminen

4.4.1 Tieteellinen tieto (SK) voi laajentaa kemian ajattelua

Tieteellinen tieto (SK) luokiteltiin, kun haastateltavat selvensivät näkemyksiään kemian ja keminformatiikan välillä. Vastaukset osoittivat, että keminformatiikka vaatii laajaa kemian tietorakenteiden hyödyntämistä ja yhdistämistä:

Haastateltava 1: Keminformatiikka vaatii fysikaalisen kemian tietämystä.

Haastateltava 3: On monia, jotka yrittävät ymmärtää mallien tulkinnaisuutta ja laajennettavuutta, näissä ongelmissa tarvitaan paljon kemian tietoa.

Haastattelija: Mitä arvoa keminformatiikalla on kaikille kemisteille?

Haastateltava 2: Nykyään ei ole järkeä syntetisoida mitään ilman vahvaa hypoteesia ja tutkimussuuntausta, tietäen mitä haluaa tehdä. Kyse on siitä, että, jos kemisti ei hyödynnä helposti käyttöönotettavia keminformatiikan työkaluja, jotka perustuvat objektiivisiin ja testattuihin malleihin, niin työskentelyssä ei ole järkeä tai se siitä tulee hyvin työlästä. Resursseja hukataan ja kilpailukyky heikkenee ... ja keminformatiikka on vain tärkeä osa tutkimusta ... kollegalta saa subjektiivista pohdintaa ja tietoa, kun taas tietokoneen mallit ovat objektiivisia.

Haastateltavien mielestä siis keminformatiikan hyödyntäminen ei ole vierasta, vaan osa kemian mallintamista ja mallien ymmärtämistä. Kemian tietorakenteiden vahva hyödyntäminen keminformatiikassa voi olla mahdollisuus kemisteille tutustua uusiin menetelmiin ratkaista kemian ongelmia. Oppimisesta voi edistyä, jos aiempi tietämys on tarpeeksi kattava. Haastateltavalta 2 kysyttiin, miten keminformatiikka muuttaa tiedettä ja onko kyse niin sanotusta neljännestä tieteellisestä paradigmasta. Hänen mielestään kyse on enemmän metakognitiivisella tasolla tapahtuvista ajatusprosesseista ja niiden muuttamisesta:

Haastateltava 2: [Keminformatiikka] on mielestäni ideoita herättävä kone ... se on vain tärkeä osa sitä [ongelmanratkaisua]. Siihen ei tarvitse luottaa ... mutta tietokoneen mielipide on objektiivinen ... ja ottaa huomioon ideaalissa tapauksessa koko tietorakenteen ja antaa riippumattoman ennusteen.

Haastateltava 1: Siinä työskentelee metakognitiivisella tasolla. En manuaalisesti rakenna molekyylejä, vaan tässä yhdistyy luova prosessi kysymyksen asetteluun eikä se ole teknistä puurtamista. Ja tässä on kanssa hauskaa, että ongelmat, joita ihmiset on yrittäneet aiemmin ratkoa manuaalisesti voi ollakin keminformatiikan ongelmia.

Kun kemistit arvioivat omia tietokokonaisuuksiaan, voi se edellä mainittujen sitaattien perusteella johtaa uudenlaisiin ajatusprosesseihin kemian menetelmällisyydestä sekä tavastaan ratkaista ongelmia. Opetus voi siis saada aikaan tutkimuksen tekemiseen ja tutkimus ongelmien ajatteluun liittyviä uusia ajattelumalleja ja näin tukea kemian oppimista.

4.4.2 Teknologinen tieto (TK) täydentää tieteellistä menetelmää

Haastateltavien vastauksista ilmeni, mitkä teknologiat ovat keminformatiikan ytimessä. Haastateltava 1 korosti algoritmeja, kun taas haastateltava 2 reflektoi

viimeisen kahdenkymmenen vuoden teknologisia muutoksia keminformatiikassa ja miten niihin vastattiin:

Haastateltava 1: Algoritmien opettelua ... löysin paljon algoritmeja, mitä on käytetty arkkitehtuurissa optimoimaan huoneiden järjestystä ja muuta. Sitten osasin ohjelmoida ja tuoda sen kemiaan optimoimaan molekyylejä.

Haastateltava 2: Kun tein väitöskirjaani, se mihin keskityttiin konformeerien enumerointi, mikä on ydin teknologia mitä tarvitaan molekyyliden mallinnuksessa ... keminformatiikasta tuli paljon suurempaa ja nyt meillä on paljon dataa ... nyt se on datan käsittelyä.

Haastateltava 3 ajatteli, että teknisiltä taidoiltaan, kemisteille olisi hyödyllistä opettaa kemometriikan perusteet. Samalla hän epäili, tarvitseeko kemistien syventyä koneoppimisen menetelmiin.

Haastateltava 2: Yliopistoissa on usein kemometriikan kursseja, ja kyllä perusteet, kuten koneoppimisen ja tilastollisten menetelmien soveltamista kemialliseen dataan, kuten spektreihin pitää olla olemassa. Siinä tärkeää on esittää kemiallista tietoa ja sovittaa se kemiallisiin ongelmiin tiedetyistä rakennemalleista. Mutta tämä ei tarkoita, että pitäisi syventyä syväoppimisen menetelmiin jne. Mielestäni se on toisenlainen maailma.

Kun haastateltavilta kysyttiin, mitä ohjelmistoja nykyään käytetään. Lisäksi kysyttiin haastattelijalle tunnetusta ohjelmiston Open Babelin käytöstä. Haastateltavan 1 mielestä oli sukupolvikysymys mitä ohjelmistokieltä tai ohjelmistoja käyttää:

Haastateltava 1: Se on sukupolvikysymys. Huomaan että ihmiset, jotka sanotaan, aloitti 2005, silloin kun laskemia vaiheita läpi, niin ne käyttävät Open Babelia. Ne käyttävät C++ ja C:tä ja Javaa. Mutta nyt ne, jotka on aloittanut tämän uuden tieteellisen laskennan buumin aikana, mikä on koneoppimiseen ja datatieteeseen hyvin vahvasti sitoutunutta, niin se on tosi vahvasti Python pohjaista.

Tätä vahvasti haastateltavan 2 näkemys Pythonista johtavana ohjelmistokielenä, ja hän mainitsi muutaman apupaketin ja ohjelmiston nimeltä. Samaan yhtyi haastateltava 1.

Haastateltava 2: Tarvitsen Pythonia. Sinun tarvitsee opetella SciKit-learn vai mitä? Nykyään ihmiset myös käyttävät PyTorch -sovelluskehys ja kyllä, on keminformatiikka kirjastoja käytettävissä, kuten RDKit. Nämä ovat pääkomponentit, millä pääsee hyvin pitkälle.

Haastateltava 1: Kielenä on Python ja tutkimustulokset pitkälti julkaistaan avoimen tieteen periaatteilla ... Python on koneoppimisen kieli ja se kulkee kemoinformatiikan kanssa käsi kädessä ... RDKit on liima, joka liittää kemian tähän muuhun Pythoniin.

Kemian oppimiselle voidaan pitää edellytyksenä riittäviä ohjelmointitaitoja sekä sopivia ohjelmistoja, jotka eivät täysin rajaa ohjelmointikieltä Pythoniin, vaikka koneoppimisen menetelmissä siitä on erityistä hyötyä. Samalla nousi esiin este oppimiselle: koneoppimisen menetelmät voivat olla liian haastavia rakentaa ilman laajempaa koulutusta ohjelmointiin.

4.4.3 Teknologis-tieteellinen tieto (TSK)

Teknologis-tieteellinen tieto, TSK tulkittiin haastateltavien vastauksista seuraavanlaisesti: ajattelua ja keminformatiikan soveltamista tieteelliseen menetelmään, tulkittiin muuttuneina ajatteluprosesseina ja käytännön mahdollisuuksina edistää kemian opetusta ja oppimista tieteellisen tiedon (SK) ja teknologisen tiedon (TK) yhdistelmänä (TSK).

Esimerkiksi haastateltava 1 esitti, kuinka keminformatiikalla ei haeta yhtä tapaa mallintaa kemialla, vaan se ilmenee jatkumona kehittää uusia ratkaisuja ja ideoita algoritmeja hyödyntäen. Keminformatiikka voidaan siis nähdä uutena työkaluna lähestyä kemian ongelmia. Samalla se nähtiin osana tieteellistä työskentelyä ja menetelmää: ongelman esittäminen, tutkimuskysymyksen laadinta, teorian muodostaminen ja testaus, ja ennen kaikkea, uuden ongelman kohtaaminen sopivalla menetelmällä.

Haastateltava 1: Aiemmin on ollut paljon ongelmia, mitä on ratkottu manuaalisesti, ja keminformatiikalla niistä saadaan algoritmisia ongelmia. Ne ei ole aina sellaisia, että tehdään jostakin joku yleinen versio ja sitten algoritmisesti ratkaistaan se. Siinä tulee jatkuvasti uusia ideoita mitä tutkia seuraavaksi.

Keminformatiikka ja sen tavoitteet voivat siis kehittyä ja uudistua algoritmisella mallintamisella. Haastateltava 1 jatkaa esimerkkiin vedenpilkkomiskatalyytista:

Haastateltava 1: Ongelmana on vaihtaa ruteniumatomi keskeltä rauta-atomiin, ja mimmoisia muokkauksia pitää tehdä ligandikenttään, että se toimisi. Sitten se voi elää, että miksi yleisesti pitäisi olla metalliatomi? Voisiko boori toimia ja sitoa ligandeja? [Keminformatiikan menetelmällisyydestä] se on enemmän iteratiivinen prosessi, joka alkaa spesifistä kysymyksestä kohti yleisempää. Sitten kun on päästy siihen yleisempään, niin halutaan ratkoa uusia spesifejä ongelmia ja tilanne uusiutuu.

Haastateltavasta 3 oli aiemmin vastannut (luku 4.1.), kuinka keminformatiikka tavoittelee säästämään resursseja ja kokeellisuuden määrää, ja kuinka simuloituakin

dataa voidaan tuottaa (Luku 4.3.2). Nämä voidaan tulkita muutoksena siitä, miten kemiallista dataa hankitaan, käsitellään ja hyödynnetään datakeskeisesti.

Haastateltavan 3 aiemmat kommentit voidaan tulkita täten TSK:na.

Aiempien kommenttien lisäksi, haastateltava 3 hahmotti, miten keminformatiikka vaikuttaa muihin tieteisiin ja luo uudenlaisen monitieteellisen käsityksen kemiasta suhteessa muihin kuin kemiallisiin tieteisiin, jossa kemiallisen rakenteen tarkastelu on keskiössä.

Haastateltava 3: Näin itse näin asiat. Laskennallinen kemia lähestyy koko ajan keminformatiikkaa tai QSAR-mallinnusta. Se on ytimessä, kun hyödynnetään koneoppimisen menetelmiä yhdisteiden ominaisuuksien ja biologisen aktiivisuuden mallinnukseen. Ja koska tarvitaan dataa, täytyy sinun saada dataa laskennallisesti eli simuloitua dataa mitä QSAR tarvitsee.

Haastateltava 3: Alasta tulee sulatusuuni, jossa tarvitaan matematiikkaa ja fysiikkaa laskentaan, kuten DfT:hen ja muihin menetelmiin tarvitaan datatieteilijöitä ja tietojenkäsittelytieteiden asiantuntijoita. Kaikki sekoittuu.

Haastateltava 3 jatkaa siihen, kenelle suosittelee keminformatiikan opintoja yliopistossa:

Haastateltava 3: Sen pitäisi olla joku, jolla on vahva perusta tietojenkäsittelytieteessä ja matematiikassa, mutta myös fysiikassa. [Osaamiseen] Tarvitaan paljon vähemmän kokeellisuutta.

Haastateltava 2 mainitsi jo aiemmin, kuinka keminformatiikka on datakeskeistä ja sen menetelmällisyys on samanlaista eri aloilla, mutta tämän lisäksi hän korostaa kuinka keminformatiikan tekeminen ja hyödyntäminen avoimeen lähdekoodiin perustuen luo puitteet tieteelliselle keskustelulle ja oppimiselle:

Haastattelija: Sanoisitko että keminformatiikan ohjelmistokehitys on ajurina, kun tekijät ovat tutkijoita?

Haastateltava 2: Se ei ole niinkään ohjelmistokehitys, vaan kun olet käyttäjä ja jos sinulla on kysyttävää, niin on yhteisö, jolta kysyä ... tarvitaan yhteisöä, joka auttaa käyttäjiä. Ja jos käytät esimerkiksi RDKit-ohjelmistoa, on hyvin todennäköistä, että joku keskustele kanssasi ja auttaa. Se on ratkaisevaa.

Tätä puoltaa myös se, että haastateltava 2 katsoo, että keminformatiikasta voidaan saada saavutettavaa, helposti käytettävää ja että se on itsessään arvokasta, vaikka toteaa, ettei tämä ole vielä toteutunut. Tämä korostaa erityisesti luvuissa 4.3.2. ja

4.4.1 esiintyneitä haastateltavan näkemyksiä menetelmien objektiivisuudesta ja hyödynnettävyydestä:

Haastateltava 2: Suuri apu on siinä, että on kääntäjiä [keminformaattikkoja] ja kemistejä. Keminformaattikot auttavat kemistejä, että saadaan ohjelmistoista saavutettavaa. Kemistit voisivat hyödyntää vain intuitiota, mutta ohjelmistot tuovat vain yhden lisäaskeleen [objektiivinen malli] prosessiin. Mutta voin kertoa, että en ole vielä tavannut kemistiä, joka hyödyntäisi keminformatiikan potentiaalia täysin. Toisaalta en tunne ketään, joka luopuisi näistä menetelmistä myöskään. He jotka ovat kokeilleet ymmärtävät kuinka yksinkertaista sen käyttö on, eivät tee reaktioita ilman sitä ... Kaiken lisäksi on vielä halpaa.

Lopuksi haastateltava 2 nostaa, kuinka keminformatiikka on vain osa laajempaa taitoa hyödyntää dataa ja tämä taito on helposti yleistettävissä muihin ongelmiin. Keminformatiikan opetus voi mahdollistaa oppimisen hyödyntämisen myös muilla aloilla.

Haastateltava 2: Opit vain kemiallisen tiedon esimerkein, ja samat ongelmat toimisivat esimerkiksi kaupallisella alalla tai muualla vai mitä? Tarvitset vain peruskäsityksen [data-analyysistä] ... se on monipuolinen taito.

5 Keskustelu ja johtopäätökset

Tässä luvussa vastataan tutkimuskysymyksiin sekä keskustellaan tutkimuksen tuloksista ja esitetään johtopäätökset. Tuloksista keskustellaan peilaamalla hyödyntäen aiempaa kirjallisuuteen ja tutkimukseen keminformatiikasta sekä sen opetuksesta ja oppimisesta. Lopuksi arvioidaan tutkimuksen luotettavuutta ja jatkotutkimustarpeita.

5.1 Keminformatiikka kemian tutkimuksessa (TK 1)

Kemian asiantuntijat hyödynsivät keminformatiikan dataa ja menetelmiä erilaisiin kemian ongelmiin ja prosesseihin. Keskeisenä toimintamallina korostui kemian tiedon datatieteellinen soveltaminen mallinnuksiin, jotka pyrkivät pelkistämään selkeitä trendejä suurista aineistoista. Keminformaatikkojen ajattelua ohjasi saatavilla oleva data, standardoidut menetelmät, validoinnin merkitys ja tiedon kriteerit sekä kriittinen arviointi mallintamisen tarkkuudesta. Avainasemassa oli kemian tietotekninen kuvailu, jolle kirjallisuudessa Fourches ym. (2010) ajatukset tiedon ja menetelmien validoinnista, ja Birdin & Freyn (2010) kuvaama tiedonkäsittelyn kulttuuri ja datan löydettävyyss vastasivat odotuksia.

Hyödynnetty keminformatiikka näyttäytyi hyvin datariippuvaiselta, ja alan suurimmaksi ongelmaksi koettiin kemiallisen tiedon niukkuus. Toisaalta pieniin aineistoihin data-analyyseja ei nähty tarvetta, mutta keminformatiikka yleisesti sovelletaan kuitenkin laajoihin aineistoihin (Gromski ym., 2019; Saifi ym., 2023). Syitä tähän oli avoimen datan puute, datan kustannukset ja se, ettei tutkijat tyypillisesti omista dataansa. Tämä rajoittaa keminformatiikan menetelmällisyyttä, mutta samalla ohjaa muutosta keminformatiikan suhdetta kokeelliseen kemiaan, joka vaikuttaa myös tiedon muuttumiseen. Kokeellisesta kemiasta tulee tapa validoida tuloksia, ja toimia pohjana kemian simulaatiolle uusiksi aineistoiksi. Toisaalta myös kokeellisuuden automaatio, kuten HTS-menetelmä ja sen tulosten analyysi koneoppimisen ja tekoälyn mallien avulla (Attene-Ramos ym., 2014; Chen ym., 2018; Dreiman ym., 2021) nähtiin tutkijoista mielekkäänä ratkaisuna aineistojen laajentamiselle kustannustehokkaasti. Tämän lisäksi osoittautui, että haastateltavan 2 mielestä keminformatiikan tutkimuksen avoimuutta ajaa voimakkaasti ulkoiset tekijät, kuten säätely, mistä esimerkiksi QSAR-mallien kriteerit kuvailusta, menetelmällisyydestä julkistamisesta ovat peräisin (Diderichs, 2010; ECHA, 2008).

Tämä mielenkiintoinen näkemys siitä, kuinka kemiallisen datan hyödyntämistä voi lisätä teollisuuden säätelijän omaksumat arvot sen sijaan, että avainasemassa olisi tutkimusyhteisö, jonka avoimen lähdekoodin ja avoimen tieteen periaatteista on puhuttu jo pitkään (O'Boyle, 2011).

Ongelmanratkaisussa korostui keminformaattikkojen pyrkimykset laajentaa malleja sekä simuloimaan dataa. Tarkoitus tälle on ennustaa olemassa olevasta datasta jotakin, mistä ei ole vielä dataa tai jotakin ilmiöstä, jota ei tunneta. Koska nämä ovat tietojenkäsittelytieteen sovelluksia (Tolk, 2018), vahvistaa tämä käsitystä keminformatiikasta hyvin keskeisenä tietojenkäsittelytiede painotteisena. Tätä vahvisti myös se, että haastateltavista 2 ja 3 olivat samaa mieltä syistä simulaatioille eli datan puutteesta, ja mihin malleilla pyritään. Haastateltava 2 painotti kuitenkin kokeellisuuden validoivaa vaikutusta malleille, mutta haastateltava 3 korosti ajattelussaan kemian simuloinnin merkitystä kokeellisen datan korvaamiselle luottaen, että menetelmä voi korvata olennaisesti osan kokeista. Hän näki avoimen datatieteelliset ja monitieteelliset periaatteet ratkaisuna, kun taas haastattelija 1 keskittyi enemmän simuloitujen mallien laajennettavuuteen. Tämä kuvaa keminformaattikkojen erilaisia tapoja tehdä tiedettä, joka heijastuu myös kokemuksessa: haastateltava 1 on vertailussa uusin tutkija alalla, haastateltava 2 kokenut akatemiassa ja teollisuudessa, kun taas haastateltava 3 muutamassa vuodessa erikoistunut moninaisimmin menetelmällisyyteen akatemiassa.

Järkeilyltään keminformaattikot perustelivat alansa merkittävyyttä kemiassa eri tavoin: haastateltava 3 perustelu kokeellisuuden korvaamiseksi perusteltiin datan hinnalla. Tämä voidaan nähdä näkemyksenä, jossa tutkija osoittaa keminformatiikan olevan hyödynnettävissä datakeskeisenä tieteenä, jota tukee Heyn ym. (2009) tieteen neljännen paradigman ajattelu ja Reisigin (2020) käsitys tietojenkäsittelytieteen ja luonnontieteiden samanlaisesta mallinnuksesta. Haastateltava 2 oli varovaisempi, ja luotti kokeellisten kemistien yhteistyöhön mallien validoinnissa ja korosti myös, että in silico menetelmät eivät koskaan ole luotettavia ja vain tiedettyyn tietoon voidaan vastata keminformatiikalla varmasti. Muissa tapauksissa malli antaa arvion pitääkö se paikkaansa vai tarvitaanko kokeellisuutta. Tähän yhtyi myös haastateltava 1.

Mielenkiintoisesti haastateltava 1 sanoitti keminformatiikan edistävän hänen metatason ajatteluaan ja luovaa ongelmanratkaisua, mikä voi tehdä menetelmistä houkuttelevia kemian oppimiselle ja työskentelyn tukemiselle. Erityisesti datan

suurien määrien käsittelyn ongelmaan, mihin vain keminformatiikan perusteilla tiedonhallinnasta (Gasteiger, 2006) sekä koneoppimisen ja tekoälyn datankäsittely tehon perusteella voidaan vastata (Zohuri & Sadeh, 2020). Haastateltava 1 suhteutti keminformatiikan etuja erityisesti laskennalliseen kemiaan, jota kuvaili ”manuaaliseksi”, jossa perinteisesti tutkitaan yhtä tai muutamaa molekyyliä, kun taas keminformatiikka tutkii usein tuhansia sen jälkeen, kun ne on seulottu vielä suuremmasta joukosta yhdisteitä. Merkittävää eroavaisuutta laskennalliseen kemiaan ei kuitenkaan selvitetty haastattelussa, vaikka tämä olisi ollut kiinnostavaa tietoa siitä, miten keminformaatikot eroavat laskennallisen kemian asiantuntijoista ja työskentelytavoista.

Konkreettiset hyödyt, kuten yhdisteiden ominaisuuksien ennustaminen ennen laboratoriotyöskentelyä tai sen täydentämiseksi voivat olla käyttökelpoisia sovelluksia, mutta toisaalta empiiristä dataa löytyy myös erilaisista tietokannoista jo nyt monille yhdisteille (Esim. Kim ym., 2019). Samalla, mutta eri mittakaavassa ja kompleksisuudessa, yhdisteiden ennusteet ainutlaatuisista tai monimutkaisista yhdisteistä ei ole yksinkertaista. Keminformatiikan käyttökelpoisuus näihin ongelmiin näyttäytyi asiantuntijoissa vahvana eetosena luottaa siihen, että olemassa olevista malleista voidaan saada jotakin empiiristä selville, mitä ihmismielen prosessointi kyky ei riitä samassa ajassa ja yhtä suurilla data määrillä pysty saavuttamaan. Tutkimusta kuitenkin tekevät ihmiset, jotka laativat tieteelliselle työskentelylle periaatteita (March & Smith, 1995) ja luovat tietoa yhteisössä eikä kokeellisuutta tai keminformatiikkaa erikseen tule tehdä. Keminformatiikalla pyritään muuttamaan kemian asiantuntijuutta hyödyntämään menetelmällisyyttä, jossa tietokoneiden laskentatehoa ja optimointia mallinnukseen laajennetaan. Tällä ei kuitenkaan tarkoiteta sitä, että keminformatiikalla pyrittäisiin korvaamaan kokeellisuus. Kokeellisuus nähdään roolissa, joka validoi malleja, jota keminformatiikka pyrkii approksimoimaan aiemmasta datasta paremmaksi malliksi, johon ei päästä yhtä nopeasti kuin kokeellisuudella.

5.2 Keminformatiikka on ohjelmointia ja kemian mallintamista (TK 2)

Asiantuntijoiden mielestä, keminformatiikassa on olennaista hyödyntää ohjelmointia ja erilaisia algoritmeja sekä niiden muokkausta. Haastattelijoiden mielestä keskeisintä on hallita Python-ohjelmointikielen perusteet, että sitä voi lukea ja

parhaassa mahdollisimmassa tapauksessa, muokata ja kirjoittaa. Tätä korostettiin työpaikkailmoituksilla sekä mallintamisen taidoilla.

Asiantuntijoissa näyttäytyi ero taitojen priorisoinnissa: haastateltavan 2 mielestä datan luotettavuus, arviointi ja tiedonhankinta nähtiin olennaisena tapana muokata dataa omaan käyttöön sekä valitsemaan oikeat menetelmät suhteessa datan määrään. Kun taas haastateltavat 1 ja 3 nojasivat lausunnossani enemmän omien menetelmien ja uusimpien ideoiden hyödyntämiseen. Datan tiedonhallinta ja sen menetelmällisyys, joihin esimerkiksi molekulaariset deskriptorit (Cereto-Massagué ym., 2015) lukeutuvat, eivät haastateltavan 3 mielestä riitä enää tutkimuksen tarpeisiin. Haastateltavan 3 sanoma osoittaa, että kemia ja ohjelmointitaidot eivät riitä yksin keminformatiikan ongelmiin, vaan monitieteellinen työskentely on olennaista. Tämä tulkittiin alan nopeaksi kehitystahdiksi, mutta samalla aiheuttaa epäselvyyttä kuinka syvällistä tietoa ja taitotasoa keminformatiikka vaatii.

Suomen koulutusjärjestelmässä kemian yliopisto-opetuksessa voi olla tämän perusteella ohjelmointitaitojen kehittämiseen mahdollisuuksia keminformatiikan opetuksessa. Aiemman tutkimuksen tai kurssien näkyvyyden puute voi selittää, miksi keminformatiikka ei esiinny nimellä kemian opetuksen tutkimuksessa tai suomenkielisessä kirjallisuudessa Wildin (2021) teoksen suomennosta lukuun ottamatta. Hardyn ym. (2021) tutkimuksessa on mainittu keminformatiikan mahdollisuuksia yliopisto-opetukseen perustuen sen monitieteellisyyteen, kun taas teollisuudella on tavoitteena hyödyntää kemian datasta tekoälyllä ratkaisuja erilaisiin ongelmiin (Takeda ym., 2020). Tämän voi tulkita haastatteluaineiston ja tutkimuskirjallisuuden samana tarpeena muuttaa opetussuunnitelmia ottamaan huomioon nykyajan työpaikkojen tarpeita. Taitoja oppii työskentelyllä ja ratkaisu voi esimerkiksi olla pilotoida keminformatiikan kursseja erilaisille kohderyhmille, kuten alumneille tai jatko-opiskelijoille. Tämä tukee suomalaisen yliopistokoulutuksen tavoitteessa taata elinikäistä oppimista (Louko & Blomqvist, 2018). Keminformatiikan verkkokurssit, kuten Kim ym. (2021) sopisivat myös hyvin taitojen ja käytänteiden haltuunottoon, koska ne tehdään etänä ja tietokoneella.

Toisaalta toivottujen taitojen ja tutkimuksen muuttuvat luonteen välille kuitenkin voidaan todeta ongelma: vaikka datatieteellisyys näyttäytyy aineistossa, on keminformatiikan toteutuminen tulosten valossa toisenlaista, kuin Hey ym., (2009) lupasivat paradigmaattisesta muutoksesta tieteiden alalla. Dataa ei ole runsaasti,

vaan niukasti, ja keminformaatikot haastavat tämän aselman kehittämällä menetelmiä datakeskeisesti eikä datalähtöisesti. Data on keskiössä, koska se on kallisarvoista. Toisaalta (Hey ym., 2009) kirjan väite neljännestä tieteen paradigmasta voidaan nähdä kemian datan ja keminformatiikan suhteen toteutumattomana, koska dataa ei ole runsaasti asiantuntijoiden käytössä, kuten Heyn ym. (2009) malli pyrkii selittämään. Toisaalta joissain soveltavissa tieteissä, kuten farmasiassa on kiistatonta, että keminformatiikan malleilla voidaan ratkaista lääkekehityksen ongelmia ennen kokeellisuutta perustuen koneoppimiseen ja tekoälyyn (Brown, 1998, s. 375; Lo ym., 2018). Keminformatiikka ja taito hyödyntää sitä näyttäytyy asiantuntijoille kemian kokeellisuutta täydentävänä tai kokonaan korvaavana menetelmänä pakotetusti, johon etsitään alati monimutkaisia ja monitieteellisiä ratkaisuja.

5.3 Keminformatiikka ja TPASK: mahdollisuudet kemian opetukseen (TK 3)

TPASK-mallin hyödyt opetukselle tapahtuvat alaluokkien ja teemojen vuorovaikutuksessa, ja erityisesti teknologiset-tieteellinen tieto, TSK kehitti keminformatiikan omaksumista.

Haastateltavien kyky hyödyntää teknologiaa nojaa aiempiin tietorakenteisiin kemiallisesta tiedosta, joka näyttäytyi osin fysikaalisen kemian ja mallien ymmärtämisenä. Tämä vastaa aiempaa käsitystä keminformatiikasta laajalti laskennalliseen kemian malleihin perustuvana soveltavana tieteenä (McArdle ym., 2020). Mallien objektiivisuus, yksinkertaisuus ja järkeily, joka ei perustu enää vain ihmisen ajatteluun, vaan vuorovaikutukseen mallien ja käyttäjän kanssa korostui vastauksissa. Tätä tuki väitteet siitä, kuinka keminformatiikan hyödyntäminen tapahtuu metakognitiivisen tason ajattelulla herättäen ideoita. Nämä prosessit osoittavat tieteellisen tiedon, SK:n piirteiden ymmärtämistä, joka voi myös konkretisoitua oppimisessa, joka kehittää mallinnuksen taitoja ja pyrkii objektiivisuuteen. Toisaalta on muistettava, että haastateltavat ovat asiantuntijoita, joiden tiedot ja taidot ovat työn avulla jo kehittyneet, eli tietorakenteiden ero opiskelijoihin voi olla suuri.

Keminformatiikan käyttöönotto nojaa taidoiltaan vahvasti valmiuksiin ohjelmoida. Ohjelmointitaitojen keskeisyys, vaikkakin haastateltavista eri tasoilla, vastaa myös muitten tutkimusten tuloksia keminformatiikan soveltamisesta yliopisto-opetukseen

(Esim. Hardy ym., 2021). Toimenkuva, jossa keminformaattikot ohjelmointia hyödyntävät on selkeästi Python-ohjelmointia, eli se vaatii teknologisen tiedon, TK:n osaamista. Riittävät taidot keminformatiikan hyödyntämiseen tulee siis ottaa huomioon oppimisen ja ohjelmistojen soveltamisessa. Oppimisessa ohjelmointi voidaan nähdä teknisen tiedon, TK:n edistämiseen, mutta kun se keminformatiikka ohjaa tutkimusta, siitä voi tulla teknologis-tieteellisen tiedon, TSK:n osaamista. Jimoyianniksen (2010) mukaan tätä hyödynnetään oppiminen, jossa teknologia muuttaa prosesseja, jolla tieteellistä tietoa tulkitaan, tutkitaan eri resursseilla ja työkaluilla voi muuttaa aiempia käsityksiä.

Osaamistavoitteet, joilla pyritään ottamaan käyttöön keminformatiikan työkaluja, tulee hyödyntää jossakin tutkimuskontekstissa, missä keminformatiikan rooli on olennainen. Tämän lisäksi, keminformatiikan datatieteelliset menetelmät ja niiden soveltaminen on myös poikkitieteellistä, eli niitä voi hyödyntää myös muissa tutkimusongelmissa. Lääketeollisuuden tutkimusongelma löytää suuresta ja laajalti kokeellisesti kartoittamasta joukosta molekyyliä lääkeyhdisteitä on hyvä esimerkki, joka nojaa keskeisesti keminformatiikkaan. Tällöin orgaanisten yhdisteiden enumerointi, virtuaaliseulonta, HTS-datan tulkinta voivat muuttaa käsityksiä keminformatiikan mahdollisuuksista ja rajoitteista, kuten tarpeesta validoida tulokset kokeellisesti. Tietokantojen arkipäiväinen ja tehokas hyödyntäminen sekä mahdollisuudet ennustaa yhdisteiden rakenteista niiden ominaisuuksia, voivat myös edistää tulosten tulkintaa kokeellisessa työskentelyssä.

Oppiminen on vaikuttavaa, kun teknologiaa hyödynnetään aiempien tietorakenteiden kriittisessä tarkastelussa. Ja opetus, joka pyrkii herättämään ideoita ja ohjaa tarkastelemaan metakognitiivisesti kemian tieteellistä menetelmää täyttää TPASK-mallilla erityisesti TSK:n kriteerit. Opettajankoulutuksessa, kuten Rodríguez-Becerra ym. (2020) tutkimuksessa, laskennallisella kemialla oli samanlainen päätelmä TSK:n roolista muuttaa opiskelijoiden ajattelua kemiasta ottaen käyttöön uusia niin keminformatiikan kuin laskennallisen kemian ohjelmistoja ja resursseja.

Toisaalta oppimiselle ja opetukselle voi olla myös esteitä. Haastateltavien ja kirjallisuuden suhteen keminformaattisten mallien monitieteellisyys voi osoittautua teknisiltä tiedoilta haastavaksi. Tämä muodosti vastakkaisen asetelman tuloksissa siitä, mitä alalla tarvitaan ja miten kemistien kyky ymmärtää keminformatiikkaan eroaa mm. koneoppimisen asiantuntijoiden kanssa. Tähän kuitenkin vasta-

argumenttina haastateltava 2 totesi, että yhteisön kyky auttaa kysymyksissä ja ongelmissa, sekä vastuu auttaa keminformaattisten mallien raportoinnista voi luoda yhden ratkaisumallin. Tämä ratkaisumalli tulkittiin erityisesti teknologis-tieteellisen tiedon kehittymisenä. Tulos korostaa avoimen datan ja menetelmällisyyden tarkan kuvailun arvoa. Tämän lisäksi niin säätely (Diderich, 2010; ECHA, 2008) kuin avoimeen lähdekoodiin perustuva ohjelmistokehitys voivat tehdä datasta ja keminformatiikasta saavutettavampaa myös opetukseen (Pernaa ym., 2023).

Yleisesti keminformatiikka nähtiin mielekkäänä tapana täydentää kemian tietämystä. Korkeakouluopetuksessa ja TPASK-mallilla analysoituna metakognitiivinen ajattelu ja ohjelmointitaitojen kehittäminen ovat mahdollisia kehityskohteita osaamistavoitteisiin. Ideoiden ja luovuuden kehittyminen vuorovaikutuksessa tietokoneiden kanssa tulkittiin positiivisena, mutta väitteen luotettavuudelle pitää suhtautua varoivaisesti, koska haastateltavat ovat kehittäneet tietorakenteitaan enemmän keminformatiikan myötä asiantuntijuudessaan. Saavutettavuus ja yhteisön tuki ovat merkityksellisiä askelia keminformatiikan käyttöönotossa, mikä voi selittää myös, miksi verkkokursseilla keminformatiikan asiantuntijat ovat osa opetussuunnittelua (Esim. Kim ym., 2021).

5.4 Johtopäätökset ja jatkotutkimustarve

Tutkielman toimii ensimmäisenä kemian opetuksen laadullisena tutkimuksena keminformatiikan mahdollisuuksiin ja opetukselliseen arvoon. Tutkittavien asiantuntijoiden käsityksistä keminformatiikan hyödynnettävyydestä yliopisto-opetukseen muodostui selkeitä oppimista kehittäviä kehittämistarpeita suomalaiseen yliopisto-opetukseen, jossa keminformatiikan roolia pitäisi tutkia enemmän sen arvolta ja soveltuvuudeltaan. Keminformatiikan hyödyt osana modernin teknologian ja laadukkaan opetuksen järjestämistä ovat osoittaneet ulkomailla onnistumisia koulutuspolitiikassa (). Ohjelmointitaidot, modernit datatieteelliset työskentelytavat ja tiedonhallinnalliset perusteet ovat asiantuntijoiden mukaan kiistattomasti hyödyllisiä, mutta eriävät käsitykset keminformatiikan menetelmällisyyden syvyydestä kemian opetuksessa jäi epäselväksi tutkimuksessa. Tähän liittyy erityisesti eri haastateltavien ymmärrys tieteellisessä työskentelyssä kemian ja datatieteen painoarvosta: osa näki kemian keskiössä, kun toiset datatieteen ja koneoppimisen osaajan. Tämä osoittaa haastateltavien tieteellisen työskentelyn menetelmällisyyden

monitieteellistä luonnetta, mutta samalla voidaan tulosten perusteella todeta, että ratkaisuna voi toimia paremmat käytännöt saavutettavuudessa.

Bird ja Frey (2013) ilmaisivat, kuinka kemiallisen tiedon käsittelystä tutkimuksessa on tullut entistä tärkeämpää datatieteellisten mahdollisuuksien myötä. Haasteena on kuitenkin tutkielman mukaan datan niukkuus. Tämä voi olla syy, miksi keminformatiikan opetus ei ole yleistynyt Suomessa. Keminformatiikan soveltava luonne kuitenkin mahdollistaa sen käyttöönoton erilaisissa tieteissä. Toisaalta, koska tutkielmassa haastateltiin asiantuntijoita, on mielekästä teettää kyselyitä eri kemian toimialoilla selvitystä varten, käytetäänkö keminformatiikkaa niissä kuinka paljon ja millä tavalla. Tutkielman aineisto on melko suppea, koska informanttien määrä oli suppea. Siksi kyselyt ja haastattelun toistaminen samoilla tutkimuskysymyksillä ja laadullisella sisällönanalyysillä voi lisätä luotettavuutta. Tarvitaan enemmän tietoa keminformatiikan hyödyistä kemian tiedon hyödyntämisessä ja taidoista, mitä alalla vaaditaan. Tämä auttaisi opetuspolitiikan arvioinnissa ja mahdollisten kurssien valmistelussa.

Tämä tutkimus on uskottavasti ensimmäinen laatuaan keminformatiikasta ja sen mahdollisuuksista kemian opetukseen. Tulokset tukevat aiempia perusteluja keminformatiikan integroimisessa opetukseen, mutta suomalaisissa yliopistossa tarvitaan enemmän keskustelua ja yhteistyötä kemian opetuksessa hyödynnettävän keminformatiikan osaamistavoitteiden kartoittamiseen ja ymmärtämiseen. Ilman lisätutkimusta kemian opetuksessa, ei laajaa keminformatiikan alaa pystytä ymmärtämään ja soveltamaan kattavasti.

Tutkimuksella ei kuitenkaan voida aukottomasti vastata siihen, miten keminformatiikkaa tulisi tai voi hyödyntää, vaan tutkielma toimii katsauksena aiheeseen korostaen alan taustalla tapahtuvaa muutosta: kasvavaa dataa, monitieteellistyvää tiedettä, ohjelmointitaitojen kasvavaa tarvetta sekä kemian tieteellisen tiedon muutosta. Kemian opetuksen yliopisto-opetuksen kehittämiseen näiden ulottuvuuksien alustava tarkastelu teemoittelulla luo jatkotutkimuksille hyvän pohjan.

Lähteet

- ACS. (ei pvm.). Cheminformatics. American Chemical Society. Noudettu 28. heinäkuuta 2023, osoitteesta <https://www.acs.org/careers/chemical-sciences/fields/cheminformatics.html>
- Anaconda cloud. (2023). <https://anaconda.cloud/>.
- Ash, J. R., & Hughes-Oliver, J. M. (2022). Confidence bands and hypothesis tests for hit enrichment curves. *Journal of Cheminformatics*, 14(1), 50. <https://doi.org/10.1186/s13321-022-00629-0>
- Attene-Ramos, M. S., Austin, C. P., & Xia, M. (2014). High Throughput Screening. Teoksessa P. Wexler (Toim.), *Encyclopedia of Toxicology* (Third Edition) (ss. 916–917). Academic Press. <https://doi.org/10.1016/B978-0-12-386454-3.00209-8>
- Bachelor, M. (2015). Professional Development of Continuing Higher Education Unit Leaders: A Need for a Competency-Based Approach. *The Journal of Continuing Higher Education*, 63(3), 152–164. <https://doi.org/10.1080/07377363.2015.1085799>
- Banerjee, N., & Das, S. (2021). Lung Cancer Prediction in Deep Learning Perspective. Teoksessa *Computational Analysis and Deep Learning for Medical Care* (ss. 237–255). <https://doi.org/10.1002/9781119785750.ch10>
- Baykoucheva, S. (2015). Managing Scientific Information and Research Data. *Managing Scientific Information and Research Data*, 97–102. <https://doi.org/10.13016/M2SJ3F>
- Beisken, S., Meinl, T., Wiswedel, B., de Figueiredo, L. F., Berthold, M., & Steinbeck, C. (2013). KNIME-CDK: Workflow-driven cheminformatics. *BMC Bioinformatics*, 14(1), 257. <https://doi.org/10.1186/1471-2105-14-257>
- Bento, A. P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., Bellis, L. J., De Veij, M., & Leach, A. R. (2020). An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, 12(1), 51. <https://doi.org/10.1186/s13321-020-00456-1>
- Bharti, N., Leonard, M., & Singh, S. (2016). Review and Comparison of the Search Effectiveness and User Interface of Three Major Online Chemical Databases. *Journal of Chemical Education*, 93(5), 852–863. <https://doi.org/10.1021/acs.jchemed.5b00601>

- Bird, C. L., & Frey, J. G. (2013). Chemical information matters: An e-Research perspective on information and data sharing in the chemical sciences. *Chemical Society Reviews*, 42(16), 6754–6776.
<https://doi.org/10.1039/C3CS60050E>
- Brown, F. (1998). Chemoinformatics: What is it and how does it impact drug discovery. *Teoksessa Annual reports in medicinal chemistry* (Vsk. 33, ss. 375–384).
- Chan, P., Van Gerven, T., Dubois, J.-L., & Bernaerts, K. (2021). Virtual chemical laboratories: A systematic literature review of research, technologies and instructional design. *Computers and Education Open*, 2, 100053.
<https://doi.org/10.1016/j.caeo.2021.100053>
- Chen, H., Kogej, T., & Engkvist, O. (2018). Cheminformatics in Drug Discovery, an Industrial Perspective. *Molecular Informatics*, 37(9–10), 1800041.
<https://doi.org/10.1002/minf.201800041>
- Chen, I. J., Yin, D., & MacKerell, A. D. (2002). Combined ab initio/empirical approach for optimization of Lennard-Jones parameters for polar-neutral compounds. *Journal of Computational Chemistry*, 23(2), 199–213.
<https://doi.org/10.1002/jcc.1166>
- Clark, A. M., Williams, A. J., & Ekins, S. (2013). Cheminformatics workflows using mobile apps. *Chem-Bio Informatics Journal*, 13, 1–18.
<https://doi.org/10.1273/cbij.13.1>
- Clauset, A., Larremore, D. B., & Sinatra, R. (2017). Data-driven predictions in the science of science. *Science*, 355(6324), 477–480.
<https://doi.org/10.1126/science.aal4217>
- Diderich, R. (2010). Tools for category formation and read-across: Overview of the OECD (Q)SAR application toolbox. *Teoksessa In silico toxicology: Principles and applications* (ss. 390–412).
- Dong, X., Gilbert, K. E., Guha, R., Heiland, R., Kim, J., Pierce, M. E., Fox, G. C., & Wild, D. J. (2007). Web Service Infrastructure for Chemoinformatics. *Journal of Chemical Information and Modeling*, 47(4), 1303–1307.
<https://doi.org/10.1021/ci6004349>
- Dreiman, G. H. S., Bictash, M., Fish, P. V., Griffin, L., & Svensson, F. (2021). Changing the HTS Paradigm: AI-Driven Iterative Screening for Hit Finding.

- Special Issue: Hit Discovery Methodology, 26(2), 257–262.
<https://doi.org/10.1177/2472555220949495>
- Ebejer, J.-P., Fulle, S., Morris, G. M., & Finn, P. W. (2013). The emerging role of cloud computing in molecular modelling. *Journal of Molecular Graphics and Modelling*, 44, 177–187. <https://doi.org/10.1016/j.jmgm.2013.06.002>
- ECHA. (2008). Guidance on information requirements and chemical safety assessment. Chapter R.6: QSARs and grouping of chemicals.
- Faulon, J.-L., & Bender, A. (2010). *Handbook of Chemoinformatics Algorithms*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781420082999>
- Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, 359(6379), eaao0185. <https://doi.org/10.1126/science.aao0185>
- Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling*, 50(7), 1189–1204. <https://doi.org/10.1021/ci100176x>
- Fourches, D., Muratov, E., & Tropsha, A. (2015). Curation of chemogenomics data. *Nature Chemical Biology*, 11(8), 535–535. <https://doi.org/10.1038/nchembio.1881>
- Fowler, D. A., Arroyave, R., Ross, J., Malak, R., & Banerjee, S. (2017). Looking Outwards from the “Central Science”: An Interdisciplinary Perspective on Graduate Education in Materials Chemistry. Teoksessa R. Waterman & A. Feig (Toim.), *ACS Symposium Series* (Vsk. 1248, ss. 65–89). American Chemical Society. <https://doi.org/10.1021/bk-2017-1248.ch004>
- Gasteiger, J. (2006). The central role of chemoinformatics. *Selected Papers from the International Conference on Chemometrics and Bioinformatics in Asia*, 82(1), 200–209. <https://doi.org/10.1016/j.chemolab.2005.06.022>
- Gasteiger, J., & Engel, T. (2006). *Chemoinformatics: A Textbook*. John Wiley & Sons.
- Goble, C., & Roure, D. D. (2009). The Impact of Workflow Tools on Data-centric Research. *Teoksessa The Fourth Paradigm: Data-Intensive Scientific Discovery* (ss. 137–145).
- Gressling, T. (2021). *Data Science in Chemistry Artificial Intelligence, Big Data, Chemometrics and Quantum Computing with Jupyter*. De Gruyter.

- Griffen, E. J., Dossetter, A. G., Leach, A. G., & Montague, S. (2018). Can we accelerate medicinal chemistry by augmenting the chemist with Big Data and artificial intelligence? *Drug Discovery Today*, 23(7), 1373–1384.
<https://doi.org/10.1016/j.drudis.2018.03.011>
- Gromski, P. S., Henson, A. B., Granda, J. M., & Cronin, L. (2019). How to explore chemical space using algorithms and automation. *Nature Reviews Chemistry*, 3(2), 119–128. <https://doi.org/10.1038/s41570-018-0066-y>
- Guha, R., & Willighagen, E. (2020). Learning cheminformatics. *Journal of Cheminformatics*, 12(1), 4, s13321-019-0406-z.
<https://doi.org/10.1186/s13321-019-0406-z>
- Guha, R., Willighagen, E., Zdrazil, B., & Jeliazkova, N. (2021). What is the role of cheminformatics in a pandemic? *Journal of Cheminformatics*, 13(1), 16.
<https://doi.org/10.1186/s13321-021-00491-6>
- Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*, 194(4824), 178–180.
<https://doi.org/10.1038/194178b0>
- Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., & Hutchison, G. R. (2012). Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, 4(1), 17.
<https://doi.org/10.1186/1758-2946-4-17>
- Hardy, J. G., Sdepanian, S., Stowell, A. F., Aljohani, A. D., Allen, M. J., Anwar, A., Barton, D., Baum, J. V., Bird, D., Blaney, A., Brewster, L., Cheneler, D., Efremova, O., Entwistle, M., Esfahani, R. N., Firlak, M., Foito, A., Forciniti, L., Geissler, S. A., ... Wright, K. L. (2021). Potential for Chemistry in Multidisciplinary, Interdisciplinary, and Transdisciplinary Teaching Activities in Higher Education. *Journal of Chemical Education*, 98(4), 1124–1145.
<https://doi.org/10.1021/acs.jchemed.0c01363>
- Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S.-I., Greene, C. S., & Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 18(10), 1132–1135. <https://doi.org/10.1038/s41592-021-01256-7>

- Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., & Pletnev, I. (2013). InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, 5(1), 7. <https://doi.org/10.1186/1758-2946-5-7>
- Heller, S. R., McNaught, A., Pletnev, I., Stein, S., & Tchekhovskoi, D. (2015). InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7(1), 23. <https://doi.org/10.1186/s13321-015-0068-4>
- Hey, T., Tansley, S., Tolle, K., & Gray, J. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research. <https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/>
- Hirsjärvi, S., & Hurme, H. (2018). *Tutkimushaastattelu: Teemahaastattelun teoria ja käytäntö*. Gaudeamus Helsinki University Press.
- Hoyt, C. T., Zdrazil, B., Guha, R., Jeliaskova, N., Martinez-Mayorga, K., & Nittinger, E. (2023). Improving reproducibility and reusability in the Journal of Cheminformatics. *Journal of Cheminformatics*, 15(1), 62. <https://doi.org/10.1186/s13321-023-00730-y>
- Hu, Y., & Bajorath, J. (2017). Entering the 'big data' era in medicinal chemistry: Molecular promiscuity analysis revisited. *Future Science OA*, 3(2), FSO179. <https://doi.org/10.4155/fsoa-2017-0001>
- Jarada, T. N., Rokne, J. G., & Alhajj, R. (2020). A review of computational drug repositioning: Strategies, approaches, opportunities, challenges, and directions. *Journal of Cheminformatics*, 12(1), 46. <https://doi.org/10.1186/s13321-020-00450-7>
- Jeliaskova, N., & Kochev, N. (2011). AMBIT-SMARTS: Efficient Searching of Chemical Structures and Fragments. *Molecular Informatics*, 30(8), 707–720. <https://doi.org/10.1002/minf.201100028>
- Jimoyiannis, A. (2010). Designing and implementing an integrated technological pedagogical science knowledge framework for science teachers professional development. *Computers & Education*, 55(3), 1259–1269. <https://doi.org/10.1016/j.compedu.2010.05.022>
- Jirát, J., Čech, P., Znamenáček, J., Šimek, M., Škuta, C., Vaněk, T., Dibuszová, E., Nič, M., & Svozil, D. (2013). Developing and Implementing a Combined Chemistry and Informatics Curriculum for Undergraduate and Graduate

- Students in the Czech Republic. *Journal of Chemical Education*, 90(3), 315–319. <https://doi.org/10.1021/ed3001446>
- Kadtsyna, A. S., Chubarov, A. S., & Kadtsyn, E. D. (2022). Basic Cheminformatics Course for First-Year Chemistry Students. *Journal of Chemical Education*, 99(8), 2932–2942. <https://doi.org/10.1021/acs.jchemed.2c00175>
- Kaushik, M. (2014). A review of Innovative Chemical Drawing and Spectra Prediction Computer Software. *Mediterranean Journal of Chemistry*, 3(1), 759–766. <https://doi.org/10.13171/mjc.3.1.2014.04.04.16>
- Kim, S., Bucholtz, E. C., Briney, K., Cornell, A. P., Cuadros, J., Fulfer, K. D., Gupta, T., Hepler-Smith, E., Johnston, D. H., Lang, A. S. I. D., Larsen, D., Li, Y., McEwen, L. R., Morsch, L. A., Muzyka, J. L., & Belford, R. E. (2021). Teaching Cheminformatics through a Collaborative Intercollegiate Online Chemistry Course (OLCC). *Journal of Chemical Education*, 98(2), 416–425. <https://doi.org/10.1021/acs.jchemed.0c01035>
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research*, 47(D1), D1102–D1109. <https://doi.org/10.1093/nar/gky1033>
- Kim, S., Thiessen, P. A., Bolton, E. E., & Bryant, S. H. (2015). PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. *Nucleic Acids Research*, 43(W1), W605–W611. <https://doi.org/10.1093/nar/gkv396>
- Kuhn, S., Helmus, T., Lancashire, R. J., Murray-Rust, P., Rzepa, H. S., Steinbeck, C., & Willighagen, E. L. (2007). Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML Vocabulary for Spectral Data. *Journal of Chemical Information and Modeling*, 47(6), 2015–2034. <https://doi.org/10.1021/ci600531a>
- Kučák, D., Poppleton, E., Miao, H., Šulc, P., & Barišić, I. (2022). Unified Nanotechnology Format: One Way to Store Them All. *Molecules*, 27(1). <https://doi.org/10.3390/molecules27010063>
- Kuusisto-Niemi, S., & Saranto, K. (2009). Sosiaali- ja terveydenhuollon tiedonhallinta—Paradigma tieteenalan perustana. *Finnish Journal of eHealth and eWelfare*, 1(1), 19–23.
- Landrum, G. (2019). RDKit Documentation. <https://www.rdkit.org/docs/#>

- Landrum, G. (2021, lokakuuta 19). Intro to the RDKit [Powerpoint-esitelmä]. Advanced machine learning for Innovative Drug Discovery (AIDD), On-line seminar.
- Lo, Y.-C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8), 1538–1546. <https://doi.org/10.1016/j.drudis.2018.05.010>
- Louko, S., & Blomqvist, C. (2018). Report on the referencing of the Finnish National Qualifications Framework to the European Qualifications Framework and the Framework for Qualifications of the European Higher Education Area (s. 59). Opetushallitus.
- Luechtefeld, T., Rowlands, C., & Hartung, T. (2018). Big-data and machine learning to revamp computational toxicology and its use in risk assessment. *Toxicology Research*, 7(5), 732–744. <https://doi.org/10.1039/c8tx00051d>
- March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15(4), 251–266. [https://doi.org/10.1016/0167-9236\(94\)00041-2](https://doi.org/10.1016/0167-9236(94)00041-2)
- Mazanetz, M. P., Marmon, R. J., Reisser, C. B. T., & Morao, I. (2012). Drug discovery applications for KNIME: an open source data mining platform. *Curr Top Med Chem*, 12(18), 1965–1979. PubMed. <https://doi.org/10.2174/156802612804910331>
- McArdle, S., Endo, S., Aspuru-Guzik, A., Benjamin, S. C., & Yuan, X. (2020). Quantum computational chemistry. *Reviews of Modern Physics*, 92(1), 015003. <https://doi.org/10.1103/RevModPhys.92.015003>
- McEwen, L., & Li, Y. (2014). Academic librarians at play in the field of cheminformatics: Building the case for chemistry research data management. *Journal of Computer-Aided Molecular Design*, 28(10), 975–988. <https://doi.org/10.1007/s10822-014-9777-4>
- McGuire, R., Verhoeven, S., Vass, M., Vriend, G., de Esch, I. J. P., Lusher, S. J., Leurs, R., Ridder, L., Kooistra, A. J., Ritschel, T., & de Graaf, C. (2017). 3D-e-Chem-VM: Structural Cheminformatics Research Infrastructure in a Freely Available Virtual Machine. *Journal of Chemical Information and Modeling*, 57(2), 115–121. <https://doi.org/10.1021/acs.jcim.6b00686>

- McKinney, J. D., Richard, A., Waller, C., Newman, M. C., & Gerberick, F. (2000). The Practice of Structure Activity Relationships (SAR) in Toxicology. *Toxicological Sciences*, 56(1), 8–17. <https://doi.org/10.1093/toxsci/56.1.8>
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., ... Leach, A. R. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1), D930–D940. <https://doi.org/10.1093/nar/gky1075>
- Mishra, P., & Koehler, M. J. (2006). Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record*, 108(6), 1017–1054. <https://doi.org/10.1111/j.1467-9620.2006.00684.x>
- Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. *WIREs Computational Molecular Science*, 4(5), 468–481. <https://doi.org/10.1002/wcms.1183>
- Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
- Moriwaki, H., Tian, Y.-S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, 10(1), 4. <https://doi.org/10.1186/s13321-018-0258-y>
- Moshawih, S., Goh, H. P., Kifli, N., Idris, A. C., Yassin, H., Kotra, V., Goh, K. W., Liew, K. B., & Ming, L. C. (2022). Synergy between machine learning and natural products cheminformatics: Application to the lead discovery of anthraquinone derivatives. *Chemical Biology & Drug Design*, 100(2), 185–217. <https://doi.org/10.1111/cbdd.14062>
- Muegge, I., & Mukherjee, P. (2016). An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*, 11(2), 137–148. <https://doi.org/10.1517/17460441.2016.1117070>
- Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., Isayev, O., Curtarolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D. A., Agrafiotis, D., Cherkasov, A., & Tropsha, A. (2020). QSAR without borders. *Chem Soc Rev*, 49(11), 3525–3564. PubMed. <https://doi.org/10.1039/d0cs00098a>

- Nagy, Z. (2018). Artificial Intelligence and Machine Learning Fundamentals: Develop real-world applications powered by the latest AI advances. Packt Publishing Ltd.
- Nayarisseri, A., & Hood, E. A. (2019). Advancement in Microbial Cheminformatics. *Current Topics in Medicinal Chemistry*, 18(29), 2459–2461. <https://doi.org/10.2174/1568026619666181120121528>
- Nevalainen, E., & Toivonen, L. (2007). Syvemmät yhteydet oppimiseen ja opetukseen vievät eteenpäin tiedonhallinnan opetusta. *Signum*, 35(1). <https://journal.fi/signum/article/view/3415>
- O'Boyle, N. M. (2012). Towards a Universal SMILES representation—A standard method to generate canonical SMILES based on the InChI. *Journal of Cheminformatics*, 4(1), 22. <https://doi.org/10.1186/1758-2946-4-22>
- O'Boyle, N. M., Guha, R., Willighagen, E. L., Adams, S. E., Alvarsson, J., Bradley, J.-C., Filippov, I. V., Hanson, R. M., Hanwell, M. D., Hutchison, G. R., James, C. A., Jeliaskova, N., Lang, A. S., Langner, K. M., Lonie, D. C., Lowe, D. M., Pansanel, J., Pavlov, D., Spjuth, O., ... Murray-Rust, P. (2011). Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on. *Journal of Cheminformatics*, 3(1), 37. <https://doi.org/10.1186/1758-2946-3-37>
- Open Source Initiative. (2023, heinäkuuta 25). Open Source Initiative. <https://opensource.org/>
- Pence, H. E., & Williams, A. (2010). ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education*, 87(11), 1123–1124. <https://doi.org/10.1021/ed100697w>
- Pernaa, J., Takala, A., Ciftci, V., Hernández-Ramos, J., Cáceres-Jensen, L., & Rodríguez-Becerra, J. (2023). Open-Source Software Development in Cheminformatics: A Qualitative Analysis of Rationales. *Applied Sciences*, 13(17). <https://doi.org/10.3390/app13179516>
- Ponder, E. L., Freundlich, J. S., Sarker, M., & Ekins, S. (2014). Computational Models for Neglected Diseases: Gaps and Opportunities. *Pharmaceutical Research*, 31(2), 271–277. <https://doi.org/10.1007/s11095-013-1170-9>
- Pradeep, P., Povinelli, R. J., White, S., & Merrill, S. J. (2016). An ensemble model of QSAR tools for regulatory risk assessment. *Journal of Cheminformatics*, 8(1), 48. <https://doi.org/10.1186/s13321-016-0164-0>

- Prakash, N., & Gareja, D. A. (2010). Cheminformatics. *Journal of Proteomics & Bioinformatics*, 03(08), 249–252. <https://doi.org/10.4172/jpb.1000147>
- Randić, M. (1993). Novel molecular descriptor for structure—Property studies. *Chemical Physics Letters*, 211(4), 478–483. [https://doi.org/10.1016/0009-2614\(93\)87094-J](https://doi.org/10.1016/0009-2614(93)87094-J)
- Riniker, S., & Landrum, G. A. (2015). Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling*, 55(12), 2562–2574. <https://doi.org/10.1021/acs.jcim.5b00654>
- Rodríguez-Becerra, J., Cáceres-Jensen, L., Díaz, T., Druker, S., Bahamonde Padilla, V., Perna, J., & Aksela, M. (2020). Developing technological pedagogical science knowledge through educational computational chemistry: A case study of pre-service chemistry teachers' perceptions. *Chemistry Education Research and Practice*, 21(2), 638–654. <https://doi.org/10.1039/C9RP00273A>
- Ruddigkeit, L., van Deursen, R., Blum, L. C., & Reymond, J.-L. (2012). Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling*, 52(11), 2864–2875. <https://doi.org/10.1021/ci300415d>
- Saifi, I., Bhat, B. A., Hamdani, S. S., Bhat, U. Y., Lobato-Tapia, C. A., Mir, M. A., Dar, T. U. H., & Ganie, S. A. (2023). Artificial intelligence and cheminformatics tools: A contribution to the drug development and chemical science. *Journal of Biomolecular Structure and Dynamics*, 1–19. <https://doi.org/10.1080/07391102.2023.2234039>
- Sander, T., Freyss, J., von Korff, M., & Rufener, C. (2015). DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *Journal of Chemical Information and Modeling*, 55(2), 460–473. <https://doi.org/10.1021/ci500588j>
- Senyo, P. K., Addae, E., & Boateng, R. (2018). Cloud computing research: A review of research themes, frameworks, methods and future research directions. *International Journal of Information Management*, 38(1), 128–139. <https://doi.org/10.1016/j.ijinfomgt.2017.07.007>
- Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2), 4–14. <https://doi.org/10.3102/0013189X015002004>

- Shulman, L. S. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review*, 57(1), 1–23.
<https://doi.org/10.17763/haer.57.1.j463w79r56455411>
- Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2), 493–500. <https://doi.org/10.1021/ci025584y>
- Takeda, S., Hama, T., Hsu, H.-H., Yamane, T., Masuda, K., Piunova, V. A., Zubarev, D., Pitera, J., Sanders, D. P., & Nakano, D. (2020). AI-driven Inverse Design System for Organic Molecules. *ArXiv Preprint ArXiv*.
<https://doi.org/10.48550/arXiv.2001.09038>
- Tietoarkisto. (ei pvm.). Aineistohallinnan käsikirja. Tietoarkisto. Noudettu 18. elokuuta 2023, osoitteesta
<https://www.fsd.tuni.fi/fi/palvelut/aineistohallinta/>
- Tietosuoja-asetus. (2016). Asetus (EU) 2016/679, Euroopan parlamentin ja neuvoston asetukset, annettu 27 päivänä huhtikuuta 2016, luonnollisten henkilöiden suojelusta henkilötietojen käsittelyssä ja näiden tietojen vapaasta liikkuvuudesta sekä direktiivin 95/46/EY kumoamisesta (yleinen tietosuoja-asetus). Euroopan unionin virallinen lehti, L 119/1.
- Tietosuojalaki. (2018). Laki henkilötietojen käsittelystä poliisi- ja tulliviranomaisissa (1050/2018).
- Tolk, A. (2018). Simulation and Modeling as the Essence of Computational Science. *Proceedings of the 50th Computer Simulation Conference*, 1–8.
<https://doi.org/10.22360/SummerSim.2018.SCSC.014>
- Tosco, P., Stiefl, N., & Landrum, G. (2014). Bringing the MMFF force field to the RDKit: Implementation and validation. *Journal of Cheminformatics*, 6(1), 37.
<https://doi.org/10.1186/s13321-014-0037-3>
- Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, 29(6–7), 476–488.
<https://doi.org/10.1002/minf.201000061>
- Tuomi, J., & Sarajärvi, A. (2018). Laadullinen tutkimus ja sisällönanalyysi. Tammi.
- Walters, W. P. (2019). Virtual Chemical Libraries. *Journal of Medicinal Chemistry*, 62(3), 1116–1124. <https://doi.org/10.1021/acs.jmedchem.8b01048>

- Warr, W. A. (2012). Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of Computer-Aided Molecular Design*, 26(7), 801–804.
<https://doi.org/10.1007/s10822-012-9577-7>
- Wegner, J. K., Sterling, A., Guha, R., Bender, A., Faulon, J.-L., Hastings, J., O'Boyle, N., Overington, J., Van Vlijmen, H., & Willighagen, E. (2012). *Cheminformatics. Communications of the ACM*, 55(11), 65–75.
<https://doi.org/10.1145/2366316.2366334>
- Weininger, D., Weininger, A., & Weininger, J. L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, 29(2), 97–101. <https://doi.org/10.1021/ci00062a008>
- Wild, D. (2021). Johdatus keminformatiikkaan (J. Pernaa, Käänt.). Edumendo.
<https://doi.org/10.31885/9789529443918>
- Wild, D. J. (2009). Grand challenges for cheminformatics. *Journal of Cheminformatics*, 1(1), 1. <https://doi.org/10.1186/1758-2946-1-1>
- Wild, D. J. (2013). Cheminformatics for the masses: A chance to increase educational opportunities for the next generation of cheminformaticians. *Journal of Cheminformatics*, 5(1), 32, 1758-2946-5–32. <https://doi.org/10.1186/1758-2946-5-32>
- Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., Torrance, G., Evelo, C. T., Guha, R., & Steinbeck, C. (2017). The Chemistry Development Kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics*, 9(1), 33. <https://doi.org/10.1186/s13321-017-0220-4>
- Wishart, D. S. (2016). Introduction to Cheminformatics. *Current Protocols in Bioinformatics*, 53(1), 14.1.1-14.1.21.
<https://doi.org/10.1002/0471250953.bi1401s53>
- Wolfender, J.-L., Nuzillard, J.-M., van der Hooft, J. J. J., Renault, J.-H., & Bertrand, S. (2019). Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography–High-Resolution Tandem Mass Spectrometry and NMR Profiling, in *Silico Databases, and Chemometrics. Analytical Chemistry*, 91(1), 704–742.
<https://doi.org/10.1021/acs.analchem.8b05112>

- Xiouras, C., Cameli, F., Quilló, G. L., Kavousanakis, M. E., Vlachos, D. G., & Stefanidis, G. D. (2022). Applications of Artificial Intelligence and Machine Learning Algorithms to Crystallization. *Chemical Reviews*, 122(15), 13006–13042. <https://doi.org/10.1021/acs.chemrev.2c00141>
- Yerin, A., McNaught, A., & Heller, S. (2013). Current Status and Future Development in Relation to IUPAC Activities. *Chemistry International -- Newsmagazine for IUPAC*, 35(6), 12–15. <https://doi.org/10.1515/ci.2013.35.6.12>
- Young, D., Martin, T., Venkatapathy, R., & Harten, P. (2008). Are the Chemical Structures in Your QSAR Correct? *QSAR & Combinatorial Science*, 27(11–12), 1337–1345. <https://doi.org/10.1002/qsar.200810084>
- Zdrazil, B., & Guha, R. (2022). Diversifying cheminformatics. *Journal of Cheminformatics*, 14(1), 25. <https://doi.org/10.1186/s13321-022-00597-5>
- Zin, P. P. K., Williams, G. J., & Ekins, S. (2020). Cheminformatics Analysis and Modeling with MacrolactoneDB. *Scientific Reports*, 10(1), 6284. <https://doi.org/10.1038/s41598-020-63192-4>
- Zohuri, B., & Sadeh, S. (2020). Artificial intelligence driven by machine learning and deep learning. Nova Science Publishers.

Liitteet

Liite 1. Sisällönanalyysin pelkistetyt ilmaisut, luokat ja teemat.

Pelkistetty ilmaisu	Alaluokka	Yläluokka			Pääluokka	Teema
		SK	TK	TSK		
Keminformatiikka biologisissa tieteissä	Monitieteellisyys	X			Keminformatiikan muuttuva luonne	Tieteellisen tiedon monitieteellisyys
Keminformatiikka tietojenkäsittelytieteen ja kemian leikkauksena				X		
Monitieteelliset menetelmät				X		
Automatisoitu kokeellisuus		X	X	X		
Keminformatiikka kasvaa alana datan kasvun mukana	Datan kartoitus ja mallintaminen	X	X	X	Keminformatiikan muuttuva luonne	Keminformatiikan datakeskeinen luonne
Molekyylien ohjelmallinen muokkaaminen		X	X	X		
Kemian tiedon tiedonhallinta; datan seulonta, järjestely, aineistonhallinta ym.		X	X	X		
Datatieteellisten mallien kasvava rooli menetelmällisyydessä		X	X	X		
Datan niukkuus	Tiedon saavutettavuus	X			Keminformatiikan menetelmällisyys	Ohjelmointitaidot
Algoritmien optimointi ja hyödyntäminen	Kemian simulointi		X			
Python-koodin merkitys alalla		X	X			
Menetelmän suunnittelu		X	X	X		Keminformatiikan hyödyntäminen
Mallien laajennettavuus		X				
Fysikaalisen kemian merkitys		X				
Luova ongelmanratkaisu		X				
Hypoteesien testaaminen		X				
Kemian simuloiminen		X	X	X		

Metatason ajattelun merkitys		X				
Yhteisöllinen ja avoin tiede	Tiedon saavutettavuus	X				