TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Lisette Pihor

Tunnuste valik närvivõrguga akuutse mürgisuse prognoosimisel

Bakalaureusetöö (9 EAP)

Juhendajad:
Sulev Sild, Uko Maran

Tartu 2025

Tunnuste valik närvivõrguga akuutse mürgisuse prognoosimisel

Lühikokkuvõte:

Bakalaureusetöö raames vaadeldi ja võrreldi tunnuste valikualgoritme tehislike närvivõrkude mudelite rakendamisel kemikaalide mürgisuse ($pIGC_{50}$) hindamiseks. Tunnuste valikualgoritmidest koostati ülevaade ja võrreldi nelja meetodit tehislike närvivõrkude mudelite koostamisel. Parima tulemuse andis molekulaartunnuste valik juhumetsa lähenemisega, mille R^2 väärtus treeningandmetel oli 0.9534 ja testandmetel 0.8128. Tulemused olid kooskõlas varasemate leidudega, kuid pakkus lahenduse suuremale andmehulgale vähemate molekulaartunnustega kui varasemalt, mis lõi eeldused kasutatud masinõppemeetodite tulemuste interpreteerimiseks.

Võtmesõnad: kvantitatiivne struktuurianalüüs, närvivõrk, akuutne mürgisus, *Tetrahymena pyroformis*, molekulaartunnused

CERCS: B740 Farmakoloogia, farmakognoosia, farmaatsia, toksikoloogia, P176 Tehisintellekt

Feature Selection for Evaluating Acute Toxicity with Neural Networks

Abstract:

As a part of this bachelor thesis, the application of feature selection methods for evaluating toxicity (pIGC50) of chemicals using artificial neural networks were examined. An overview of the feature selection methods was compiled, and four different methods were analysed while building neural network models. The best results were achieved with a random forest based selection method. The best model had the R^2 value of 0.9534 for the training and 0.8128 for the test set. The results were consistent with previous findings and provided a solution for a larger dataset using fewer molecular descriptors than before, therefor creating an opportunity for interpreting the machine learning results.

Keywords: QSAR, neural network, acute toxicity, *Tetrahymena pyroformis*, molecular descriptors

CERCS: B740 Pharmacological sciences, pharmacognosy, pharmacy, toxicology, P176 Artificial intelligence

Sisukord

1. Sissejuhatus

Keemiliste ühendite akuutne mürgisus on oluline eelkõige ühendi keskkonnaohutuse määratlemisel. See valdkond on pidevalt arenev ja kulutuste vähendamiseks hakati mürgisuse ennustamiseks otsima järjest odavamaid viise. Vaadeldav organism *Tetrahymena pyroformis* on selles valdkonnas oluline, sest ta on laialdaselt leitav, lihtne kasvatada ja tundlik keskkonna muutuste suhtes (Maurya & Pandey, 2020). Selle organismi põhiste meetodite arendamisesse on suurima panuse andnud Terry W. Schultz ja tema uurimisrühm. Masinõppe kasutamist mürgisuse tagapõhjade selgitamiseks kasutab ka juba tema oma publikatsioonides (Schultz, 1997).

Masinõppe väärtus keemias tuleneb mudelite võimekusest leida andmetes seoseid. Kvantitatiivselt molekuli aktiivsuse prognoosimiseks (QSAR, ingl *quantitative structure–activity relationship*), on seda kasutatud juba aastast 1963 (Aoyama jt, 1990). Närvivõrkude kui masinõppe meetodi sobivuse eelmainitud ülesandeks on oma töös tõestanud näiteks Kahn jt (2007) ja Xu jt (1994). Tunnuste valik on seejuures mudeli koostamisel oluline, sest sedasi on mudelid täpsemad, vähem üle sobitatud ja lihtsamini kirjeldatavad (Khan & Roy, 2018). Ülevaate tunnuste valiku algoritmidest QSAR analüüsil annab oma töös Shahlaei (2013).

Käesoleva töö eesmärgiks oli luua mudel kemikaalide akuutse toksilisuse ennustamiseks ja leida, kuidas selle korral tunnuste valikut teostada. Samuti uurida, kuidas mürgisus sõltub kemikaali omadustest. Mürgisust on püüdnud mudeldada mitmed uurimisgrupid. Mitmetes töödes on välja tulnud, et närvivõrgud sobivad hästi mürgisuse modelleerimismeetodiks. Hiliseim neist on näiteks Pushkarova jt (2022) koostatud võrdlus fenoolide mürgisuse ennustamisel.

Eksperimentaalsel mürgisuse mõõtmisel on probleemiks selle protseduuri kulukus. Varasemalt on seda üritatud lahendada arvutusmudelitega. Käesolevas töös püstitati hüpotees, et võimalik on saada varasemate mudelitega vähemalt samaväärne tulemus ka suurema andmehulga ja väiksema arvu molekulaartunnuste korral. Samuti loodeti näha trendi, et tunnuste vähendamisel mudeli ennustusvõime kasvab ja mudelist on kergem aru saada.

Töö koosneb kirjanduse ülevaatest, metoodika kirjeldusest ja tulemustest. Kirjanduse ülevaade kirjeldab täpsemalt masinõppe kasutust keemias nii üldiselt kui ka närvivõrkude abil. Lisaks sisaldab see teavet, kuidas *Tetrahymena pyroformis*t on varem kasutatud akuutse mürgisuse uurimiseks.

Metoodika peatükis kirjeldatakse täpsemalt kasutatud andmeid ja kuidas neid on võimalik esitada. Tunnuste valiku algoritmid valiti Teixeira jt (2013), Belfield jt (2023) Rahangdale ja Raut (2019) ning De Sá (2019) tööde põhjal. Algoritmid püüti valida võimalikult erinevad, kuid siiski ka võrreldavad. Viimaks kirjeldatakse metoodika peatükis ka närvivõrgu ehitamise protsessi.

Tulemustes tuuakse välja iga algoritmi korral leitud parimad mudelid ja võrreldakse neid teiste meetoditega saadud tulemustega. Viimaseks tuuakse välja parim mudel ja võrreldakse seda ning selles kasutatud tunnuseid kirjandusest leitavatega. Lisadesse kuuluvad koodi sisaldav andmehoidla ja parima mudeli tunnuste kirjeldused.

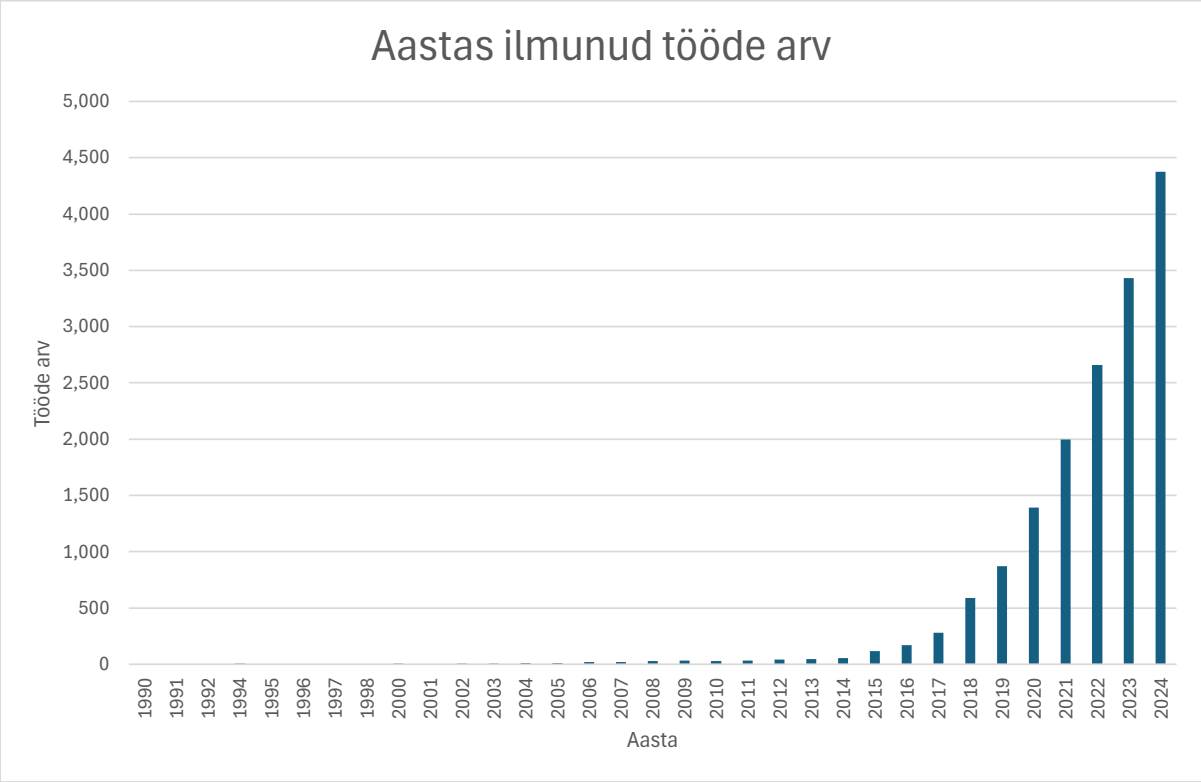
2. Kirjanduse ülevaade

Selles peatükis selgitatakse täpsemalt käesoleva töö tausta ja antakse ülevaade samas valdkonnas varasemalt tehtud uurimistöödest.

2.1 Masinõpe keemias

Masinõpe on kujunenud iseseisvaks distsipliiniks (Barbierato jt, 2025), mis suudab parandada teatud ülesande sooritust läbi andmetest kogutud teadmuse (Goodfellow jt, 2016).

Kirjanduses on selgelt märgatav trend, et alates aastast 2016 on teadusartiklite arv hüppeliselt kasvanud ja see on siiani järjest suurenemas. Erinevates andmebaasides otsingu tulemused natuke erinevad üksteisest, kuid trendid on sarnased. Eeltoodud kasvutrend on näha näiteks Ameerika Keemiaühingu (*American Chemical Society*) poolt loodud andmebaasis SciFinder[1] keemia ja masinõppe alaseid töid otsides (vt joonis 1).



Joonis 1. Andmebaasi SciFinder otsing „machine learning AND chemistry“

Masinõppe meetodid on keemiaalaselt eriti kasulikud, sest need oskavad leida andmetest mustreid ja nende abil saab leida sõltuvusi uuritava omaduse ning keemilise struktuuri vahel.

[1] https://scifinder-n.cas.org/

Saadud mudelid aitavad näiteks nii uute kemikaalide disainimisel (Tkatchenko, 2020) ja reaktsioonide uurimisel (Meuwly, 2021).

2.1.1 Kvantitatiivne struktuurianalüüs

Molekulide nii füüsikaliste, keemiliste, kui ka bioloogiliste omaduste kvantitatiivsele ennustamisele struktuuri põhjal viidatakse keemias lühendiga QSAR, eesti keeles kvantitatiivne struktuur-aktiivsus sõltuvus (Gini, 2018). Keemilise ühendi struktuuri kirjeldamiseks arvutatakse tavaliselt selle keemilise ühendi valemi või muu molekulaarse esitluse põhjal erinevaid molekulaartunnuseid. Samas viimasel ajal on välja mõeldud ka närvivõrke kasutavaid meetodeid, mis suudavad ennustada ka näiteks otse molekuli struktuuri pildi põhjal (Tan jt, 2023).

QSAR analüüsil on mitmeid erinevaid rakendusi, millest üks olulisemaid on ravimite disainimine. Tänapäeval on QSAR tähtsal kohal ravimite laadsete ühendite identifitseerimisel, optimeerimisel ja nende bioloogilise mõju määratlemisel, kuid ka keskkonnariskide hindamisel ja ka tööstuslike protsesside optimeerimisel (Khan & Roy, 2018).

Aoyama jt (1990) sõnul pärinevad esimesed QSAR meetodi kirjeldused juba aastast 1963, millal selle kallal töötasid Hansch ja tema kolleegid (Aoyama jt, 1990: Hansch jt, 1963). Esimesed tööd keskendusid peamiselt lineaarsetele ennustusmudelitele (Khan & Roy, 2018). Peale metoodika kasutuselevõttu on QSAR uuringuid koostatud mitmete keemiliste ühendite kohta ja erinevate masinõppe meetoditega kaasaarvatud närvivõrkudega. Selles uurimistöös kasutatakse masinõpet, et leida kemikaali struktuuri põhjal selle akuutset mürgisust, mis kuulub samuti molekuli keemiliste omaduste alla.

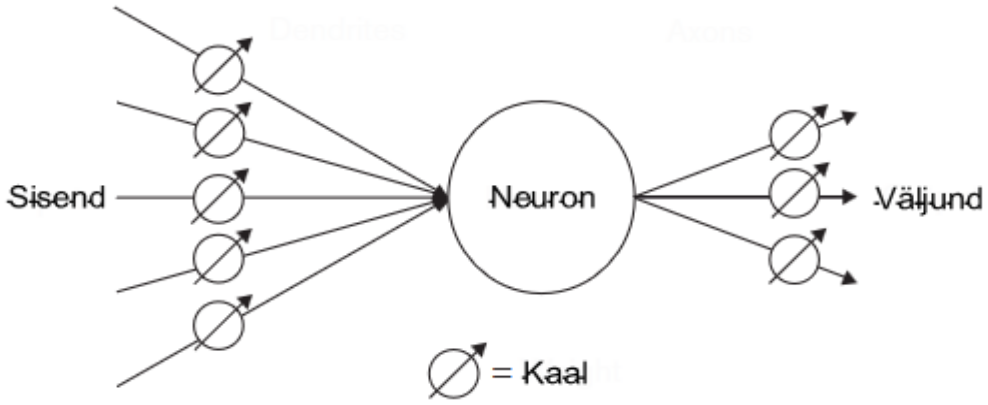
2.1.2 Närvivõrgud

Närvivõrke jaotatakse mitmete erinevate omaduste põhjal ning jaotiste puhul on võimalik tekitada veel mitmeid alamjaotuseid seega sõltub täpne grupeering tihtipeale konkreetsest tööst ja autorist. Peamised kasutusel olevad närvivõrkude tüübid Sharkawy (2020) sõnul on ühe- ja mitmekihilne pärilevivõrk, rekurrentne võrk, radiaalse baasfunktsiooniga võrk, üldine regressiooniline närvivõrk, tõenäosuslik närvivõrk ja täiendav närvivõrk.

Tetrahymena pyroformis organismile mürgisuse ennustamisel suurtes andmekomplektides on närvivõrgud näidanud oma head sooritust võrreldes multilineaarsete mudelitega nii Kahn jt (2007) töös kui ka Xu jt (1994) välja pakutud lahenduses. Närvivõrgud koos teiste mittelineaarsete meetoditega on oma võimekust näidanud näiteks aromaatsete ühendite

mürgisuse ennustamisel (Ren, 2003). Samuti on uuritud fenoolide mürgisuse ennustamist (Pushkarova jt, 2022), mille käigus võrreldi erinevaid masinõppe meetodeid. Nende töös osutusid parimateks pärilevi- ja kaskaadvõrgud, mille mõlema puhul antakse infot ainult edasi. Ka uurimuses, mis käsitles parameetrite valikut erinevate masinõppe meetodite jaoks, olid parimad tulemused saadud närvivõrguga (Belfield jt, 2022). Varasematele töödele põhinedes ning vähese informatsiooni tõttu, kuidas nende korral süstemaatiliselt molekulaartunnuste valikut läbi viia, otsustati ka käesolevas töös keskenduda närvivõrkudele.

Tehisnärvivõrkude ehitus on inspireeritud ajust. Grossi & Buscema (2007) sõnul koosnevad närvivõrgud põhiliselt neuronitest, neist igal ühel on oma sisend ja väljund, mille abil see arvutab ja suhtleb teiste neuronitega (vt joonis 2).



Joonis 2. Üksiku neuroni ehitus (Grossi & Buscema, 2007)

Neuron arvutab antud sisendi põhjal väljundi järgneva valemi 1 järgi, kus x on sisend, y on väljund, θ on mudeli parameetrite hulk, f on funktsioon, mille abil leitakse sisendist väljund ja ε on mudeli juhuslik viga (Grossi & Buscema, 2007).

$$y = f(\theta, x) + \varepsilon \quad (1)$$

Grossi & Buscema (2007) selgitavad, et igat neuronite vahelist ühendust iseloomustab teatud kaal, mis väljendab, kui palju seda ühendust arvesse võetakse. Treenimise käigus neid väärtuseid muudetakse. Kaalusid kohendatakse treenimisel tagasilevi algoritmiga, mis kõige lihtsamalt töötab kõige järsema laskumise meetodil (Svozil jt, 1997).

$$\omega_{ij}{}^{(k+1)} = \omega_{ij}{}^{k} - \lambda \left(\frac{\delta E}{\delta \omega_{ij}} \right)^{k} \quad (2)$$

Välja toodud valemis 2 arvutatakse uus kaal lahutades eelnevast kaalust õpisammu ja tõusu korrutis. Tõus arvutatakse diferentseerides eesmärkfunktsiooni. Sisuliselt annab tõus kaalu muutusele suuna ja õpisamm järgmise kaalu kauguse eelmisest (Svozil jt, 1997). Uuemad optimeerivad algoritmid suudavad õpisammu muuta ning seeläbi leida kiiremini ja parema koonduvusega parimad kaalud. Üheks selliseks populaarseks algoritmiks on näiteks Adam (Bock & Weis, 2019). Näiteks toodud optimeerimisalgoritmi kasutatakse ka käesolevas uurimuses Bock ja Weis (2019) välja toodud universaalsuse, kiiruse ja hea lokaalse koonduvuse tõttu.

Närvivõrkude eelised teiste masinõppe meetodite ees tulevad välja andmete hulga suurenemisel. Ng jt (2018) tõid välja, et mida suurem on närvivõrk ja mida rohkem on andmeid, seda täpsemaks saab närvivõrke treenida võrreldes traditsiooniliste masinõppe meetoditega. Nende sõnul pole veel leitud universaalselt rakendatavaid piire, millest alates mudel on liiga suur või andmeid liiga palju ette antud. Need sõltuvad paljuski konkreetsest andmekomplektist ja süsteemi võimsusest ning ilmnevad tavaliselt mudeli treenimise käigus.

2.2 Tetrahymena pyroformis

Akuutset mürgisust on algselt uuritud kalade peal (Seward jt, 2002). Sewardi jt (2002) sõnul on see aga kulukas ettevõtmine, mille tõttu hakati asendama neid teiste, nn alamorganismidega, mis on kaladega piisavalt tugevalt seotud. Selleks kujunes akuutse mürgisuse jaoks *Tetrahymena pyroformis*.

QSAR analüüsil püütakse üldjuhul mudelite abil prognoosida antud keemilise ühendi omaduse olemasolu või selle väärtust. Käesolevas töös on uuritavaks ja ennustatavaks eksperimentaalseks omaduseks akuutne mürgisus, mis on varasemalt kogutud ja kureeritud (Ruusmann & Maran, 2013).

2.2.1 Akuutne mürgisus

Akuutne mürgisus väljendab aine ühekordset lühiajalist mõju veekeskkonnale (Elhag, 2023). *Tetrahymena pyroformis*e puhul määratakse mürgisust populatsiooni kasvu aeglustumise või letaalsuse järgi, kuid surnud ja elus rakkude raske eraldamise tõttu on esimene neist soositud (Schultz, 1997). Schultz (1997) kirjeldab, et populatsiooni kasvu vähenemist mõõdetakse kindla aja, temperatuuri ja muude varasemalt paika pandud tingimuste juures. Ta mainib ka, et tulemused viiakse lõpuks sellisele kujule, et saadud tulemus väljendaks kontsentratsiooni, mis põhjustab viiekümneprotsendilise populatsiooni kasvu vähenemise. Seda märgitakse

kirjanduses lühendiga IGC_{50} ja mõõdetakse üldiselt millimoolides. Mõõdikuna kasutatakse tihti negatiivset logaritm eelmainitud kujust.

2.2.2 Tetrahymena pyroformis mudelorganismina

Tetrahymena pyroformis on ripsloom, kes kuulub *Oligohymenophorea* klassi, *Hymenostomia* alamklassi, *Hymenostomatida* hõimkonda ja *Tetrahymenina* alamhõimkonda. Ta on ainurakne, kes elab mageveekogudes (Shultz, 1997). Ainuraksed on eükarüoodid ja neid leidub peaaegu kõikjal (Sauvant jt, 1999). Sauvant jt (1999) toovad välja, et ainuraksete seas on ripsloomad, mille alla kuulub ka *Tetrahymena pyroformis*, kõige tihedamini uuritud organismid. *Tetrahymena pyroformis* on ainuraksete seas üks populaarsemaid mudelorganisme (Sauvant jt, 1999). Tema eelised teiste uuritavate objektide ees on see, et lühikese aja vältel on võimalik jälgida palju iseseisvaid organisme, millel on nii eukarüootide kui ka hulkraksete omadused, mis suurendab tema põhjal tehtud järelduste adekvaatsust (Schultz, 1997). Samuti on teda Schultzi (1997) sõnul ökonoomne laboris kasvatada.

Seost kalade ja Tetrahymena pyroformise vahel on varasemalt palju uuritud ning leitud, et mürgisuse mõjud näiteks gupidele (*Poecilia reticulata*) ja *Tetrahymena pyroformis*ele on tugevalt seotud (Seward jt, 2002). *Tetrahymena pyroformis*e populatsiooni uurimise on standardiseerinud Schultz (1997). Ta on ka selle protsessi valideerinud ning on üks suurimaid panustajaid *Tetrahymena pyroformis*e uuringutesse (Seward jt, 2002).

2.2.3 Närvivõrgud *tetrahymena pyriformis*e uurimisel

Uurimisgruppe, kes on tegelenud mürgisuse uurimisega läbi masinõppe mudelite *Tetrahymena pyriformis*e näitel, on mitmeid. Tihti erinevad uuringud kasutatud andmete koguse ja iseloomu tõttu. Kuna käesolevas töös on andmete hulk võrreldes varasemate uuringutega pigem suur, siis vaadatakse võrdluseks sarnase andmehulgaga tegelenud teadlaste töid. Huvi korral saab ülevaate teistsuguste andmehulkade korral parimatest kasutatud meetoditest Ghosh jt (2024) töö sissejuhatusest.

Tabel 1. Varasemad närvivõrgu mudelid

Mudel	Test R2	Test RMSE	Tunnuseid	Andmeid	Autorid
pärilevinärvivõrk	0.948	0.059	66	824	(Yu, 2020: Yu jt, 2010).
üldine regressiooniline närvivõrk	0.80	0.41	4	1164	(Yu, 2020)
assiotsiatiivne närvivõrk	0.87	–	58	1093	(Zhu jt, 2008)
süvanärvivõrk	0.806 (ristvalideerides)	–	447	1994	(Belfield jt, 2023)

Varasema kirjandusest leitud täpseim mudel oli pärilevinärvivõrk, mille koostasid Yu jt (Yu, 2020: Yu jt, 2010). Nad kasutasid enda välja arendatud meetodit andmete eelgrupeerimiseks, mille tulemuseks saadi mudel, mille R^2 testandmetel oli 0.984. Nende andmekomplekt sisaldas 824 erinevat keemilist ühendit. Mudeli treenimiseks kasutati 66 molekulaartunnust. Veel on kasutatud *Tetrahymena pyroformise* andmete põhjal mürgisuse ennustamiseks üldist regressioonilist närvivõrku (ingl *general regression neural network*), mille R^2 testandmetel oli 0.80 (Yu, 2020). Närvivõrk näitas ka head tulemust Zhu jt (2008) töös, kus 58 tunnuse pealt ennustades saadi mudel, mille R^2 testandmetel oli 0.87. Käesoleva tööga sama allika põhjal koostatud andmekomplekti kasutanud Belfield jt (2023) said parima tulemuse kasutades süvanärvivõrku. Täpsem närvivõrgu arhitektuur ja parameetrid on neil enda töös välja toodud. Kokkuvõtvalt on väljatoodud varasemast kirjandust pärit närvivõrgud koondatud lõigu all nähtavasse tabelisse (vt tabel 1). Varasemalt saadud tulemused on üsnagi head, kuid neis on kasutatud kas palju tunnuseid või on tulemused saadud käesolevast tööst väiksemal andmehulgal.

3. Metoodika

Metoodika välja töötamisel võeti arvesse eelnevaid töid ja andmete iseloomu. Samuti mängis rolli ka käesoleva uurimistöö raamistik ja maht.

3.1 Andmed

Käesolevas uurimistöös kasutatud andmed pärinevad uurimusest, milles Ruusmann ja Maran (2013), koostasid teaduskirjanduses toodud andmete kureerimise töövoo, et luua kasutatav andmebaas. See realiseeriti tarkvaralahendusena ja valideeriti *Tetrahymena pyroformise* mürgisuse uuringute peal. Selle tulemusena saadi andmekomplekt, mida pandi kokku 86 erineva teadustöö mõõtmistulemustest. Tööd pärinevad vahemikust 1980 - 2011 ning suure osa neist on autori või kaasautorina kirjutanud Theodore William Schultz. Lõplik käesolevas töös kasutatud andmekomplekt sisaldas 2027 keemilise ühendi mürgisuse mõõtmiseid, millest 1994 olid kasutatavad molekulaartunnuste arvutamisel.

Vaadeldud kemikaalid on üldjuhul orgaanilised ühendid, millel on tõenäosus sattuda veekogudesse ja sealset elukeskkonda kahjustada. Orgaanilised ühendid on kasutusel peaaegu igal pool: nii meditsiinis, toidutehnoloogias, põllumajanduses, kosmeetikas jne (Kumar jt, 2022). Keskkonda saastavaid aineid on palju ja uuritud ei ole neist kindlasti kõiki. Põhiline suund mürgisuse uuringutel on aromaatsed ühendid. Kumari jt (2022) sõnul on aromaatsed ühendid nagu benseen ja selle derivaadid põhiliselt kasutusel kütuste tootmisel. Samuti leidub nende sõnul benseeni derivaate ka värvides, lakkides, lahustites ja paljudes muudes kohtades. Aromaatsed ühendid võivad inimestel põhjustada aneemiat, leukeemiat, vähki ja immuunsüsteemi häireid (Kumar jt, 2022). Vees lahustuvad ained tekitavad suuri probleeme veekeskkonna elustikule (Mollaei jt, 2009). Samas pole vaadeldavate kemikaalide mõjuala kindlasti piiratud ainult veekeskkonnaga (Kumar jt, 2022). Keshavarz jt (2021) toovad välja, et põhilised uuritavad aromaatsed ühendid on nitraadid, benseen, nitrobenseen ja fenool ning nende derivaadid, aromaatsed aldehüüdid jms.

3.2 Molekulaarstruktuuride kirjeldamine

Molekulide esitamine masinloetaval kujul ei ole lihtne. Esitused peaksid olema kanoonilised ja unikaalsed ehk ühele molekulile vastab täpselt üks esitus (Wigh jt, 2022). Nad toovad välja järgmised põhilised murekohad esituse koostamisel: tsüklid, ebastandardsed sidemed, anorgaanilised ühendid ja sümmeetria. Nende ebakorrektne või ebapiisav esitus võib viia

valede otsingutulemusteni andmebaasides. Lahenduseks on välja pakutud mitmeid erinevaid esitamise meetodeid.

Molekulaarstruktuuride esitusviisid jagunevad tekstiks, tabeliteks, molekulaartunnuste põhiseks ja arvuti poolt õpitud, näiteks närvivõrgu abil graafi lugemine, viisideks (Wigh jt, 2022). Osade puhul neist ei ole molekulaarstruktuuride esitlustele ette määratud tingimused täidetud. Näiteks molekulaartunnuste kaudu molekuli kirjeldamise korral pole unikaalsuse tingimus alati täidetud, kuid see ei takista nende kasutamist modelleerimisel. Siinses töös on andmed antud tekstina seega selgitatakse seda ka pikemalt lahti. Variante esituseks tekstina on mitmeid, mille alla kuuluvad ka ainete nimetused ja molekulvõrrandid. Keerulisemates võib välja tuua sageli kasutatavad keemilise struktuuri joontähistused SMILES (ingl *simplified molecular-input line-entry system)*, WSL (ingl *Wiswesser line notation*) ning lisaks mitmed andmebaaside koodidel põhinevad esitusviisid (Wigh jt, 2022). Kõigil neil on oma eeskirjad, mille järgi molekulaarstruktuuride esitusi kirjutatakse. Kasutusel olev andmestik kasutab keemilist nimetust ning InChi, SMILES ja CAS RN süsteeme. Wigh jt (2022) selgituste põhjal kirjeldatakse käesolevas töös kasutatavaid meetodeid täpsemalt.

SMILES korral kirjutatakse molekul välja ASCII tähestikuga. Aatomid on tähistatud tähtedega, =,# ja $ märgivad erinevaid sidemeid, tsükleid väljendatakse numbritega, lahknemisi sulgudega ja aromaatsust kas väiketähtedega või vastavate sidemetüüpidega (nn Kekule esitus). InChi on keerulisem ja vähem loetavam. See koosneb kihtidest. Põhikihil on märgitud molekulvõrrand, sidemed ja vesinikud. Laengu kiht kogu süsteemi laengut ja näitab vajadusel protoneeritud aatomeid. Lisaks on võimalik kasutada kihte kirjeldamaks stereokeemiat, isotoopi, tautomeeri ja taastatud sidemeid. CAS RN ei ole üldse loetav ja sisaldab endas CAS andmebaasis ühendile omistatud unikaalset numbrit (ehk identifikaatorit. Konkreetse näite jaoks vaata tabel 2).

Tabel 2. Ditsükloveriin hüdrokloriidi erinevad esitusviisid (Wigh jt 2022)

Esitusviis	Molekuli kirjeldus
Molekulvalem	$C_{19}H_{36}ClNO_2$
SMILES	CCN(CC)CCOC(=O)C1(CCCCC1)C2CCCCC2.Cl
InChi	InChI = 1S/C19H35NO2.ClH/c1-3-20(4-2)15-16-22-18(21)19(13-9-6-10-14-19)17-11-7-5-8-12-17;/h17H,3-16H2,1-2H3;1H
WSL	L6TJA-AL6TJAVO2N2&2&GH

Molekulaartunnuste arvutamiseks on mitmeid tarkvaralisi lahendusi. Molekulaartunnuseid on võimalik arvutada tuhandeid ja neid vaadatakse lähemalt, kui on selgunud parimates mudelites kasutatavad molekulaartunnused. Käesolevas töös kasutatakse molekulaartunnuste

arvutamiseks RDKit tööriista versiooni 2024.09.06 (Landrum, 2024). Valik tulenes sellest, et toodud tarkvara on avatud lähtetekstiga, usaldusväärne ja hästi toetatud.

3.3 Tunnuste valik

QSAR analüüsil on tunnuste valik oluline, sest molekulaartunnuseid on palju ning mitmed neist ei ole tihtipeale uuritava tunnusega isegi seotud ja vähendavad hoopis mudeli ennustusvõimet. Tunnuste valikul on eesmärgiks leida tunnuste alamhulk, mis kirjeldaks uuritavat objekti minimaalse koguse tunnustega muutmata seejuures oluliselt kirjelduse kvaliteeti (Musil jt, 2021).

Shahlaei (2013) sõnul teeb QSAR analüüsil tunnuste valiku keeruliseks suur algoritmide valik ja algoritmide sobivuse sõltuvus koostatava mudeli tüübist. Samas on see mudeli loomisel äärmiselt vajalik etapp, sest väiksem tunnuste arv aitab ehitada täpsemaid, vähem ülesobitatud, lihtsamini kirjeldatavaid mudeleid ning vähendab ka arvutusmahtu (Khan & Roy, 2018).

Keerukaks teeb tunnuste valiku ka see, et tehnoloogiliste arengute tõttu on ühe keemilise ühendi kohta võimalik arvutada tuhandeid molekulaartunnuseid (Khan & Roy, 2018). Khan ja Roy (2018) rõhutavad oma töös, et viimasel ajal on just selle tõttu interpreteeritava mudeli loomiseks tunnuste valik väga oluline. Tunnuste valiku olulisust kinnitab ka selles valdkonnas kinnitust saanud teadmine, et QSAR analüüsi puhul annab vajaliku informatsiooni edasi juba väike alamhulk arvutatud molekulaartunnustest, teised on pigem üleliigsed ja suurendavad müra (Shahlaei, 2013).

QSAR metoodikat kasutades on tunnuste valiku algoritmi valides lisaks veel tähtis, et hiljem oleks valitud tunnuseid võimalik tõlgendada. QSAR analüüsides kasutatakse tihti nii klassikaliseid, tehisarul põhinevaid kui ka muid tunnuste valiku meetodeid (Shahlaei, 2013). Kõigi nende meetodite kirjeldamine ja analüüs ei kuulu käesoleva töö raamidesse, huvi korral annab Shahlaei (2013) oma töös põhilistest kasutusel olevatest tunnuste valiku meetoditest hea ülevaate.

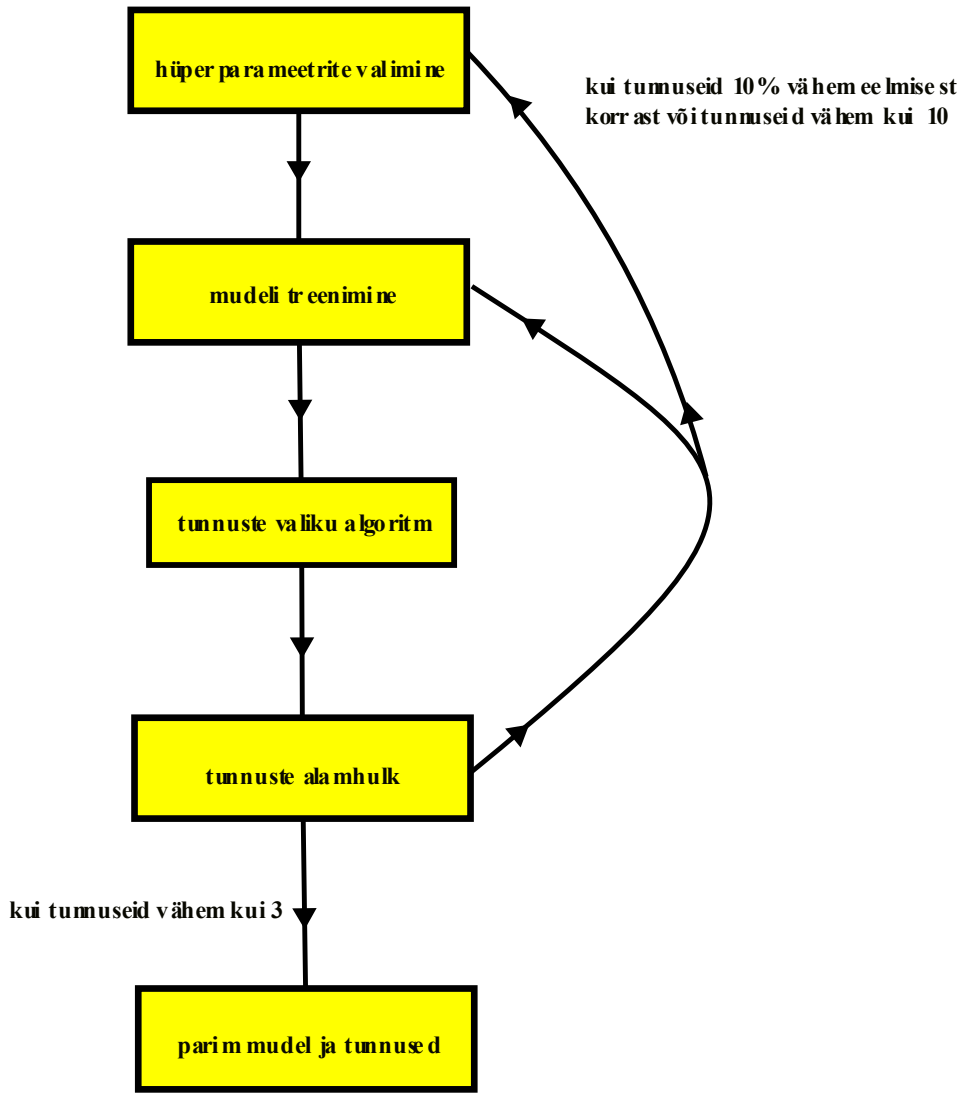
Üldiselt jagatakse tunnuste valiku meetodid ümbritsevateks, filtreerivateks ja sisseehitatuteks, millel kõigil on omad eelised ja vead, tihti arvutusvõimekusega ja stabiilsusega seotult (Li jt, 2017). Nad mainivad ka, et mainitud vigade leevendamiseks kasutatakse neid tihti põimitult ehk hübriidselt.

Lisanüanss, mida tuleb arvesse võtta tunnuste valiku meetodeid valides, tuleneb andmete enda iseloomust. Molekulaartunnuste ja uuritavate omaduste vahel on tihti väga keerulised

mittelineaarsed seosed, mille tõttu on, isegi varasemate uuringute olemasolu korral, peaaegu võimatu valida tunnuseid manuaalselt (Teixeira jt, 2013). Samuti eeldab see lineaarsusest sõltumatute meetodite kasutamist tunnuste hindamisel. Teixeira jt (2013) toovad ka veel välja, et suure arvu keemiliste ühendite ja nende molekulaartunnuste tõttu on mõttekam kõiki kombinatsioone mitte läbi proovida ja leida erinevate meetodite abil kompromiss arvutuskulukuse ja kõigi variantide kaalumise vahel.

Käesolevasse töösse valiti kõige eelneva põhjal järgnevad tunnuste valiku meetodid: Teixeira jt (2013) tööst inspireeritult juhumetsal põhinev tunnuste järjestamise meetod, üks Belfield jt (2023) kasutatud meetoditest, et tulemusi oleks võimalik võrrelda varasemalt sama andmekomplekti peal läbi viidud uuringuga, üks reguleerimise põhjal toimiv viis (Rahangdale & Raut, 2019) ning võrdluseks ka üks närvivõrkude kaalude põhjal töötav meetod (De Sá, 2019). Kõiki nelja meetodit kasutati närvivõrku ümbritsevalt, vähendades järjest tunnuste arvu, hinnates tunnuseid uuesti vastava meetodi abil iga kord, kui tunnuseid vähendatakse (vt joonis 3).

Märgates , et teatud arvu tunnuste vähendamisel mudeli täpsus järsult langeb, otsustati, et kui tunnuste arv väheneb kümme protsenti või tunnuseid on kokku alla kümne, siis valitakse mudeli parameetrid uuesti. Kuna mudel oli optimeeritud kasutama alguses suuremat andmehulka siis aegajalt parameetrite uuesti optimeerimine aitas muutuva andmehulgaga kohaneda. Selline lähenemine parandas märkimisväärselt väikese arvu tunnustega mudelite täpsust.



Joonis 3. Tunnuste valiku algoritm

Järgnevalt kirjeldatakse kasutatud algoritmide tööpõhimõtteid ja tausta.

- Esimene kasutatavatest meetoditest põhineb juhumetsal. Selleks treenitakse andmetel kõigepealt juhumets ning seejärel arvutatakse hinnangud tunnuste tähtsusele. Teixeira jt (2013) avastasid, et juhumetsa abil tunnuseid vähendades suudeti saada lausa 23 protsenti madalam keskmine ruutviga.
  Esimesena tuli juhumetsa ideega välja Breiman (2001). Need koosnevad mitmetest puudest, mis on treenitud andmete alamhulkade peal ning nende saadud tulemustest võetakse keskmine (Pedregosa jt, 2011). Nende sõnul see aitab parandada täpsust ja hoida ära ülesobitamist. Nad selgitavad ka puude enda algoritmi: need töötavad

jagades rekursiivselt tunnuste hulka kaheks nii, et sarnaste väärtustega tunnused oleksid koos. Jagamise kvaliteeti hinnatakse kaofunktsiooni abil, milleks Scikit learn lahenduses on keskmine ruutviga (Pedregosa jt, 2011). Edasi selgitavad nad, et kõige vähem kahju omav jaotus valitakse liites kokku parema ja vasaku jaotuse keskmised ruutvead ning arvestades sisse ka tunnuste arvu. Tunnuse tähtsus arvutatakse Gini meetodil: iga puu jaoks arvutatakse jaotuse kriteeriumite vähenemised tunnuse eemaldamise korral ning normaliseeritakse saadud väärtus (Pedregosa jt, 2011).

- Teiseks testiks valiti üks klassikalistest tunnuste valiku meetoditest. Ka selle autoriks on Breiman (2001), kes kasutas seda oma juhumetsa korral tunnuste hindamiseks. See töötab segades tunnuse väärtused ning arvutades kui palju mudeli tulemus halveneb (Pedregosa jt, 2011). Nad selgitavad, et kuna algoritm töötab juhuslikult siis täpsuse suurendamiseks arvutatakse tihti mitmete arvutuste keskmised.
- Kolmandas meetodis kasutatakse L1 regulatsiooni. Seda kasutatakse üle sobitamise vähendamiseks lisades igale kaalule närvivõrgu kaofunktsioonis juurde lisatingimuse (Rahangdale & Raut, 2019). L1 regulatsiooni lisatingimus arvutatakse nende sõnul summeerides iga tunnuse kaalude absoluutväärtused. Nad toovad ka välja, et kasutades seda sisendkihil pakub see hea meetodi tunnuste valikuks.
- Viimaseks tunnuste mõju kirjeldavaks algoritmiks valiti VIANN (ingl *variance-based feature importance of artificial neural networks*). Garson (1991) väidab, et kuna treenimise käigus kaalusid muudetakse nii, et need sobituksid paremini andmetega siis on loomulik oletada, et lõplikud väärtused annavad aimu sellest, kui tähtsad on kasutatavad tunnused. Paljud meetodid põhinevad kaalude lõplikelt väärtustel, üks neist on ka Garsoni enda välja pakutud lahendus (Garson, 1991). VIANN võtab arvesse ka kaalude muutmisi. See tähendab, et mida rohkem kaalu sätitakse seda rohkem see järelikult tulemusele mõju avaldab (De Sá, 2019). De Sá kirjeldab, et igal treening epohhil uuritakse kaalusid ja arvutatakse tähtsus Welfordi dispersiooni (Welford, 1962) abil.

3.4 Närvivõrgu mudelite ehitamine

Koodi realiseerimiseks kasutati Tensorflow masinõppe platvormi, sest Tensorflow on laialdaselt kasutatud, üsnagi intuitiivne pärilevivõrkude ehitamiseks ning omab väga head dokumentatsiooni (Tensorflow Developers, 2025).

Andmete eeltöötluse alustuseks jagatakse andmed juhuslikult treening ja test andmeteks ning eemaldatakse liigselt korreleeritud tunnused (skaalal ühest nullini rohkem kui 0.9). Seejärel andmed, mille põhjal hakatakse ennustama, skaleeritakse, sest mitmed masinõppe algoritmid on tundlikud andmete jaotuse suhtes (Pedregosa jt, 2011). Nende sõnul peaks jaotus olema Gaussian, keskmine väärtus null ja dispersioon ühe ühiku raames. Praktikas see eriti suurt mõju ei avalda, kaasaarvatud närvivõrkude korral, jaotust tihtipeale eiratakse ja arvutatakse uued väärtused lahutades keskmise ja jagatakse standardhälbega (Pedregosa jt, 2011). Seejärel eemaldatakse liiga vähese varieeruvusega tunnused. Siis hakatakse andmete põhjal närvivõrku treenima.

Belfield jt (2023) on uurinud erinevaid parameetreid närvivõrgu loomisel kasutades sama andmekomplekti, mida kasutatakse ka käesolevas töös. Nende leitud parameetreid arvestati käesolevas töös parameetrite väärtuste vahemike defineerimisel. Kasutatud parameetrite vahemike ja valikuid saab näha käesoleva töö GitHubi repositooriumist (vt lisa 1). Optimeerimise käigus avastati, et ka etteantud vahemikud ja valikud mõjutavad leitava mudeli kvaliteeti arvestatava R^2 muutuse võrra. Seetõttu parameetrite vahemikke kohandati hiljem manuaalselt vaadates parimate treenitud mudelite parameetreid ja kohendades selle järgi võimalikke valikuid ja vahemikke.

Parameetrite valikuks kasutati Optuna optimeerijat, sest see omab head dokumentatsiooni, töötab kiiresti ja on võrreldes teiste variantidega intuitiivne kasutada (Akiba jt, 2019). Konkreetsete parameetritega mudeli kvaliteedi hindamisel kasutati ristvalideerimist jagades algse andmekomplekti kümneks osaks nagu soovitasid Belfield jt (2019). Hüperparameetrite optimeerimisel kasutati töös mudelite hindamisel keskmist ruutviga (RMSE, ingl *root mean squared error*).

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt[2]{\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

RMSE väärtus arvutatakse valemi 3 alusel, kus *n* on andmete koguarv, *y* väljendab tõelisi ning $\hat{y}$ ennustatud väärtuseid[2]. Selleks, et tulemused oleksid reprodutseeritavad kasutati tulenevalt Bergstra jt (2025) saadud tulemustest puustruktuurilist Parzeni hinnangu põhist parameetrite valiku algoritmi, mille puhul määrati ära juhuarvu (ingl *seed*) abil.

Järgnevalt kirjeldatakse loodud närvivõrkude arhitektuuri. Kasutatakse jadamisi mudelit, mille esimene kiht on sisendkiht. Peale sisendkihti lisatakse peidetud kihid, mis kasutavad ReLU aktivatsiooni funktsiooni (Tensorflow Developers, 2025). Neile omakorda järgnes igaühele üks väljajätu kiht. Mudel lõppeb lineaarset aktivatsiooni funktsiooni kasutava väljundkihiga.

Mudeli täpsuse hindamiseks kasutati R^2 väärtust, sest seda on kasutatud ka varasemates töödes mistõttu on hiljem käesoleva töö tulemusi paremini varasemate uuringutega võrrelda. R^2 ehk determinatsioonikordaja väljendab kui suur osa dispersioonist on seletatav tunnuste abil[3].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(tõeline_i - ennustatud_i)^2}{\sum_{i=1}^{n}(tõeline_i - kõigitõelistekeskmine)^2} \quad (4)$$

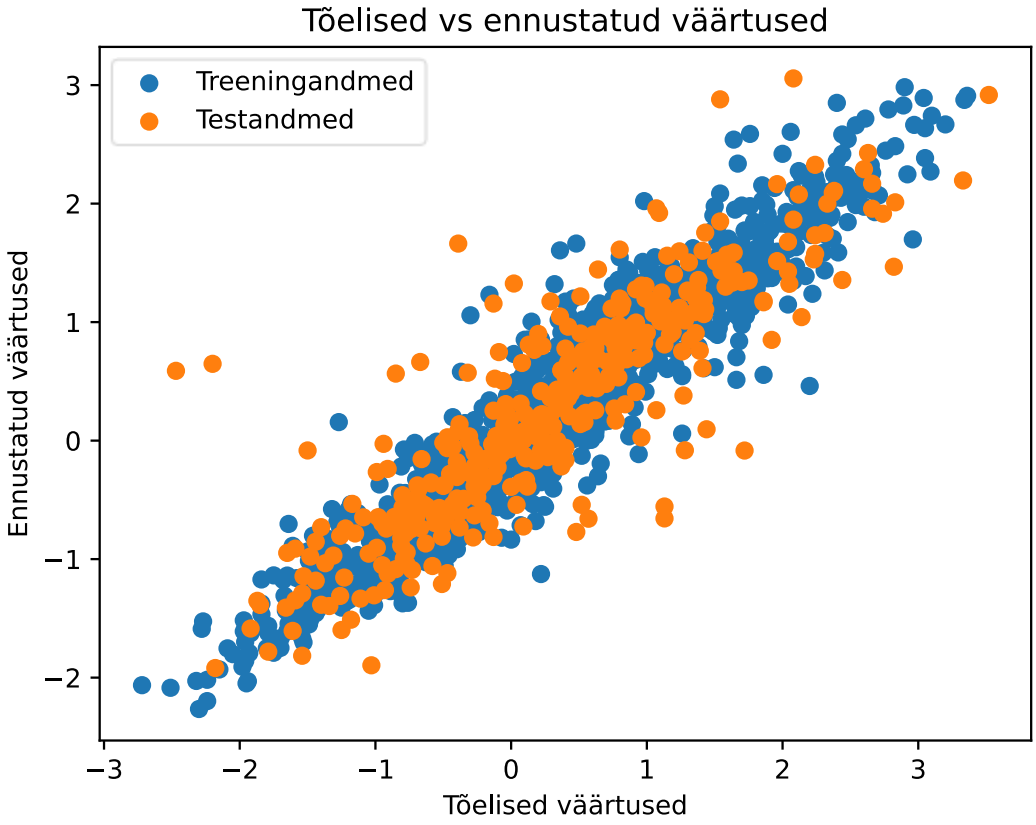
R^2 väärtus arvutatakse valemi 4 järgi, kus *n* on kogu väärtuste arv ja *i* vaadatav konkreetne väärtus.

[2] https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error
[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

4. Tulemused

Andmete eeltöötlusel eemaldati tugeva tunnuste vahelise korrelatsiooni tõttu (rohkem kui 0.9) 90 tunnust. Peale töötlust oli tunnuseid 169. Andmed jagati juhuslikult treening- ja testandmeteks. Kokku oli treeningandmeid 1595 ja testandmeid 399.

Kõiki töödeldud tunnuseid kasutades saadi mudel, mille ristvalideeritud R^2 treeningandmestikul oli 0.9382 ja testandmestikul 0.7875. Vastavad RMSE väärtused olid 0.2571 ja 0.4733. Visuaalselt on R^2 treening ja testandmetel välja toodud joonisel 4.



Joonis 4. Algse mudeli ennustatud ja tõelised väärtused

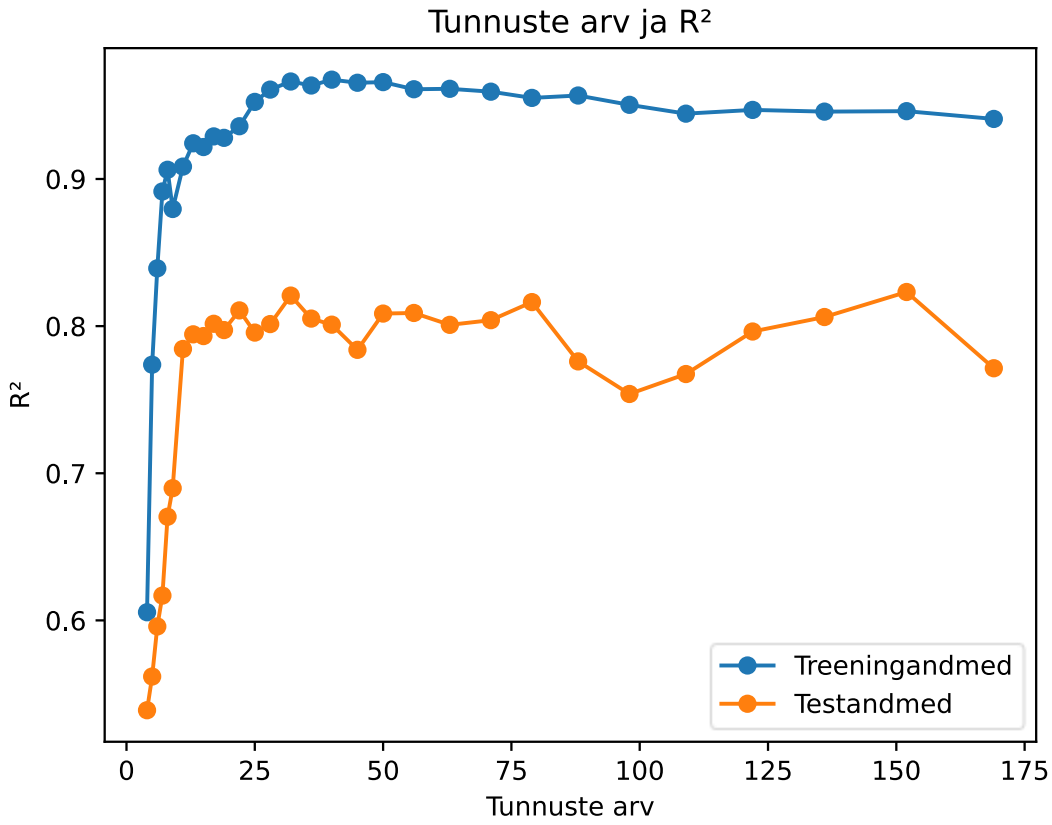
Võrreldes seda mudelit tabelis 1 välja toodud varasemate mudelitega võib leitud mudelit pidada konkurentsivõimeliseks selle ennustusvõime poolest. Oma täpsuse poolest ristvalideeritud treeningkomplektil on see mudel parem kui mudel, mille Belfield jt (2023) oma töös välja pakkusid. Yu jt (2020), Yu jt (2010) ja Zhu jt (2008) töödele jääb leitud mudel küll veidi alla, kuid ühendeid on antud andmekomplektis rohkem, mille tõttu oli see oodatav, sest suuremas andmekomplektis on erinevate omadustega ühendeid rohkem ja seega ka nende ennustamine keerulisem.

4.1 Algoritmide võrdlus

Väljatoodud mudelis hakati kasutatavaid tunnuseid vähendama peatükis 3.3 kirjendatud meetodil. Parimad meetodid valiti R^2 väärtuste järgi võttes arvesse vaid mudelid, kus on vähem kui 25 tunnust. Järgnevalt tuuakse välja tulemused iga erineva lähenemisviisi korral ja tulemuste võrdlused.

4.1.1 Permutatsioonitest

Esiteks vaadati permutatsioonitesti abil hinnatud tunnuste eemaldamist. Jooniselt 5 on näha, et tunnuseid vähendades on võimalik teatud piirini jõuda ilma ennustusvõimet kaotamata, kohati lausa seda parandades.



Joonis 5. Permutatsioonitestiga tunnuste valimine

Mudelis, kuhu jäi ainult neli tunnust, olid tunnusteks MolLogP, BCUT2D_CHGLO, BCUT2D_LOGPLOW, MQN14.

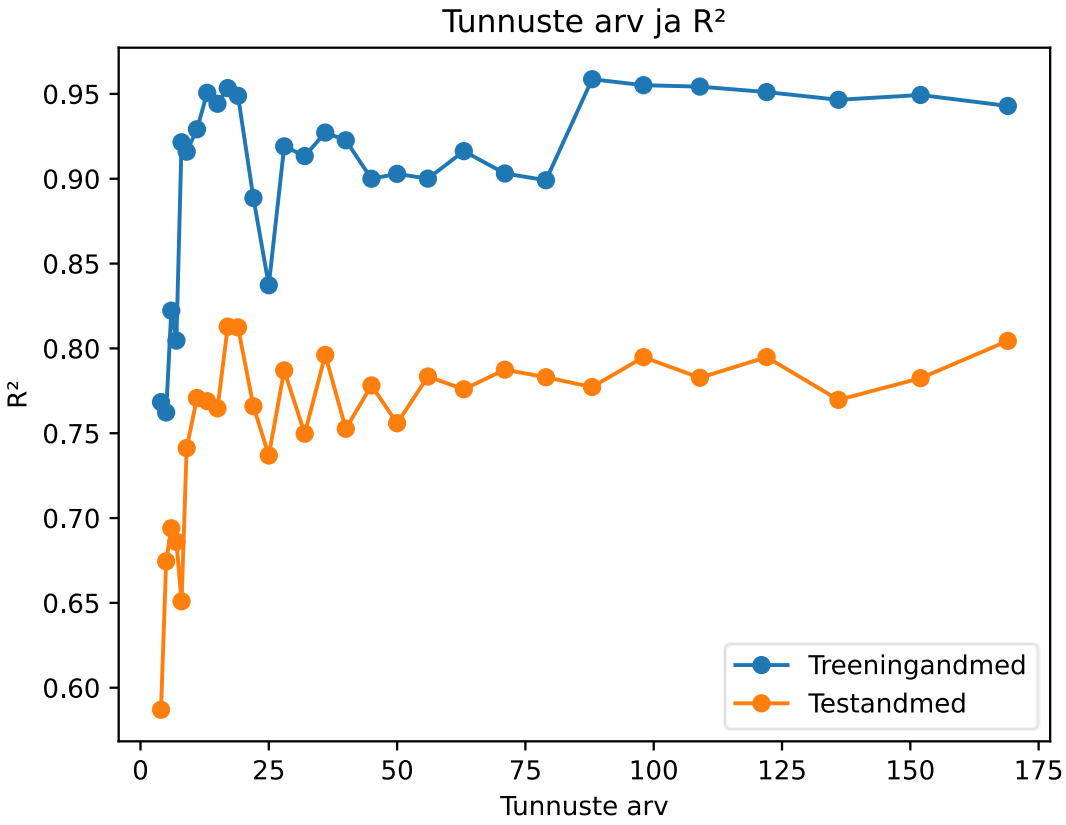
Tabel 3. Permutatsioonitestiga valitud tunnustega parimad mudelid

Treening R^2	Test R^2	Treening RMSE	Test RMSE	Tunnuste arv
0.9359	0.8107	0.9519	0.9164	22
0.9289	0.8016	0.9014	0.8463	17
0.9280	0.7973	1.0963	1.0307	19
0.9243	0.7944	0.9007	0.8351	13
0.9217	0.7932	0.9221	0.9178	15

Parimad saadud mudelid on toodud tabelis 3. Nendest parima R2 on treeningandmetel ristvalideerimisega 0.9359 ja testandmetel 0.8107. Kasutatavaid tunnuseid oli selles mudelis 22.

4.1.2 Juhumets

Teiseks uuriti juhumetsa abil valitud tunnuseid ja neile vastavaid mudeleid. Joonisel 6 on näha permutatsioonitestile sarnane olukord, kus teatud tunnuste arvuni mudelite ennustusvõime kas kahaneb minimaalselt või lausa kasvab.



Joonis 6. Juhumetsaga tunnuste valimine

Kõige viimasesse proovitud mudelisse jäid tunnused MolWt, MolLogP, MinPartialCharge, qed.

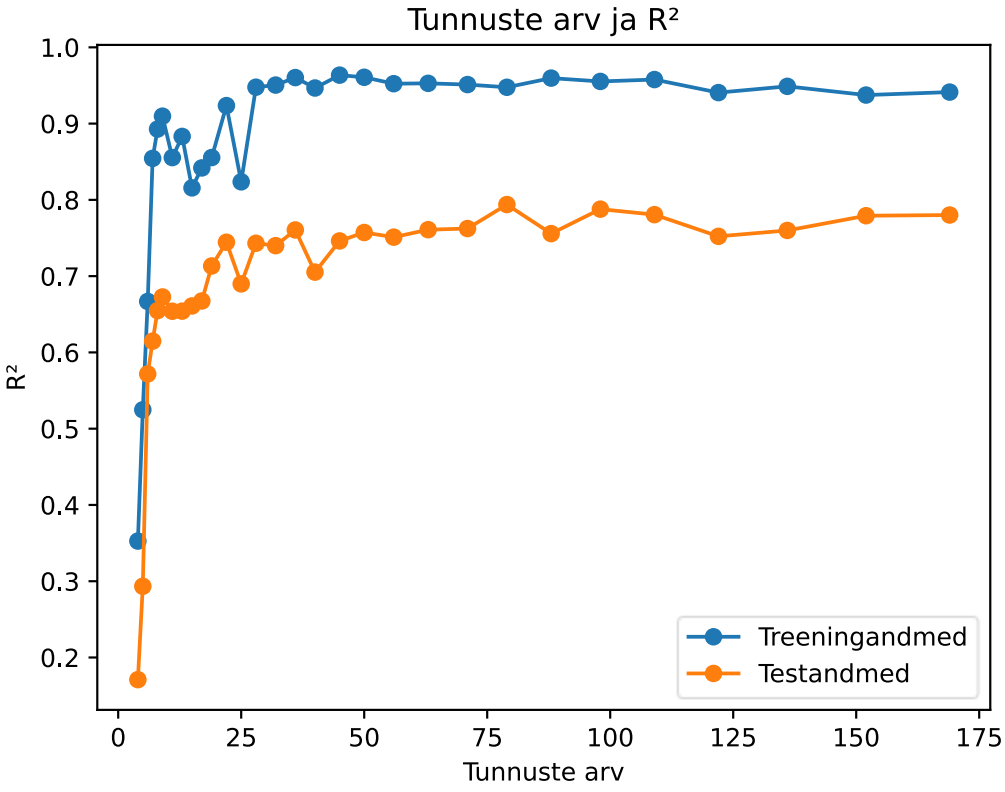
Tabel 4. Juhumetsaga valitud tunnustega parimad mudelid

Treening R^2	Test R^2	Treening RMSE	Test RMSE	Tunnuste arv
0.9534	0.8128	0.8878	0.8586	17
0.9507	0.7689	0.8325	0.8545	13
0.9489	0.8124	0.9108	0.8525	19
0.9442	0.7647	0.8074	0.8002	15
0.9292	0.7708	0.8075	0.7866	11

Tabelis 4 on välja toodud parimad juhumetsa abil korrigeeritud tunnustega mudelid, millest parimal oli alles jäetud 17 tunnust 169-st. Selle sooritus treening- ja testandmetel oli vastavalt 0.9534 ja 0.8128.

4.1.3 L1 regulatsioon

Kolmandaks analüüsitavaks meetodiks oli L1 regulatsioon. Vaadates joonist 7 käitus ka see meetod sarnaselt eelnevatega. Välja joonistub selge platoo ning järsk langus toimub umbes 20 tunnuse abil ehitatud mudeli juures.



Joonis 7. L1 regulatsiooni abil vall tunnuste valimine

Viimasesse treenitud mudelisse jäid MaxAbsEStateIndex, MinAbsEStateIndex, MinEStateIndex ja qed, millest vaid üks kattub juhumetsa poolt välja valitud viimaste tunnustega.

Tabel 5. L1 regulatsiooni abil valitud tunnustega parimad mudelid

Treening R^2	Test R^2	Treening RMSE	Test RMSE	Tunnuste arv
0.9236	0.7445	0.2864	0.5192	22
0.9097	0.6727	0.3112	0.5876	9
0.8928	0.6548	0.3390	0.6034	8
0.8832	0.6541	0.3537	0.6040	13
0.8555	0.7133	0.3927	0.5499	19

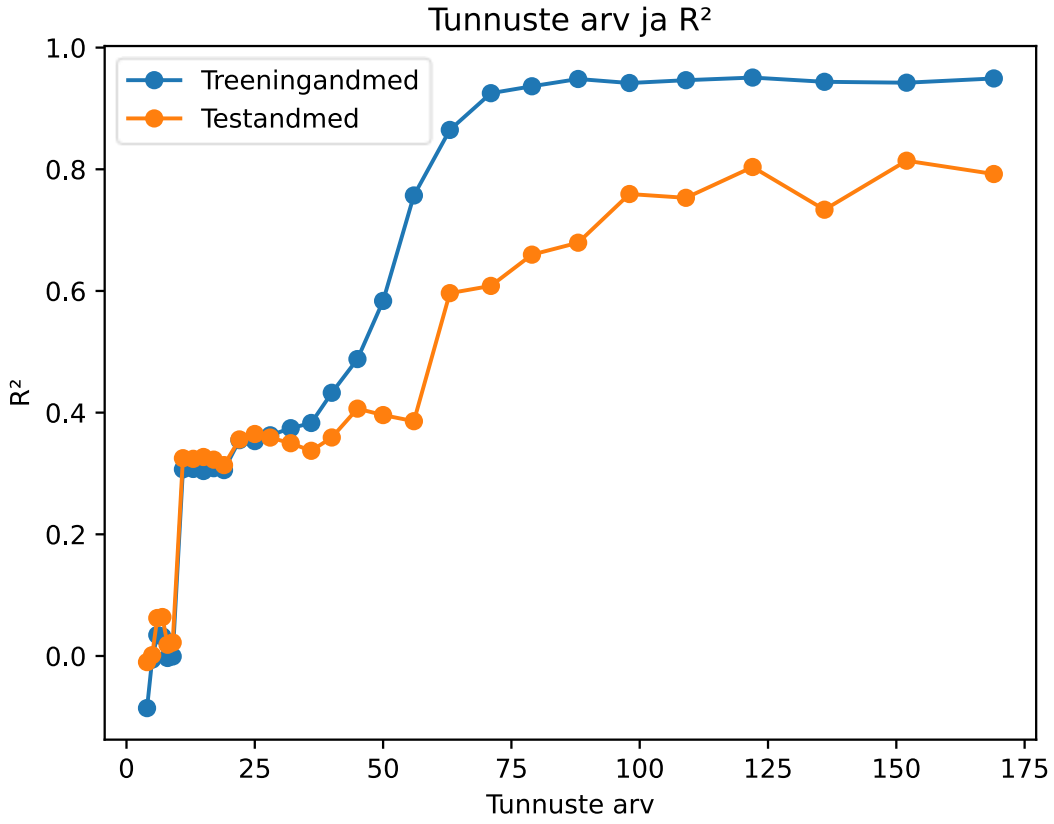
Parimate mudelite (vt tabel 5) tunnuste arvud ja R^2 väärtused treeningandmete korral on sarnased eelnevate meetoditega. Parim oli mudel 22 tunnusega, treening R^2 tulemusega 0.9236 ja test R^2 tulemusega 0.7445.

4.1.4 VIANN

Viimaseks analüüsitud meetodiks on VIANN. Ka selle puhul on märgata joonisel platood, kuid statistiku järsk vähenemine algab teistest meetoditest suurema hulga tunnuste juures (vt joonis 8).

Kõige väiksema arvuga tunnuseid kasutava mudeli halva ennustamisvõime tõttu võib eeldada, et need tunnused tegelikult mürgisuse ennustamisel suurt tähtsust ei oma. Nendeks tunnusteks on fr_HOCCN, fr_phos_acid, fr_N_O, fr_thiocyan.

Parim saadud mudel ennustas 22 tunnuse abil treeningandmetel täpsusega 0.3548 ja testandmetel 0.3560.



Joonis 8.VIANN abil vall tunnuste valimine

Tabel 6. VIANN abil valitud tunnustega parimad mudelid

Treening R^2	Test R^2	Treening RMSE	Test RMSE	Tunnuste arv
0.3548	0.3560	1.3341	1.3390	22
0.3087	0.3227	1.9963	1.9705	17
0.3074	0.3241	1.1372	1.1246	13
0.3068	0.3251	0.9004	0.8509	11
0.3056	0.3138	1.2394	1.2276	19

Tabelis 6 on välja toodud viis parimat VIANN abil leitud mudelit. Neid analüüsides võib väita, et selle meetodiga tunnuste valik jääb teistele silmnähtavalt alla. Suur langus toimub juba umbes 75 tunnuse juurde jõudes (vt joonis 8).

4.1.5 Võrdlus

Mudelite R^2 väärtuste järgi töötas kõige paremini tunnuste hindamine juhumetsa ja permutatsioonitesti kasutades. Sarnases suurusjärgus tulemused saadi ka L1 regulatsiooni abil, kuid võrreldes parimatega jäi see siiski teistele alla. Ka parimas mudelis kasutatud tunnuste arvu poolest oli parim juhumets 17 tunnusega. Järgnesid permutatsioonitest ja L1 valiku

algoritm 22 tunnusega. Tunnused, mis esinesid vähemalt kahe meetodi viimastes valitud tunnustes olid MolLogP ja qed. VIANN puhul olid tulemused kõige kehvemad. De Sá (2019) tõi ka oma töös välja, et kohati võib see meetod olla ebastabiilne, seega on meetodi teistest kehvem suutlikkus põhjendatav.

Välja toodud joonistel ja tabelites on märgata mitmete mudelite korral teatud erinevust treening- ja testandmete statistikutes. Eriti hästi on seda märgata L1 regulatsiooniga leitud mudelite R^2 väärtuste korral. Selle parandamiseks muudeti nii väljajätumeetodi, varase lõpetamise kui ka L1 regulatsiooni parameetreid. Nende muutmine ei toonud edu, seega võis treening- ja testandmetel statistikute erinevus tulla hoopis näiteks erinevate ainegruppide või ka mõõtmistulemustes sisalduvate hälvete ebaühtlasest jaotusest treening- ja testandmete vahel.

Tabel 7. Parimad mudelid

Treening R^2	Test R^2	Treening RMSE	Test RMSE	Tunnuste arv	Algoritm
0.9534	0.8128	0.8878	0.8586	17	juhumets
0.9507	0.7689	0.8325	0.8545	13	juhumets
0.9489	0.8124	0.9108	0.8525	19	juhumets
0.9442	0.7647	0.8074	0.8002	15	juhumets
0.9359	0.8107	0.9519	0.9164	22	permutatsioon

Vaadates viit kõige paremat saadud mudelit (vt tabel 7) on näha, et juhumets on olnud kõige parem, paremuselt teine on permutatsioon. Tulemus võiks viidata sellele, et mõlemad on tunnuste valikuks sobivad. Samuti, et käesoleva töö korral ei suuda närvivõrgupõhised meetodid tunnuste tähtsust paremini hinnata kui teistel masinõppe meetoditel baseeruvad tunnuste hindamise meetodid nagu seda on juhumets.

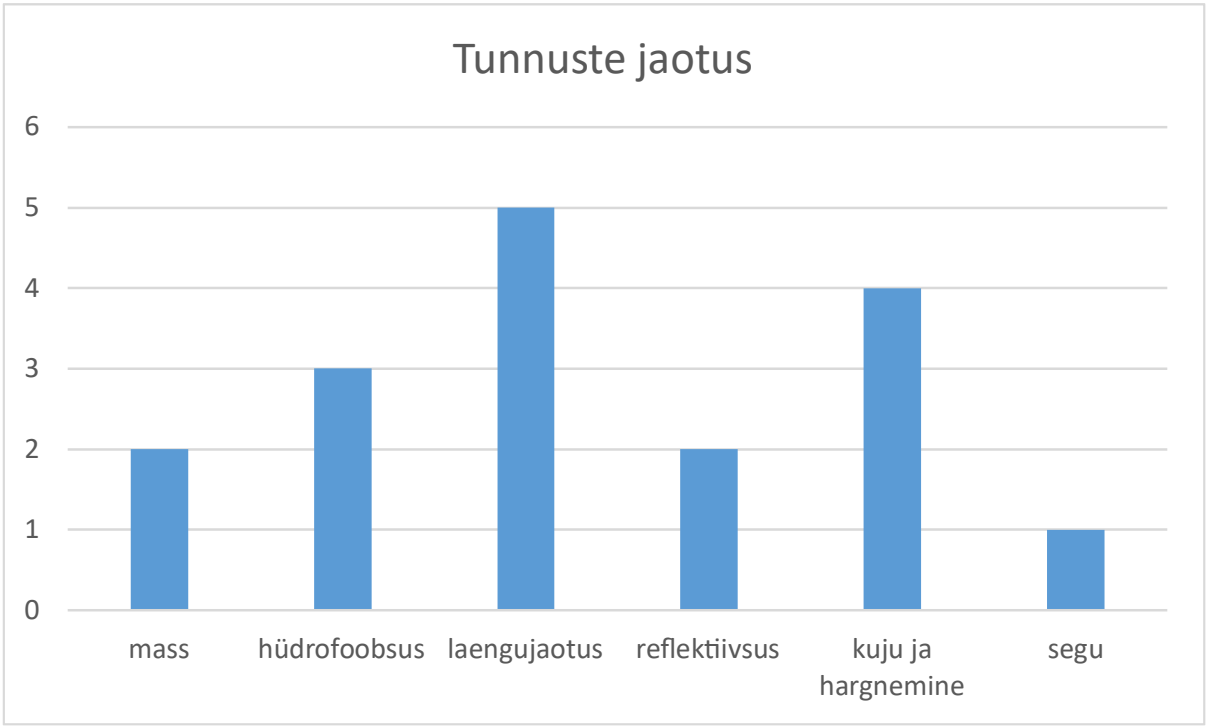
4.2 Parim mudel ja tunnused

Parimal mudelil oli statistik R^2 treeningandmetel 0.9534 ja testandmetel 0.8128 võrreldes kõiki tunnuseid kasutanud mudeli R^2 väärtusega treeningandmestikul 0.9382 ja testandmestikul 0.7875. Parim mudel oli ka testandmetel täpsem kõigi tunnustega ehitatud mudelist.

Võrreldes seda mudelit tabelis 1 välja toodud varasemate tulemustega on see igast küljest parem samal andmekomplektil tehtud mudelist. Samuti on mudel täpsem Yu (2020) töös leitud lahendusest. Teistele jääb see täpsuse poolest veidi alla, kuid täidab sama ülesande võrdväärse täpsusega kasutades mitu korda suuremat andmehulka ning vähem tunnuseid nii Yu jt (2010)

kui ka Zhu jt (2008) töödest. Vähem tunnuseid kasutas ainult Yu (2020) poolt kirjutatud uurimistöö mudel.

Parima mudeli 17 tunnuse keemilist sisu kirjeldati Landrum (2024) RDKit dokumentatsiooni põhjal (vt lisa 2) ja grupeeriti nende kirjelduse põhjal viite kategooriasse (vt joonis 9).



Joonis 9. Tunnuste jagunemine

Tunnused jagati nende olemuse põhjal algselt viite gruppi: hüdrofoobsust, kuju ja hargnemised, laengujaotust, massi ja reflektiivsust kirjeldavateks molekulaartunnusteks. Osad tunnustest olid keerukamad ja neid oleks saanud mitmesse gruppi lisada. Selliste puhul valiti grupp, mida see enim kirjeldas. Qed ehk kvantitatiivne ravimisarnasuse indeks, mis iseloomustab molekuli võimet omada bioloogilist aktiivsust, vajas siiski oma keerukuse tõttu lisakategooriat: segu ehk molekulaartunnus, mis kirjeldab peaaegu kõiki teisi gruppe.

Joonise 9 ja molekulaartunnuste kirjelduste alusel saab väita, et mürgisust mõjutab enim just molekuli laengujaotus ja molekuli hargnemised ning kuju, sidemed ja nende vahelised seosed. Lisaks omavad veidi rohkem rolli ka molekuli mõõtmed ja hüdrofoobsus. Need leiud on kooskõlas ka teiste leitud seostega näiteks Yu (2020) mudeliga, mis kasutas molekulaartunnuseid, mis kirjeldasid hüdrofoobsust, molekulmassi, polariseeritust molekulis ja molekulaartunnust, mis kirjeldas nii molekuli kuju, suurust kui ka lipofiilsust.

5. Kokkuvõte

Keemiliste ühendite keskkonnamõju määratlemisel on mürgisuse uuringud tähtsal kohal. Uute ühendite mõju prognoosimine mudelite abil aitab probleeme ette näha ja näiteks ühendite arendamisel kulutusi vähendada. Käesoleva töö eesmärgiks oli leida varasematest töödest täpsem viis mürgisuse prognoosimiseks ning analüüsida erinevaid tunnuste hindamise algoritme. Samuti läbi selle leida sõltuvusi mudeli omaduste ja mürgisuse vahel.

Parimaks tunnuste valiku algoritmiks oli juhumets. Sellega saadud parim mudel ennustas treeningandmetel täpsusega R^2 0.9534 ja testandmetel 0.8128. Mõõdikuks kasutati R^2 ja RMSE väärtuseid. See mudel kasutas 17 tunnust, mis kirjeldasid molekuli hüdrofoobsust, kuju ja hargnemist, laengujaotust ja massi. Ka varasemad tulemused näitasid sarnaseid sõltuvusi.

Töö algul seati eesmärgiks luua varasemast parem mudel. Tulemuseks saadud mudeli täpsus oli parem samal andmekomplektil varasemalt loodud mudelist. Ka andmehulk ja tunnuste arv olid paremad seega võib mudeli loomise lugeda õnnestunuks. Eesmärk saada infot kõige vajalikumate tunnuste kohta sai samuti täidetud ning kõige olulisemaks osutusid molekuli kuju ja hargnemised, laengujaotus ja hüdrofoobsus.

Käesolev töö annab suunitlusi edasisteks uuringuteks närvivõrkude kasutamise ja tunnuste valiku algoritmide kohta QSAR analüüsides. Samuti pakub käesolev töö tõestust, et närvivõrkude kasutamine mürgisuse ennustamiseks on sobiv. Edasiarendusena oleks näiteks võimalik närvivõrke proovida kasutada ilma molekulaartunnuseid arvutamata, et vähem informatsiooni läheks analüüsi käigus kaotsi.

Kasutatud kirjandus

Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. (2019). Optuna. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701.

Aoyama, T., Suzuki, Y. & Ichikawa, H. (1990). Neural Networks Applied to Quantitative Structure-Activity Relationship Analysis. *Journal of Medicinal Chemistry,* vol 33, issue 9, 2583-2590. https://doi.org/10.1021/jm00171a037.

Barbierato, E., Gatti, A., Incremona, A., Pozzi, A. & Toti, D. (2025). Breaking Away from AI: The Ontological and Ethical Evolution of Machine Learning. *IEEE*, vol 13, 55627-55647. https://doi.org/10.1109/access.2025.3553032

Belfield, S. J., Cronin, M. T., Enoch, S. J. & Firman, J. W. (2023). Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs). *PLoS ONE, vol 18, issue 5*. https://doi.org/10.1371/journal.pone.0282924.

Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. (2025) Algorithms for Hyper-Parameter Optimization. https://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf (13.05.2025).

Bock, S. & Weis, M. (2019). A proof of local convergence for the Adam Optimizer. *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/ijcnn.2019.8852239.

Breiman, L. (2001). Random forests. *Machine Learning, vol 45, issue 1*, 5–32. https://doi.org/10.1023/a:1010933404324.

De Sá, C. R. (2019). Variance-Based feature importance in neural networks. *Lecture notes in computer science*, 306–315. https://doi.org/10.1007/978-3-030-33778-0_24.

Elhag, I. Y. (2023). Role of AI in ADME/Tox toward formulation optimization and delivery. *Elsevier eBooks*, 301–345. https://doi.org/10.1016/b978-0-323-89925-3.00011-3.

Garson, G. D. (1991). Interpreting neural-network connection weights. *AI Expert Archive*, vol 6, issue 4, 46–51. https://doi.org/10.5555/129449.129452.

Ghosh, V., Bhattacharjee, A., Kumar, A. & Ojha, P. (2024). q-RASTR modelling for prediction of diverse toxic chemicals towards T. pyriformis. *SAR And QSAR in Environmental Research*, vol 35, issue 1, 11–30. https://doi.org/10.1080/1062936x.2023.2298452.

Gini, G. (2018). QSAR: What else? *Methods in Molecular Biology*, 79–105. https://doi.org/10.1007/978-1-4939-7899-1_3.

Goodfellow, I., Bengio, J. & Courville, A. (2016). Deep learning. Ameerika Ühendriigid: MIT Press. https://www.deeplearningbook.org/ (20.11.2024).

Grossi, E. & Buscema, M. (2007). Introduction to artificial neural networks. *European Journal of Gastroenterology & Hepatology*, *19*(12), 1046–1054. https://doi.org/10.1097/meg.0b013e3282f198a0.

Hansch, C., Muir, R., Fujita, T., Maloney, P. F., Geiger, F. & Streich, M. (1963). *Journal of the American Chemical Society,* vol 86, 2817.

Kahn, I., Sild, S. & Maran, U. (2007). Modeling the Toxicity of Chemicals to Tetrahymena pyriformis Using Heuristic Multilinear Regression and Heuristic Back-Propagation Neural Networks. *Journal of Chemical Information and Modeling*, vol 47, issue 6, 2271–2279. https://doi.org/10.1021/ci700231c.

Keshavarz, M. H., Shirazi, Z. & Sheikhabadi, P. K. (2021). Risk assessment of organic aromatic compounds to Tetrahymena pyriformis in environmental protection by a simple QSAR model. *Process Safety and Environmental Protection*, vol 150, 137–147. https://doi.org/10.1016/j.psep.2021.04.011.

Khan, P. M. & Roy, K. (2018). Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). *Expert Opinion on Drug Discovery*, vol 13, issue 12, 1075–1089. https://doi.org/10.1080/17460441.2018.1542428.

Kumar, A., Podder, T., Kumar, V. & Ojha, P. K. (2022). Risk assessment of aromatic organic chemicals to T. pyriformis in environmental protection using regression-based QSTR and Read-Across algorithm. *Process Safety and Environmental Protection*, vol 170, 842–854. https://doi.org/10.1016/j.psep.2022.12.067.

Landrum, G. RDKit: Open-source cheminformatics. https://www.rdkit.org (14.05.2025).

Lee, Y.-R., Lai M.-C., Liu H.-H. & Chen J.-R. (2023). The Tetrahymena Paradigm: Genetic Insights and Ecotoxicity Assessment. *Journal of Environmental Toxicology*, vol 9, issue 1. Ameerika Ühendriigid: Austin Publishing Group, 1044. https://austinpublishinggroup.com/environmental-toxicology/fulltext/ajet-v9-id1044.php (06.05.2025).

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. & Liu, H. (2017). Feature selection. *ACM Computing Surveys*, vol 50, issue 6, 1–45. https://doi.org/10.1145/3136625.

Maurya, R. & Pandey, A. K. (2020). Importance of protozoa Tetrahymena in toxicological studies: A review. *The Science of the Total Environment*, vol 741, 140058. https://doi.org/10.1016/j.scitotenv.2020.140058.

Meuwly, M. (2021). Machine learning for chemical reactions. *Chemical Reviews*, vol 121, issue 16, 10218–10239. https://doi.org/10.1021/acs.chemrev.1c00033.

Mollaei, M., Abdollahpour, S., Atashgahi, S., Abbasi, H., Masoomi, F., Rad, I., Lotfi, A. S., Zahiri, H. S., Vali, H. & Noghabi, K. A. (2009). Enhanced phenol degradation by Pseudomonas sp. SA01: Gaining insight into the novel single and hybrid immobilizations. *Journal of Hazardous Materials*, vol 175, issue 1–3, 284–292. https://doi.org/10.1016/j.jhazmat.2009.10.002.

Musil, F., Grisafi, A., Bartók, A. P., Ortner, C., Csányi, G. & Ceriotti, M. (2021). Physics-Inspired Structural Representations for Molecules and Materials. *Chemical Reviews*, vol 121, issue 16. Ameerika Ühendriigid: American Chemical Society, 9759-9815. https://doi.org/10.1021/acs.chemrev.1c00021.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & teised. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, vol 12(Oct), 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf (06.05.2025).

Pushkarova, Y., Zaitseva, G. & Saker, M. A. (2022). Prediction of toxicity of phenols using artificial neural networks. *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*, 493–496. https://doi.org/10.1109/acit54803.2022.9913174.

Rahangdale, A. & Raut, S. (2019). Deep Neural Network Regularization for Feature Selection in Learning-to-Rank. *IEEE Access*, vol 7, 53988–54006. https://doi.org/10.1109/access.2019.2902640.

Ren, S. (2003). Modeling the Toxicity of Aromatic Compounds to Tetrahymena pyriformis: The Response Surface Methodology with Nonlinear Methods. *Journal of Chemical Information and Computer Sciences*, vol 43, issue 5, 1679–1687. https://doi.org/10.1021/ci034046y.

Ruusmann, V. & Maran, U. (2013). From data point timelines to a well curated data set, data mining of experimental data and chemical structure data from scientific articles, problems and possible solutions. *Journal of Computer-Aided Molecular Design*, vol 27, issue 7, 583–603. https://doi.org/10.1007/s10822-013-9664-4.

Sauvant, N., Pepin, D. & Piccinni, E. (1999). Tetrahymena pyriformis: A tool for toxicological studies. A review. *Chemosphere*, vol 38, issue 7, 1631–1669. https://doi.org/10.1016/s0045-6535(98)00381-6.

Schultz, T. W. (1997). TETRATOX: TETRAHYMENA PYRIFORMIS POPULATION GROWTH IMPAIRMENT ENDPOINTA SURROGATE FOR FISH LETHALITY. *Toxicology Methods*, vol 7, issue 4, 289–309. https://doi.org/10.1080/105172397243079.

Seward, J. R., Hamblen, E. & Schultz, T. W. (2002). Regression comparisons of tetrahymena pyriformis and poecilia reticulata toxicity. *Chemosphere*, vol 47, issue 1, 93–101. https://doi.org/10.1016/s0045-6535(00)00473-2.

Shahlaei, M. (2013). Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: a review study. *Chemical Reviews*, vol 113, issue 10, 8093–8103. https://doi.org/10.1021/cr3004339.

Sharkawy, N. A. (2020). Principle of neural network and its main types: review. *Journal of Advances in Applied & Computational Mathematics*, vol 7, 8–19. https://doi.org/10.15377/2409-5761.2020.07.2.

Svozil, D., Kvasnicka, V. & Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems*, vol 39, issue 1, 43–62. https://doi.org/10.1016/s0169-7439(97)00061-0.

Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., Öberg, T., Dao, P., Cherkasov, A. & Tetko, I. V. (2008). Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena pyriformis. *Journal of Chemical Information and Modeling*, vol 48, issue 4, 766–784. https://doi.org/10.1021/ci700443v.

Tan, H., Jin, J., Fang, C., Zhang, Y., Chang, B., Zhang, X., Yu, H. & Shi, W. (2023). Deep learning in Environmental Toxicology: current progress and open challenges. *ACS ES&T Water*, vol 4, issue 3, 805–819. https://doi.org/10.1021/acsestwater.3c00152.

Teixeira, A. L., Leal, J. P. & Falcao, A. O. (2013). Random forests for feature selection in QSPR Models - an application for predicting standard enthalpy of formation of hydrocarbons. *Journal of Cheminformatics*, vol 5, 9. https://doi.org/10.1186/1758-2946-5-9.

TensorFlow Developers. (2025). TensorFlow (v2.19.0). Zenodo. https://doi.org/10.5281/zenodo.15009305.

Tkatchenko, A. (2020). Machine learning for chemical discovery. *Nature Communications*, vol 11, issue 1. https://doi.org/10.1038/s41467-020-17844-8.

Wigh, D. S., Goodman, J. M. & Lapkin, A. A. (2022). A review of molecular representation in the age of machine learning. *Wiley Interdisciplinary Reviews Computational Molecular Science*, vol 12, issue 5. https://doi.org/10.1002/wcms.1603.

Xu, L., Ball, J., Dixon, S. & Jurs, P. (1994). Quantitative structure-activity relationships for toxicity of phenols using regression analysis and computational neural networks. *Environmental Toxicology and Chemistry*, vol 13, issue 5, 841–851. https://doi.org/10.1002/etc.5620130520.

Yu, X. (2020). Prediction of chemical toxicity to Tetrahymena pyriformis with four-descriptor models. *Ecotoxicology and Environmental Safety*, vol 190, 110146. https://doi.org/10.1016/j.ecoenv.2019.110146.

Yu, Y.-J., Liu, R., Su, R.-X., Wang, L.-B., Qi, W. & He, Z.-M. (2010). Artificial neural network approach for prediction of toxicity of organic compounds based on an improved group contribution method. *Fresenius Environmental Bulletin*, vol 19, issue 12, 2777–2782.

Lisad

1. GitHubi repositoorium

https://github.com/LisettePihor/Acute_toxicity_modeling.git

2. Tunnuste kirjeldused

Molekulaartunnus	Tähendus	Grupeering
MolWt	molekuli kaal	mass
MolLogP	logP väärtus	hüdrofoobsus
qed	ravimisarnasuse indeks	segu
MinPartialCharge	väikseim osalaeng	laengujaotus
SMR_VSA10	molekulaarse töökeskkonna ingl Molecular Operating Environment (MOE) molekuli murdumisnäitaja ja van der Waalsi pinna (VSA) mõõtmete molekulaartunnus 10 (4.00 <= x < inf)	reflektiivsus
BertzCT	Bertzi molekuli keerukuse hinnang	kuju ja hargnemine
BCUT2D_LOGPHI	suurim Burdeni maatriksi omaväärtus logPjärgi	hüdrofoobsus
BCUT2D_LOGPLOW	väikseim Burdeni maatriksi omaväärtus logP järgi	hüdrofoobsus
SPS	ruumilise keerukuse skoor	kuju ja hargnemine
MinEStateIndex	minimaalne elektrotopoloogilise staadiumi indeks	laengujaotus
BCUT2D_MRLOW	väikseim Burdeni maatriksi omaväärtus molekuli murdumisnäitaja järgi	reflektiivsus
VSA_EState3	van der Waalsi pinna (VSA) ja elektrotopoloogilise staadiumi molekulaartunnus 3 (5.00 <= x < 5.41)	laengujaotus
VSA_EState2	van der Waalsi pinna (VSA) ja elektrotopoloogilise staadiumi molekulaartunnus 2 (4.78 <= x < 5.00)	laengujaotus
BCUT2D_MWLOW	väikseim Burdeni maatriksi omaväärtus molekulmassi järgi	mass
MQN14	molekuli ahelalistele kaksiksidemete arv	kuju ja hargnemine
PEOE_VSA6	molekulaarse töökeskkonna ingl Molecular Operating Environment (MOE) laengu ja van der Waalsi pinna (VSA) molekulaartunnus 6 (-0.10 <= x < -0.05)	laengujaotus
BalabanJ	Balabani topoloogiline indikaator	kuju ja hargnemine

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Lisette Pihor ,
(autori nimi)

annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

Tunnuste valik närvivõrguga akuutse mürgisuse prognoosimisel ,
(lõputöö pealkiri)

mille juhendaja(d) on Sulev Sild ja Uko Maran ,
(juhendaja nimi)

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;

olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;

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

Lisette Pihor

14.05.2025