PROGRAM ORANGE KOT ORODJE ZA NAPOVED LASTNOSTI MOLEKUL

PROGRAM ORANGE KOT ORODJE ZA NAPOVED LASTNOSTI MOLEKUL

Področje: **KEMIJA**

Vrsta naloge: **Raziskovalna naloga**

Dijakinja: **Špela Žunec, 2. G**

Mentorica: **Mag. Mojca Podlipnik, GJPL**
Somentor: **Dr. Črtomir Podlipnik, FKKT, Ljubljana**

2020

Gimnazija Jožeta Plečnika Ljubljana

KAZALO

SEZNAM PRILOG

Priloga 1: Skeletne formule in kode SMILES obravnavanih policikličnih aromatskih ogljikovodikov.
Priloga 2: Podatki o toksičnosti pesticidov (pLC_{50}) in ostali opisniki ter napovedane vrednosti z metodama *LR* in *RF* (preverjanje napovedne moči je bilo izvršeno z načinom *LOO*).

NASLOV NALOGE: Program Orange kot orodje za napoved lastnosti molekul

RAZISKOVALKA: Špela Žunec, 2. G

ŠOLA: Gimnazija Jožeta Plečnika Ljubljana

MENTOR: mag. Mojca Podlipnik – Gimnazija Jožeta Plečnika Ljubljana

SOMENTOR: dr. Črtomir Podlipnik – Fakulteta za kemijo in kemijsko tehnologijo, Ljubljana

KLJUČNE BESEDE: program Orange, napoved vrelišč, napoved toksičnosti, topološki indeksi

POVZETEK NALOGE:

Namen raziskovalne naloge je bil preveriti možnost uporabe odprtokodnega programa Orange, ki je namenjen podatkovnemu rudarjenju, vizualizaciji podatkov in modeliranju povezave med strukturo in lastnostmi molekul. Najprej sem zbrala podatke, tj. strukture molekul in pripadajoče lastnosti, ki so potrebne za izdelavo modelov. S pomočjo zunanjega mentorja sem s programom Canvas izračunala opisnike, ki so predstavljali neodvisne spremenljivke pri izdelavi modelov, odvisno spremenljivko je predstavljala lastnost. S programom Orange sem izdelala modele za napoved vrelišč alkanov, aromatskih ogljikovodikov in alkoholov ter modela za napoved akutne toksičnosti pesticidov za organizem *Daphnia Magna*. Ugotovila sem, da je delo s programom Orange enostavno, saj omogoča vizualno programiranje, tj. programiranje, ki poteka z zlaganjem in povezovanjem gradnikov na delovni površini grafičnega vmesnika brez pisanja zapletenih algoritmov. Za izdelavo modelov sem uporabila dve metodi: metodo večparametrske linearne regresije in metodo naključnega gozda, za preverjanje napovedne moči modelov pa metodo "izpusti enega". Pri modeliranju vrelišč alkanov sem uporabila tudi testni set spojin, ki ni bil vključen v izdelavo modela. Ugotovila sem, da je za zadovoljiv opis vrelišč potrebno le majhno število opisnikov; na primer pri vreliščih alkanov sem dobila dobre modele s tremi opisniki, tj. relativno molekulsko maso in dva topološka indeka (Randićev, Wienerjev), s katerima opišemo razvejanost molekul alkanov. Podobno dobre rezultate sem dobila tudi za napoved vrelišč aromatov in alkoholov. V skladu s pričakovanji so bili modeli za napoved akutne vodne toksičnosti manj zanesljivi, a vendar dovolj uporabni, da lahko pesticide razvrstimo na manj in bolj toksične.

1 UVOD

1.1 Namen

Namen raziskovalne naloge je ugotoviti, ali nam lahko računalniški program kot je Orange pomaga pri določanju lastnosti spojin. V ta namen smo izbrali dve lastnosti: temperaturo vrelišča in toksičnost. Temperaturo vrelišč bomo preučevali na skupini alkanov, aromatskih ogljikovodikov in alkoholov. Toksičnost pa bomo obravnavali na primeru vplivov nekaterih pesticidov na vodne bolhe.

1.2 Hipoteze

Pred uporabo programa Orange za napoved lastnosti smo postavili naslednje hipoteze:

- Na podlagi znanih vrelišč alkanov, aromatskih ogljikovodikov in alkoholov ter ustreznih molekulskih opisnikov lahko z uporabo računalniškega programa Orange napovemo vrelišča teh spojin.
- Vrelišča spojin so odvisna od velikosti molekul in njihove razvejenosti, kar bo razvidno tudi iz računalniškega modela za napoved vrelišča.
- S primernimi molekulskimi opisniki bo možno napovedati tudi vodno toksičnost (vrednost pLC_{50}) izbrane skupine pesticidov.

1.3 Raziskovalne metode

Na računalnik smo naložili program Orange. Program Orange smo uporabili za izdelavo računalniških modelov za napoved izbrane lastnosti določene skupine spojin. Spoznali smo zapis kod SMILES molekul različnih spojin. Na osnovi kod SMILES smo z ustreznimi programi računali nekatere molekulske opisnike. Modele za napoved lastnosti snovi smo izdelali z metodama linearne regresije in naključnega gozda. Pri obeh metodah smo v analizo večinoma zajeli vse podatke oziroma smo uporabili način vzorčenja »izpusti enega« (ang. leave one out).

2 TEORETIČNI DEL

2.1 Fizikalne in druge lastnosti snovi

Poznamo več fizikalnih lastnosti snovi. Mednje sodijo temperatura vrelišča, temperatura tališča, gostota, električna in toplotna prevodnost, topnost v polarnih in nepolarnih topilih, barva… Na fizikalne in druge lastnosti snovi vpliva več dejavnikov. Ti dejavniki so lahko zunanji dejavniki, kot so npr. tlak, temperatura, prostornina...

Temperatura vrelišča je odvisna od zunanjega tlaka. Tekočina zavre, ko je njen parni tlak enak zunanjemu tlaku. Parni tlak je tlak, ki ga povzročajo molekule tekočine nad tekočino. Vrelišče je po dogovoru temperatura, pri kateri tekočina, pri zunanjem tlaku 100,325 kPa, zavre.

Topnost snovi lahko opišemo s porazdelitvenim koeficientom *P*. Porazdelitveni koeficient je razmerje med množinskima koncentracijama neke snovi (topljenca), ki se ob stiku dveh topil, ki se med seboj ne mešata, razporedi med obema topiloma. Običajno je eno topilo oktan-1-ol, drugo topilo pa voda.

$P = \frac{c_1}{c_2}$, kjer sta c_1 množinska koncentracija v oktan-1-olu in c_2 množinska koncentracija v vodi. Pogosto porazdelitveni koeficient podajamo v obliki *logP*.

Na lastnosti snovi vplivajo predvsem vezi oz. privlačne sile med delci snovi. Pri snoveh zgrajenih iz molekul na lastnosti snovi vplivajo predvsem privlačne sile med molekulami. Jakost teh privlačnih sil je odvisna od velikosti molekul (t.j. relativne molekulske mase), strukture molekul (razvejene ali nerazvejene molekule) in vrste ter števila funkcionalnih skupin.

Med nepolarnimi molekulami delujejo disperzijske sile, med polarnimi molekulami pa disperzijske in orientacijske sile. Med molekulami alkoholov, ki imajo v molekuli eno ali več hidroksilnih skupin, pa ob disperzijskih in orientacijskih silah delujejo tudi vodikove vezi.

2.1.1 Vrelišča alkanov

Molekule alkanov so nepolarne molekule, med katerimi delujejo disperzijske sile. Jakost disperzijskih sil je odvisna od velikosti in razvejenosti molekul alkanov. Vrelišča podobno razvejenih alkanov naraščajo z večanjem mase molekul. Med večjimi molekulami so prisotne močnejše disperzijske vezi, torej so vrelišča večja. Primerjava vrelišč alkanov z enako maso molekul in različno razvejenostjo molekul nam pove, da se vrelišča večajo z večanjem

razvejenosti molekul. Stična površina med bolj razvejenimi molekulami je manjša, zato med njimi delujejo šibkejše disperzijske vezi.

Molekule alkanov imajo različno število strukturnih izomerov (glej tabelo 1). Večja je molekulska formula alkana, večje je število izomerov.

Tabela 1: Število strukturnih izomerov molekul alkanov

Molekulska formula	CH_4	C_2H_6	C_3H_8	C_4H_{10}	C_5H_{12}
Št. Izomerov	1	1	1	2	3
Molekulska formula	C_6H_{14}	C_7H_{16}	C_8H_{18}	C_9H_{20}	$C_{10}H_{22}$
Št. izomerov	5	9	18	35	75

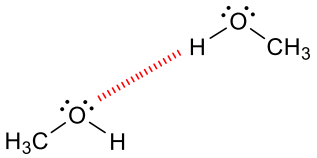
2.1.2 Vrelišča aromatskih ogljikovodikov

Osnovni aromatski ogljikovodik je benzen. Aromatski ogljikovodiki so spojine z enim ali več benzenovimi obroči. V raziskovalni nalogi sem obravnavala policiklične aromatske ogljikovodike (glej prilogo 1). Te spojine, ki jih lahko označimo s kratico PAH (iz angl. polycyclic aromatic hydrocarbons), so v vesolju zelo pogoste, najdemo jih v premogu in katranu, nastajajo tudi pri gorenju organskih snovi. Te spojine so škodljive za človeka (povzročajo raka, bolezni srca in ožilja) in okolje (onesnaženost zraka in prsti).

Molekule aromatskih ogljikovodikov se nepolarne molekule, zato tudi med njimi, tako kot med molekulami alkanov, delujejo disperzijske sile. Med večjimi molekulami aromatskih ogljikovodikov so močnejše disperzijske vezi in je vrelišče večje. Na vrelišče aromatskih ogljikovodikov vplivajo tudi alkilne skupine vezane na benzenove obroče ter način povezovanja benzenovih obročev v molekuli.

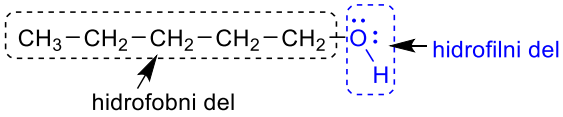
2.1.3 Vrelišča alkoholov

Molekula alkoholov so lahko polarne ali pa šibko polarne. Manjše molekule alkoholov imajo kratek nepolarni del, na katerega je vezana močno polarna hidroksilna skupina, zato so polarne. Med temi molekulami delujejo disperzijske in orientacijske sile, najmočnejše vezi med njimi pa so vodikove vezi.



Slika 1: Strukturni formuli molekul metanola in vodikova vez med molekulama

Daljše molekule alkoholov imajo daljši nepolarni del, zato so manj polarne od krajših molekul alkoholov. Tudi med daljšimi molekulami alkoholov so prisotne disperzijske in orientacijske sile ter vodikove vezi, a prispevek disperzijskih sil postane pomembnejši.



Slika 2: Strukturna formula pental-1-ola z označenim nepolarnim delom molekule in polarnim delom molekule (hidroksilno skupino)

Vpliv razvejenosti molekul alkoholov na njihova vrelišča je podoben kot je pri alkanih. Med bolj razvejenimi molekulami alkoholov delujejo šibkejše molekulske vezi.

2.1.4 Vodna toksičnost pesticidov za organizem Daphnia magna

Daphnia magna (vodna bolha) je vodni organizem, ki ga pogosto uporabljajo za testiranje vodne toksičnosti snovi. Vodne bolhe so zelo občutljive na kontaminacijo voda, zato jih pogosto uporabljamo kot indikator za kemijsko onesnaženost voda.

Slika 3: Vodna bolha

Vodno toksičnost običajno podajamo s povprečno smrtno koncentracijo LC_{50}, to je množinska koncentracija toksične snovi v vodi, ki v 14 dneh povzroči smrt 50 % preiskovanih

organizmov. Ker so vrednosti toksičnosti običajno v širokem koncentracijskem območju uvedemo pLC_{50}, ki je negativen logaritem vrednosti LC_{50}.

Dandanes se bolje, a verjetno še premalo, zavedamo vplivov na okolje, ki jih povzročamo z uporabo različnih kemikalij (pesticidi, gnojila, industrijski izpusti, itd.). Ker je kemijskih spojin, ki so potencialni onesnaževalci voda, ogromno in je ugotavljanje toksičnosti z uporabo organizmov zamudno in pogosto etično sporno, pogosto namesto testiranja na organizmih uporabljamo modele. V zadnjih letih je ob registraciji novih spojin, ki jo v EU ureja REACH (Registration, Evaluation and Autorisation of Chemicals – Regulation (EC) No 1907/2006), potrebno predložiti tudi oceno toksičnosti. Modeli, ki jih uporabijo za napoved toksičnosti, morajo ustrezati naslednjim smernicam:

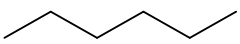
1. model mora biti znanstveno verodostojen (smernice OECD (Organisation for Economic Co-operation and Development));
2. spojina, kateri določamo toksičnost, mora biti v mejah področja uporabe;
3. rezultati morajo biti primerni za razvrščanje in označevanje in/ali oceno tveganja;
4. zagotovljena mora biti ustrezna dokumentacija o uporabljenih metodah.

2.2 Kode SMILES

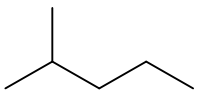
Koda SMILES (iz angl. Simplified Molecular Input Line Entry Specification) je linearni zapis strukture molekule. V tem grafičnem zapisu molekule izpustimo vodikove atome. Atome ostalih elementov, ki zgradijo organsko molekulo, zapišemo z njihovimi simboli. Za označevanje mest, kjer se glavna veriga razveja, uporabimo oklepaje. Mesta, kjer so obroči prekinjeni, se označujejo s številkami, ki omogočajo povezovanje teh atomov. Enojnih vezi med atomi v molekuli ne zapišemo, za dvojne vezi uporabimo =, za trojne vezi pa #. Na spletu je dostopnih več programskih orodij za generiranje kode SMILES.

PRIMERI:

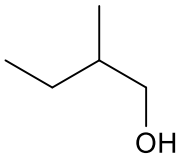
Heksan:

SMILES koda: CCCCCC Skeletna formula: 

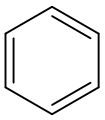
2-metilpentan:

SMILES koda: CC(C)CCC Skeletna formula: 

2-metilbutan-1-ol:

SMILES koda: CCC(C)CO Skeletna formula: 

Benzen:

SMILES koda: c1=cc=cc=cc1 Skeletna formula: 

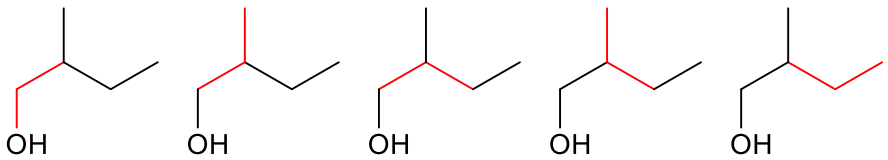
2.3 Molekulski opisniki - topološki indeksi

Razvejeno zgradbo molekul lahko opišemo s ti. topološkimi indeksi. V moji raziskovalni nalogi bom uporabila dva topološka indeksa: Wienerjev in Randićev indeks. Oba indeksa lahko izračunamo sami s kalkulatorjem, lahko pa uporabimo ustrezne računalniške programe.

2.3.1 Wienerjev indeks

Za izračun Wienerjevega indeksa moramo poznati skeletno ali racionalno formulo molekule, za katero računamo indeks. V skeletni formuli napišemo vse vezi med atomi, razen vezi z atomi vodika. Ko računamo Wienerjev indeks, si predstavljamo, da se sprehajamo po skeletni formuli molekule in ugotavljamo, koliko različnih poti lahko naredimo ter po kolikih vezeh smo se sprehodili pri posamezni poti. Pri izračunu Wienerjevega indeksa seštejemo vse različne možne poti.

Podrobneje bom izračun Wienerjevega indeksa razložila na molekuli 2-metilbutan1-ola. Za molekulo 2-metilbutan-1-ol ugotovimo, da lahko opravimo 5 poti, pri katerih se sprehodimo po 1 vezi, 5 poti, pri katerih se sprehodimo po 2 vezeh, 4 poti, pri katerih se sprehodimo po 3 vezeh in 1 pot, pri kateri se sprehodimo po 4 vezeh. Potem vse to seštejemo in dobimo vrednost indeksa. $WI = 5 \times 1 + 5 \times 2 + 4 \times 3 + 1 \times 4 = 31$. Wienerjev indeks za 2-metilbutan-1-ol ima vrednost 31.

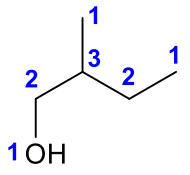


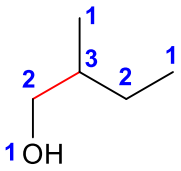
Slika 4: Prikaz vseh poti preko dveh vezi (označenih z rdečo) na skeletni formuli molekule 2-metilbutan-1-ola

Wienerjev indeks je primeren opisnik razvejenosti molekule, saj ta indeks razlikuje med različnimi strukturnimi izomeri. Npr. relativna molekulska masa vseh možnih strukturnih izomerov neke spojine je enaka (npr. ni razlike med butanom in metilpropanom).

2.3.2 Randićev indeks

Tako kot za izračun Wienerjevega indeksa je tudi tukaj potrebno znanje zapisovanja skeletnih formul za molekule, katerih indeks želimo izračunati. Randićev indeks izračunamo tako, da na skeletni formuli označimo za vsak ogljikov atom, koliko vezi tvori z ostalimi atomi, razen z vodikovimi atomi. Potem pa za vsako vez posebej ugotovimo med katerima atomoma je in koliko vezi tvori vsak atom posebej. Nato izračunamo za vsako vez izraz $\frac{1}{\sqrt{d_i \times d_j}}$, pri čemer sta d_i – število vezi, ki jih tvori prvi atom, d_j – število vezi, ki jih tvori drugi atom. Vrednost Randićevega indeksa je vsota takih izrazov za vse vezi v neki molekuli. Računanje indeksa bom ponovno razložila na primeru molekule 2-metilbutan-1-ola.

a) 

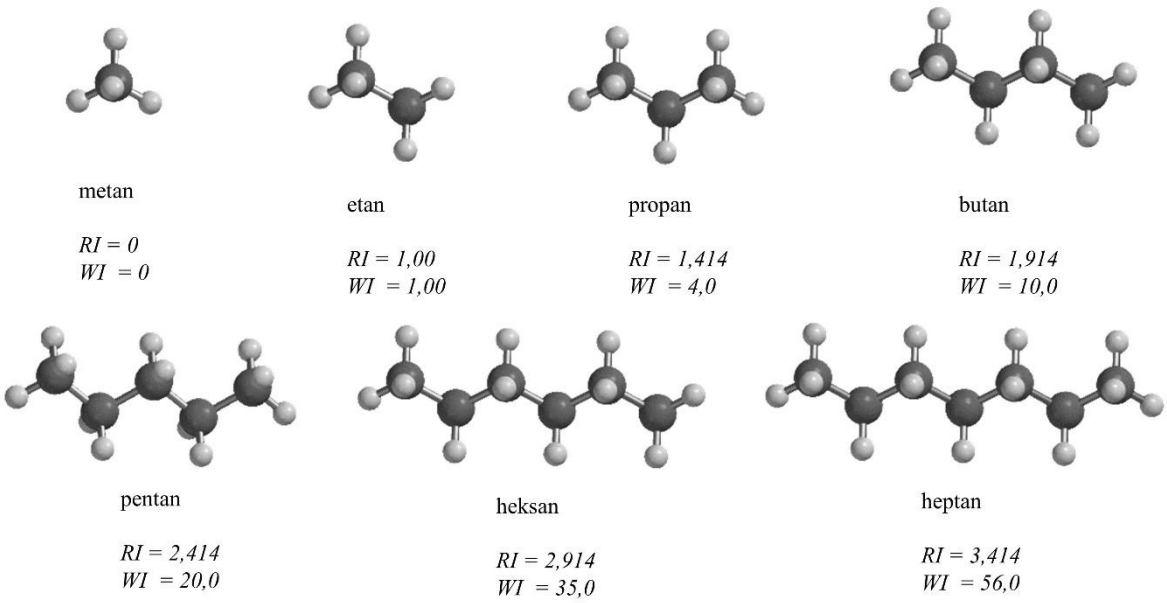
b) 

Slika 5: a) Skeletna formula molekule 2-metilbutan-1-ola z označenim številom vezi na vsakem atomu. b) Za vez označeno z rdečo (prvi ogljikov atom tvori dve vezi, drugi ogljikov atom pa tri) je izraz $\frac{1}{\sqrt{d_i \times d_j}}$ enak $\frac{1}{\sqrt{2 \times 3}}$

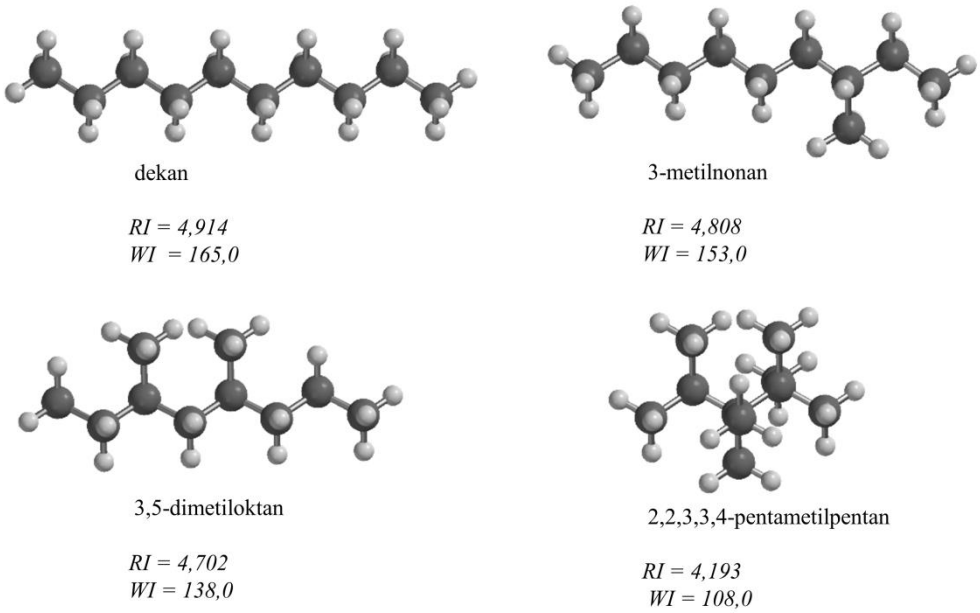
Formula za izračun Randićevega indeksa za molekulo 2-metilbutan-1-ol se glasi tako:

$$RI = \frac{1}{\sqrt{1 \times 2}} + \frac{1}{\sqrt{2 \times 3}} + \frac{1}{\sqrt{3 \times 1}} + \frac{1}{\sqrt{3 \times 2}} + \frac{1}{\sqrt{2 \times 1}} = 2{,}8081$$

Vrednost topoloških opisnikov (*WI* in *RI*) se spreminja z velikostjo in razvejenostjo molekul. Na sliki 6 je prikazano spreminjanje obeh opisnikov pri nerazvejenih alkanih, na sliki 7 pa njuno spreminjanje z večanjem razvejenosti molekul alkanov. Opazimo lahko, da z večanjem dolžine verige v molekulah nerazvejenih alkanov narašča vrednost obeh topoloških opisnikov, z večanjem razvejenosti molekul pa se njune vrednosti zmanjšujejo.



Slika 6: Vrednosti obeh topoloških indeksov v skupini nerazvejenih alkanov

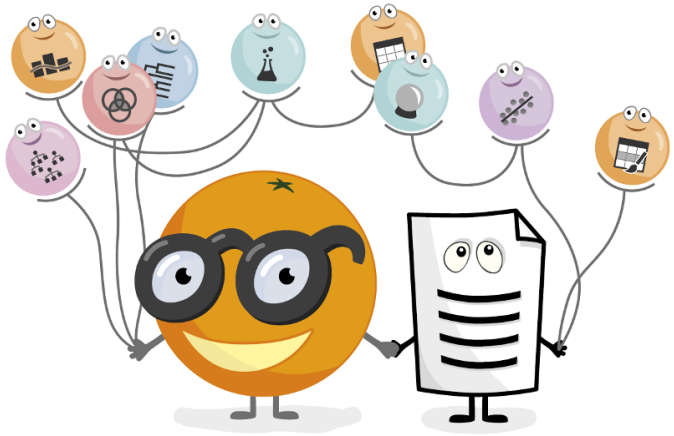


Slika 7: Vrednosti obeh topoloških indeksov štirih strukturnih izomerov z molekulsko formulo $C_{10}H_{22}$

2.4 Program Orange

V okviru raziskovalne naloge sem za izdelavo napovednih modelov uporabila program Orange, ki so ga razvili na Fakulteti za računalništvo Univerze v Ljubljani. Program omogoča enostavno analizo in pregledno vizualizacijo velike množice podatkov na različnih področjih znanosti. Orange lahko uporabimo na primer za analizo človeškega genoma, različne

sociološke in antropološke raziskave, analizo podatkov iz velikega hadronskega trkalnika v Cernu ali analizo posnetkov iz Hubblovega teleskopa. Vizualno programiranje je ena izmed poglavitnih prednosti programa Orange – namesto dolgotrajnega pisanja zapletenih programov in skript lahko v Orange-u naredimo analizo podatkov tako, da uporabimo gradnike, ki jih smiselno povežemo v shemo.



Slika 8: Orange – Odprtnokodno strojno učenje in vizualizacija podatkov za začetnike in strokovnjake

Raziskala sem, kako lahko program Orange uporabim za izdelavo modelov za napoved nekaterih lastnosti molekul. Sam postopek izdelave modela je sestavljen iz več korakov:

1. Najprej sem v literaturi in/ali na spletu poiskala podatke o izbranih molekulah. Za izdelavo napovednega modela sem potrebovala tako podatke o molekulski strukturi (npr. kode SMILES), kot informacije o lastnosti, ki sem jo želela napovedati.
2. Nato sem uporabila molekulske strukture (kode SMILES) za izračun molekulskih opisnikov (ti. deskriptorjev). Opisnike, ki sem jih uporabila v svojem delu, smo izračunali s programom Canvas, ki je del programskega paketa Schrodinger Suite 2018.4.
3. Opisnike in lastnost za vsako izmed molekul v skupini sem zbrala v Excelovi preglednici, ki sem jo nato izvozila v obliki csv datoteke (iz angl. comma separated value, slov. vrednosti, ločene z vejico).
4. Podatke sem nato uvozila v shemo, ki sem jo naredila s programom Orange.
5. Izračunane modele sem nato s pomočjo mentorjev analizirala.

Orange je program, ki na osnovi opisnikov izbrane skupine molekul in njihovih lastnosti izračuna iskano lastnost skupine molekul ali posamezne molekule. Je zelo uporaben

predvsem, ko želimo raziskati lastnosti večje skupine spojin. Za določitev lastnosti vsake izmed spojin je potrebno izvesti več poskusov, s programom Orange pa lahko v skupini strukturno podobnih lastnosti te lastnosti napovemo. V program Orange preprosto vnesemo imena spojin, nekaj njihovih lastnosti ter opisnikov in program Orange na podlagi tega izračuna oz. napove lastnost, ki nas zanima.

2.5 Modeli za napoved lastnosti snovi

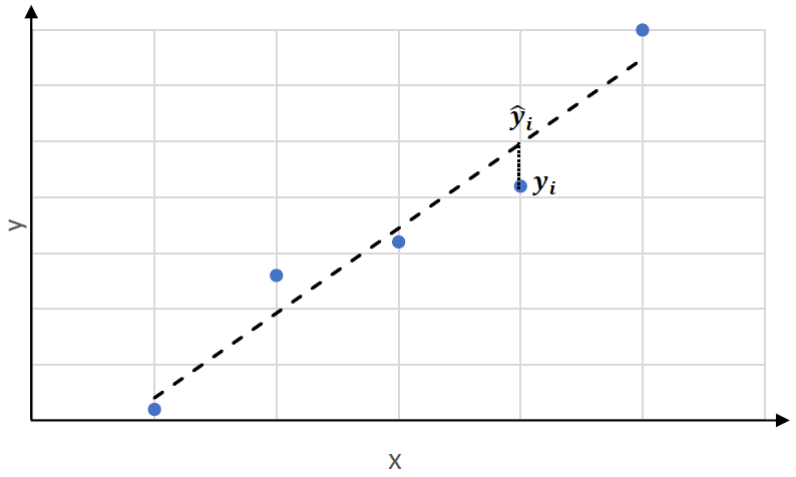
Modele za napoved lastnosti snovi lahko izdelamo na različne načine. V moji raziskovalni nalogi sem se po nasvetu zunanjega mentorja odločila za uporabo dveh metod: linearna regresija in naključni gozd.

2.5.1 Linearna regresija

Pri izdelavi modela za napoved lastnosti lahko uporabimo metodo linearne regresije *LR*. Pri tej metodi napovedane vrednosti lastnosti y ($\hat{y}_i$) v odvisnosti od x izrazimo z regresijsko premico:

$\boldsymbol{\hat{y}_i = bx + a}$, kjer je b regresijski koeficient, a pa regresijska konstanta.

Na grafu najprej narišemo vse točke, ki predstavljajo dejansko vrednost spremenljivke y (npr. temperature vrelišča) v odvisnosti od x (npr. molske mase spojine ali katerega od topoloških indeksov molekul), ti. razsevni ali korelacijski diagram. Nato med točkami narišemo regresijsko črto – tj. premico, ki je v povprečju najmanj oddaljena od točk (kriterij je minimalna vsota kvadratov vseh odmikov/odklonov točk od premice (y_i-$\hat{y}_i$)). Premico izračunamo tako, da je srednja kvadratna napaka MSE najmanjša.



Slika 9: Prikaz vrednosti spremenljivke y od spremenljivke x. S točkami so označene dejanske vrednosti spremenljivke y(y_i) , na premici označeno črtkano pa so z linearno regresijo napovedane lastnosti spremenljivke y ($\hat{y}_i$).

Spodaj so zapisane enačbe za računanje statističnih vrednosti MSE, RMSE, MAE in R^2, v katerih so y_i dejanska vrednost spremenljivke y, $\hat{y}_i$ napovedana vrednost spremenljivke y in $\bar{y}$ povprečna vrednost spremenljivke y.

Povprečno vrednost spremenljivke y izračunamo z uporabo enačbe:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Srednjo kvadratno napako MSE (iz angl. mean squared error) izračunamo z uporabo enačbe:

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

Celotno napako (koren srednje kvadratne napake) RMSE (iz angl. root mean squared error) izračunamo z uporabo enačbe:

$$RMSE = \sqrt{MSE}$$

Srednjo absolutno napako MAE (iz angl. mean absolute error) izračunamo z uporabo enačbe

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Koeficient determinacije R^2 pa izračunamo z uporabo enačbe:

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

Ta koeficient nam pove, kolikšen delež vseh napovedanih vrednosti je pojasnjen z regresijo.

In kdaj je model za napoved lastnosti dober?

Čim večja je vrednost koeficienta determinacije R^2, tem boljša je napovedna sposobnost modela. Čim manjše so vrednosti MSE, RMSE in MAE, tem boljši je model.

MSE je običajno večji od MAE, v primeru dejanskih vrednosti y, ki so zelo daleč od napovedanih vrednosti y, je MSE precej večji od MAE. Celotna napaka RMSE in srednja absolutna napaka MAE imata enake enote kot lastnost y. Ti dve napaki sta povezani z odmiki napovedanih vrednosti neke lastnosti od pravih vrednosti te lastnosti.

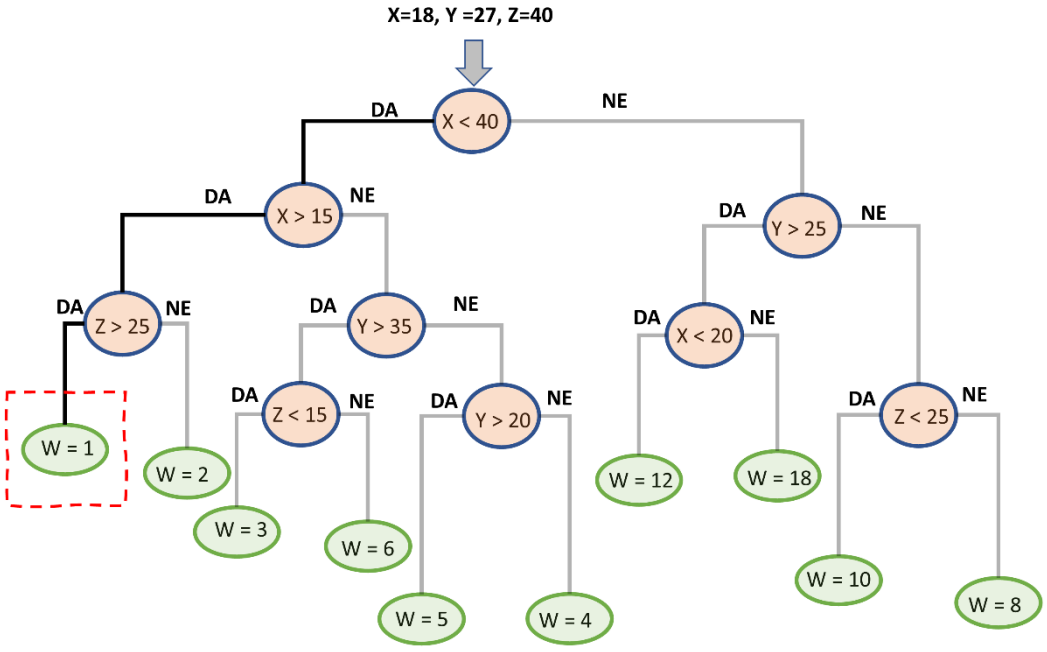
2.5.2 Naključni gozd (angl. Random Forest)

Pri izdelavi modela za napoved lastnosti lahko uporabimo metodo naključnega gozda. Za razumevanje te metode je potrebno najprej razložiti, kaj so odločitvena drevesa.

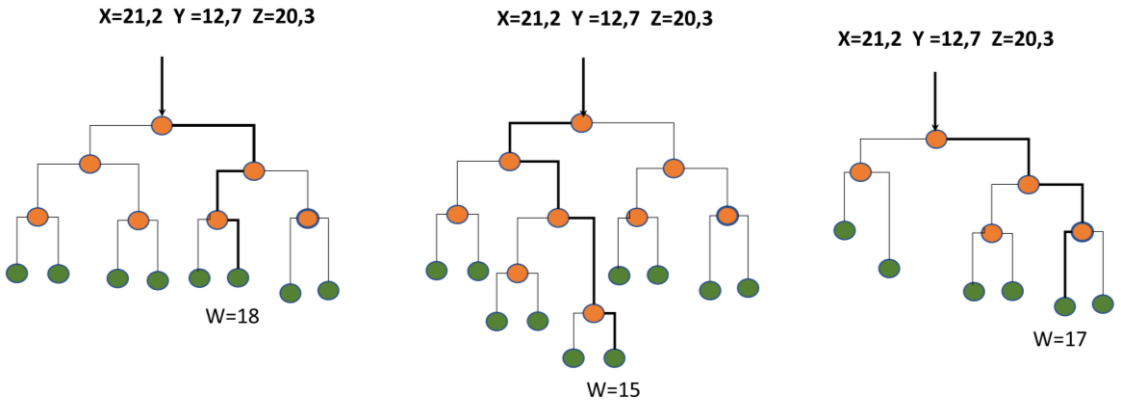
Odločitvena drevesa so hierarhične podatkovne strukture, ki so sestavljene iz notranjih in zunanjih vozlišč, ki v analogiji z drevesom predstavljajo veje in liste. Vsak list drevesa vsebuje napovedi vrednosti ciljne spremenljivke. Če ima ciljna spremenljivka zvezne vrednosti, govorimo o regresijskem drevesu, v primeru, da so ciljne vrednosti nezvezne, pa o klasifikacijskem drevesu.

Odločitveno drevo ponazarja graf odločitev in možnih posledic in opisuje odnos med odvisno (v našem primeru lastnost molekule) in neodvisnimi (molekulski opisniki) spremenljivkami. Interpretacija dreves je enostavna in lahko razumljiva (glej primer odločitvenega drevesa na sliki 10).

Naključni gozd predstavlja zbirko dreves (glej sliko 11). Drevesa se med seboj razlikujejo in vsako drevo, ki predstavlja svoj model, vrne drugačno vrednost ciljne spremenljivke. Končno vrednost ciljne spremenljivke dobimo z računanjem povprečne vrednosti te spremenljivke po vseh odločitvenih drevesih v gozdu.



Slika 10: Primer odločitvenega drevesa, pri katerem so X, Y in Z neodvisne spremenljivke, W pa je odvisna spremenljivka.



$$\overline{W} = \frac{W_1 + W_2 + W_3}{3} = \frac{18 + 15 + 17}{3} = 16,66$$

Slika 11: Primer naključnega gozda s tremi odločitvenimi drevesi

Tudi pri modelih naključnega gozda *RF* računamo statistične vrednosti *MSE*, *RMSE*, *MAE* in R^2.

2.6 Načini preverjanja napovedne moči modela za napoved lastnosti snovi

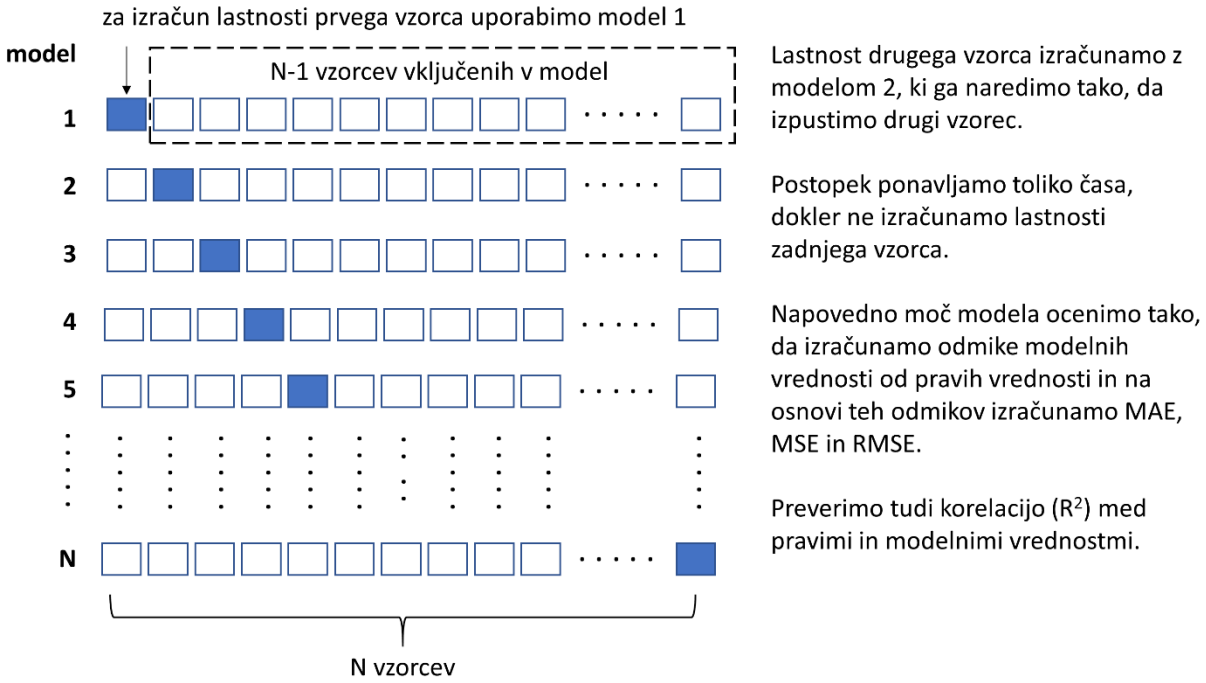
Modele dobljene z zgoraj opisanimi metodami pogosto uporabljamo za napoved lastnosti drugih spojin (v splošnem vzorcev), zato je potrebno pred samo uporabo modela preveriti njegovo napovedno moč. Preverjanje napovedne moči modelov lahko naredimo na več načinov. V moji raziskovalni nalogi sem uporabila dva načina, ki sta opisana v nadaljevanju.

2.6.1 Zunanje preverjanje napovedne moči modela s testno skupino vzorcev

Prvi način preverjanja napovedne moči modela je uporaba zunanje testne skupine spojin. Nekaj spojin (recimo 20 %) izločimo iz skupine, preostale uporabimo za izdelavo modela, nato model uporabimo za izračun lastnosti spojin znotraj testne skupine spojin. Za vsako izmed spojin v testni skupini preverimo, koliko izračunana vrednost odstopa od prave, ter uporabimo statistične metode za izračun *MSE*, *RMSE*, *MAE* in R^2.

2.6.2 Metoda "izpusti enega" (iz angl. leave one out LOO)

Drugi način preverjanja je metoda "izpusti enega" (angl. leave one out), ki se pogosto uporablja za preverjanje notranje napovedne moči modela. Postopek opisan na sliki 12, predstavlja metodo "izpusti enega" za N vzorcev. Iz celotne skupine spojin (vzorcev) postopoma izločamo po eno spojino (en vzorec), zgradimo model in napovemo lastnost iz modela izpuščene spojine. Na koncu napovedno moč ocenimo, tako da izračunamo odmike izračunanih od pravih vrednosti, na osnovi teh odmikov lahko izračunamo *MAE*, *MSE*, *RMSE* in R^2 .



Slika 12: Shematski prikaz metode "izpusti enega".

2.7 Oznake, uporabljene v raziskovalni nalogi

V raziskovalni nalogi sem uporabljala nekatere oznake in kratice, te so zbrane v tabeli 2.

Tabela 2: Oznake in njihov pomen

Oznaka	Pomen oznake
***T_v*, *BP* (ang. boiling point)**	Temperatura vrelišča
c	Množinska koncentracija
LC_{50}	Povprečna smrtna koncentracija

WI	Wienerjev indeks
RI	Randićev indeks
MSE	Srednja vrednost kvadratov napak
LR	Linearna regresija
RMSE	Koren srednje vrednosti kvadratov napak
MAE	Srednja vrednost absolutnih napak
R^2	Koeficient determinacije
M_r	Relativna molekulska masa
RF	Metoda naključnega gozda
***MW* (ang. molecular weight)**	Molekulska masa
LOO	Metoda izpusti enega
Δ_{LR}	Odmik linearne regresije
Δ_{RF}	Odmik metode naključnega gozda
P	Porazdelitveni koeficient
pLC_{50}	Negativni logaritem LC_{50}
logP	Logaritem porazdelitvenega koeficienta
HBD	Donor vodikove vezi
HBA	Akceptor vodikove vezi
RB	Število vrtljivih vezi

3 EKSPERIMENTALNI DEL

3.1 Uporaba programa Orange za napoved vrelišč alkanov

V tabeli 3 so zbrani podatki, ki sem jih potrebovala za izdelavo modelov. Med te podatke sodijo za vsako spojine v skupini spojin: koda SMILES, ki jo potrebujemo za izračun opisnikov, ime spojine, trije opisniki (relativna molekulska masa, Randićev indeks in Wienerjev indeks) ter temperatura vrelišča, tj. lastnost, ki jo želimo modelirati. Za nekaj spojin sem topološka opisnika izračunala sama z uporabo kalkulatorja, ostale opisnike je izračunal zunanji mentor s Phytonovo skripto, ki jo je napisal v ta namen. Podatke o vreliščih alkanov sem našla v znanstvenem članku objavljenem v J. Chem. Soc. Faraday Trans.

Tabela 3: Vhodni podatki izbrane skupine alkanov za izdelavo modelov za napoved vrelišč

Koda SMILES	Ime spojine	M_r	*RI*	*WI*	T_v [°C]
C	METAN	16,031	0	0	-164
CCC	PROPAN	44,063	1,414	4,000	-42,1
CC(C)C	2-METILPROPAN	58,078	1,732	9,000	-11,7
CCCC	BUTAN	58,078	1,914	10,000	-0,5
CC(C)(C)C	2,2-DIMETILPROPAN	72,094	2,000	16,000	9,5
CC(C)CC	2-METILBUTAN	72,094	2,270	18,000	27,8
CCCCC	PENTAN	72,094	2,414	20,000	36,1
CC(C)(C)CC	2,2-DIMETILBUTAN	86,110	2,561	28,000	49,7
CC(C)C(C)C	2,3-DIMETILBUTAN	86,110	2,643	29,000	58
CC(C)CCC	2-METILPENTAN	86,110	2,770	32,000	60,3
CCCCCC	HEKSAN	86,110	2,914	35,000	69
CC(C)(C)C(C)C	2,2,3-TRIMETILBUTAN	100,125	2,943	42,000	80,9
CC(C)(C)CCC	2,2-DIMETILPENTAN	100,125	3,061	46,000	79,2
CCC(C)(C)CC	3,3-DIMETILPENTAN	100,125	3,121	44,000	86,1
CCC(C)C(C)C	2,3-DIMETILPENTAN	100,125	3,181	46,000	89,8
CC(C)CC(C)C	2,4-DIMETILPENTAN	100,125	3,126	48,000	80,5
CC(C)CCCC	2-METILHEKSAN	100,125	3,270	52,000	90
CCC(C)CCC	3-METILHEKSAN	100,125	3,308	50,000	92
CC(C)(C)C(C)(C)C	2,2,3,3-TETRAMETILBUTAN	114,141	3,250	58,000	106,5
CC(C)(C)C(C)CC	2,2,3-TRIMETILPENTAN	114,141	3,481	63,000	110
CCC(C)(C)C(C)C	2,3,3-TRIMETILPENTAN	114,141	3,504	62,000	114,7
CC(C)(C)CC(C)C	2,2,4-TRIMETILPENTAN	114,141	3,417	66,000	99,2
CC(C)(C)CCCC	2,2-DIMETILHEKSAN	114,141	3,561	71,000	106,8
CCC(C)(C)CCC	3,3-DIMETILHEKSAN	114,141	3,621	67,000	112
CCC(C)C(C)CC	3,4-DIMETILHEKSAN	114,141	3,719	68,000	118,2
CC(C)C(C)C(C)C	2,3,4-TRIMETILPENTAN	114,141	3,553	65,000	113,4
CC(C)C(C)CCC	2,3-DIMETILHEKSAN	114,141	3,681	70,000	115,6
CCC(C)CC(C)C	2,4-DIMETILHEKSAN	114,141	3,664	71,000	109,4
CC(C)CCC(C)C	2,5-DIMETILHEKSAN	114,141	3,626	74,000	109
CC(C)CCCCC	2-METILHEPTAN	114,141	3,770	79,000	117,6
CCC(C)CCCC	3-METILHEPTAN	114,141	3,808	76,000	118
CCCC(C)CCC	4-METILHEPTAN	114,141	3,808	75,000	117,7
CCCCCCCC	OKTAN	114,141	3,914	84,000	125,7
CCC(C)(C)C(C)(C)C	2,2,3,3-TETRAMETILPENTAN	128,157	3,811	82,000	140,27
CC(C)(C)C(C)CCC	2,2,3-TRIMETILHEKSAN	128,157	3,981	92,000	131,7
CC(C)(C)CC(C)CC	2,2,4-TRIMETILHEKSAN	128,157	3,955	94,000	133,83
CCC(C)(C)C(C)CC	3,3,4-TRIMETILHEKSAN	128,157	4,042	88,000	140,5
CC(C)C(C)(C)C(C)C	2,3,3,4-TETRAMETILPENTAN	128,157	3,887	84,000	141,5
CC(C)C(C)(C)CCC	2,3,3-TRIMETILHEKSAN	128,157	4,004	90,000	137,7
CC(C)C(C)C(C)CC	2,3,4-TRIMETILHEKSAN	128,157	4,091	92,000	141,6
CC(C)(C)CC(C)(C)C	2,2,4,4-TETRAMETILPENTAN	128,157	3,707	88,000	122,7
CCC(C)(C)CC(C)C	2,4,4-TRIMETILHEKSAN	128,157	3,977	92,000	126,5

CC(C)(C)CCCCC	2,2-DIMETILHEPTAN	128,157	4,061	104,000	132,7
CCC(C)(C)CCCC	3,3-DIMETILHEPTAN	128,157	4,121	98,000	137,3
CCCC(C)(C)CCC	4,4-DIMETILHEPTAN	128,157	4,121	96,000	135,2
CCCC(C)C(C)CC	3,4-DIMETILHEPTAN	128,157	4,219	98,000	140,6
CC(C)C(C)CC(C)C	2,3,5-TRIMETILHEKSAN	128,157	4,037	96,000	136,73
CC(C)C(C)CCCC	2,3-DIMETILHEPTAN	128,157	4,181	102,000	140,5
CCCC(C)CC(C)C	2,4-DIMETILHEPTAN	128,157	4,164	102,000	133,5
CCC(C)CCC(C)C	2,5-DIMETILHEPTAN	128,157	4,164	104,000	133,8
CCC(C)CC(C)CC	3,5-DIMETILHEPTAN	128,157	4,202	100,000	136
CC(C)CCCC(C)C	2,6-DIMETILHEPTAN	128,157	4,126	108,000	135,2
CC(C)CCCCCC	2-METILOKTAN	128,157	4,270	114,000	142,8
CCC(C)CCCCC	3-METILOKTAN	128,157	4,308	110,000	143,3
CCCC(C)CCCC	4-METILOKTAN	128,157	4,308	108,000	143
CCCCCCCCC	NONAN	128,157	4,414	120,000	151,77
CC(C)C(C)(C)C(C)(C)C	2,2,3,3,4-PENTAMETILPENTAN	142,172	4,193	108,000	166,05
CC(C)(C)C(C)(C)CCC	2,2,3,3-TETRAMETILHEKSAN	142,172	4,311	115,000	158
CC(C)(C)C(C)C(C)CC	2,2,3,4-TETRAMETILHEKSAN	142,172	4,392	118,000	168
CCC(C)(C)C(C)(C)CC	3,3,4,4-TETRAMETILHEKSAN	142,172	4,371	111,000	170,5
CC(C)(C)C(C)C(C)(C)C	2,2,3,4,4-PENTAMETILPENTAN	142,172	4,155	111,000	159,29
CCC(C)(C)C(C)C(C)C	2,3,4,4-TETRAMETILHEKSAN	142,172	4,415	116,000	162,2
CC(C)(C)C(C)CC(C)C	2,2,3,5-TETRAMETILHEKSAN	142,172	4,337	123,000	148,4
CC(C)(C)C(C)CCCC	2,2,3-TRIMETILHEPTAN	142,172	4,481	130,000	158
CC(C)(C)CC(C)CCC	2,2,4-TRIMETILHEPTAN	142,172	4,455	131,000	159
CCC(C)(C)C(C)CCC	3,3,4-TRIMETILHEPTAN	142,172	4,542	123,000	164
CCC(C)(C)CC(C)CC	3,3,5-TRIMETILHEPTAN	142,172	4,515	126,000	165
CC(C)C(C)(C)C(C)CC	2,3,3,4-TETRAMETILHEKSAN	142,172	4,425	115,000	164,59
CCCC(C)(C)C(C)CC	3,4,4-TRIMETILHEPTAN	142,172	4,542	122,000	164
CCC(C)C(C)C(C)CC	3,4,5-TRIMETILHEPTAN	142,172	4,629	125,000	170
CC(C)C(C)(C)CC(C)C	2,3,3,5-TETRAMETILHEKSAN	142,172	4,360	120,000	153
CC(C)C(C)(C)CCCC	2,3,3-TRIMETILHEPTAN	142,172	4,504	127,000	160,1
CC(C)C(C)C(C)CCC	2,3,4-TRIMETILHEPTAN	142,172	4,591	128,000	169
CCC(C)C(C)CC(C)C	2,4,5-TRIMETILHEPTAN	142,172	4,575	130,000	174
CCC(C)(C)CC(C)(C)C	2,2,4,4-TETRAMETILHEKSAN	142,172	4,268	119,000	153,3
CC(C)(C)CC(C)C(C)C	2,2,4,5-TETRAMETILHEKSAN	142,172	4,327	124,000	148,2
CC(C)(C)CCC(C)CC	2,2,5-TRIMETILHEPTAN	142,172	4,455	134,000	147
CC(C)(C)CCC(C)(C)C	2,2,5,5-TETRAMETILHEKSAN	142,172	4,207	127,000	137,46
CCC(C)(C)CCC(C)C	2,5,5-TRIMETILHEPTAN	142,172	4,477	131,000	152,8
CC(C)(C)CCCC(C)C	2,2,6-TRIMETILHEPTAN	142,172	4,417	139,000	148,2
CC(C)(C)CCCCCC	2,2-DIMETILOKTAN	142,172	4,561	146,000	155
CCC(C)(C)CCCCC	3,3-DIMETILOKTAN	142,172	4,621	138,000	161,2
CCCC(C)(C)CCCC	4,4-DIMETILOKTAN	142,172	4,621	134,000	157,5
CCC(C)C(C)CCCC	3,4-DIMETILOKTAN	142,172	4,719	137,000	163,8
CCCC(C)C(C)CCC	4,5-DIMETILOKTAN	142,172	4,719	135,000	166,3
CC(C)C(C)C(C)C(C)C	2,3,4,5-TETRAMETILHEKSAN	142,172	4,464	121,000	161

CC(C)C(C)CC(C)CC	2,3,5-TRIMETILHEPTAN	142,172	4,575	131,000	164
CCCC(C)(C)CC(C)C	2,4,4-TRIMETILHEPTAN	142,172	4,477	127,000	163
CC(C)C(C)CCC(C)C	2,3,6-TRIMETILHEPTAN	142,172	4,537	136,000	155,7
CC(C)C(C)CCCCC	2,3-DIMETILOKTAN	142,172	4,681	143,000	164,31
CC(C)CC(C)CCCC	2,4-DIMETILOKTAN	142,172	4,664	142,000	166
CCCC(C)CC(C)CC	3,5-DIMETILOKTAN	142,172	4,702	138,000	167
CC(C)CC(C)CC(C)C	2,4,6-TRIMETILHEPTAN	142,172	4,520	135,000	144,8
CCCC(C)CCC(C)C	2,5-DIMETILOKTAN	142,172	4,664	143,000	160
CCC(C)CCCC(C)C	2,6-DIMETILOKTAN	142,172	4,664	146,000	159,7
CCC(C)CCC(C)CC	3,6-DIMETILOKTAN	142,172	4,702	141,000	160
CC(C)CCCCCCC	2-METILNONAN	142,172	4,770	158,000	167
CCC(C)CCCCCC	3-METILNONAN	142,172	4,808	153,000	167,8
CCCC(C)CCCCC	4-METILNONAN	142,172	4,808	150,000	165,7
CCCCC(C)CCCC	5-METILNONAN	142,172	4,808	149,000	165,1
CCCCCCCCCC	DEKAN	142,172	4,914	165,000	174,12

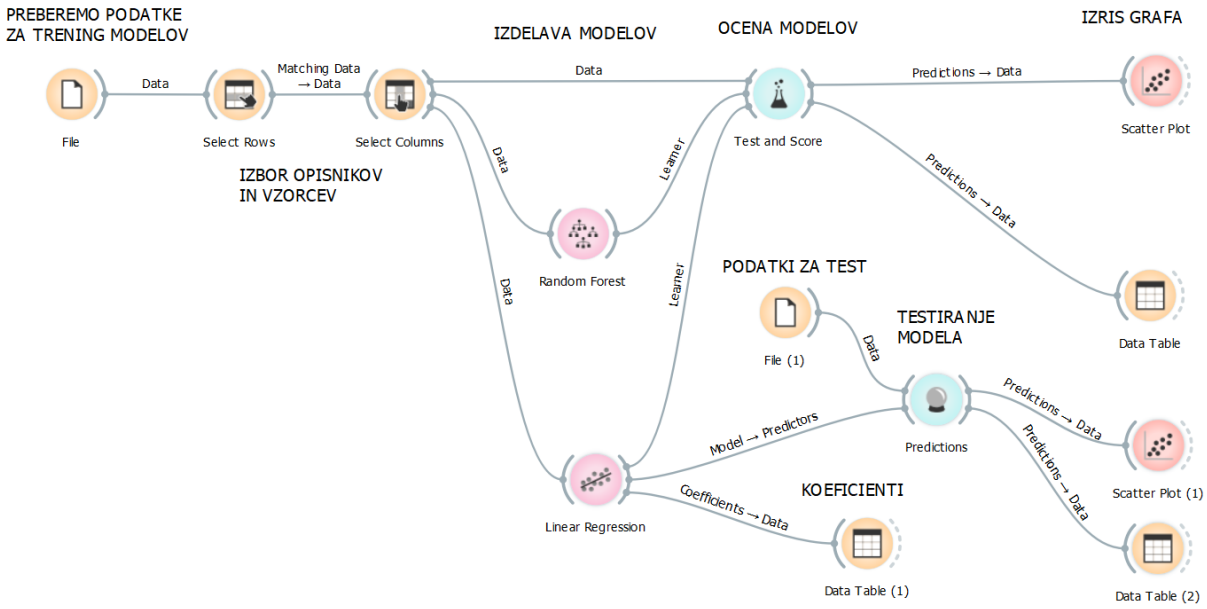
Nekaj alkanov sem izločila iz skupine alkanov. Podatke o tej testni skupini alkanov sem zbrala v tabeli 4. Te podatke sem uporabila za testiranje modelov za napoved vrelišč, ki so izdelani na podlagi podatkov o alkanih, ki so zbrani v tabeli 3.

Tabela 4: Vhodni podatki testne skupine alkanov za napoved vrelišč. Podatki teh alkanov niso bili uporabljeni pri izdelavi modela, s temi podatki le preverimo napovedno moč izdelanih modelov

Koda SMILES	Ime spojine	M_r	*RI*	*WI*	T_v [°C]
CC	ETAN	30,047	1,000	1,000	-88,6
CCC(C)CC	3-METILPENTAN	86,110	2,808	31,000	63,3
CCCCCCC	HEPTAN	100,125	3,414	56,000	98,4
CCC(C)CC(C)C	2,4-DIMETHILHEKSAN	114,141	3,664	71,000	115,6
CC(C)C(C)C(C)(C)C	2,2,3,4-TETRAMETILPENTAN	128,157	3,854	86,000	133
CC(C)CCC(C)(C)C	2,2,5-TRIMETILHEKSAN	128,157	3,917	98,000	124
CCCC(C)CC(C)C	2,4-DIMETILHEPTAN	128,157	4,164	102,000	138
CCCCC(C)CCC	4-METILOKTAN	128,157	4,308	108,000	142,4
CC(C)C(C)CC(C)(C)C	2,2,4,5-TETRAMETILHEKSAN	142,172	4,327	124,000	155,3
CC(C)C(C)C(C)C(C)C	2,3,4,5-TETRAMETILHEKSAN	142,172	4,464	121,000	169,4
CCCC(C)(C)CC(C)C	2,4,4-TRIMETILHEPTAN	142,172	4,477	127,000	153
CCCC(C)C(C)CCC	4,5-DIMETILOKTAN	142,172	4,719	135,000	167
CC(C)CC(C)CC(C)C	2,4,6-TRIMETILHEPTAN	142,172	4,520	135,000	157
CCCC(C)CC(C)CC	3,5-DIMETILOKTAN	142,172	4,702	138,000	162
CC(C)CCCCC(C)C	2,7-DIMETILOKTAN	142,172	4,626	151,000	159,9

Za izdelavo in validacijo napovednih modelov za vrelišča alkanov smo uporabili program Orange. Program omogoča možnost enostavne izdelave modelov brez dolgotrajnega

programiranja. Delo s programom Orange ni zapleteno, večino nalog lahko opravimo z enostavnim zlaganjem in povezovanjem gradnikov na delovni površini. Na spodnji sliki 13 je predstavljena shema narejena s programom Orange. Na sliki opazimo, da so različni gradniki različno obarvani. Tako so gradniki, ki so namenjeni za vnos, iznos in predobdelavo podatkov, obarvani oranžno, gradniki namenjeni za izdelavo modelov (linearna regresija, naključni gozd) so obarvani svetlo vijolično, svetlo modro sta obarvana gradnika, ki smo ju uporabili za oceno kvalitete modelov in testiranje modelov, svetlo rdeče so označena orodja, ki služijo za vizualizacijo modela.



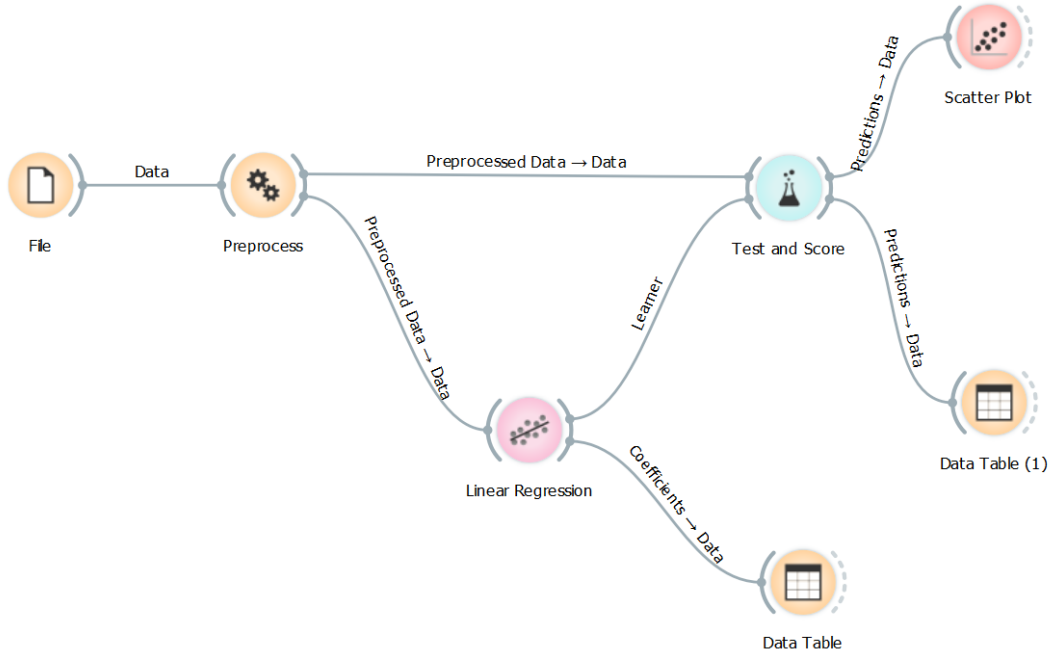
Slika 13: Shema povezanih gradnikov, ki sem jo uporabila pri izdelavi modelov za napoved vrelišč alkanov.

Izdelala sem različne modele za napoved vrelišč alkanov (tri modele z le eno spremenljivko (odvisnost T_v od relativne molekulske mase, odvisnost T_v od vrednosti Wienerjevega indeksa, odvisnost T_v od Randićevega indeksa), model z dvema spremenljivkama (odvisnost T_v od relativne molekulske mase in Randićevega indeksa) in model s tremi spremenljivkami.

3.2 Uporaba programa Orange za napoved vrelišč aromatskih ogljikovodikov

V okviru raziskovalne naloge sem izdelala tudi model za napoved vrelišč policikličnih aromatskih ogljikovodikov. Strukture in vrelišča sem našla v znanstvenem članku objavljenem v reviji *Research Journal of Pharmaceutical, Biological and Chemical Sciences.*

Zaradi nekaterih zapletenih in dolgih imen teh nisem navedla v tabeli 5, v tabeli so napisane le kode SMILES, skeletne formule policikličnih aromatskih ogljikovodikov pa so zbrane v prilogi 1 raziskovalne naloge. V tabeli so zbrane tudi vrednosti molekulskih opisnikov, eksperimentalna vrelišča ter vrelišča dobljena na osnovi dveh modelov. Molekulske opisnike za to skupino spojin je izračunal zunanji mentor s svojim programom. Za izdelavo in analizo modelov sem uporabila program Orange (glej shemo na sliki 14).



Slika 14: Shema povezanih gradnikov, ki sem jih uporabila za izdelavo modelov za napoved vrelišč aromatskih ogljikovodikov.

V tabeli 5 so zbrani vhodnimi podatki (vrelišča, kode SMILES, relativne molekulske mase in Randićevi indeksi), v njej pa so tudi rezultati (napovedana vrelišča in odmiki napovedanih vrelišč od dejanskih vrelišč). Rezultate bom komentirala v poglavju rezultati.

Tabela 5: Vhodni podatki (Randićev indeks, relativna molekulska masa in vrelišče) izbrane skupine aromatskih ogljikovodikov za izdelavo modelov za napoved vrelišč ter napovedana vrelišča z metodama linearne regresije in naključnega gozda ter odmiki napovedanih vrednosti od pravih vrednosti.

Koda SMILES	*RI*	M_r	T_v [°C]	LR	Δ_{LR}	RF	Δ_{RF}
Cc1ccc2ccccc2c1	5,36	142,08	241,00	246,46	5,46	240,30	0,70
c1ccc2c(c1)cc3c4ccccc4c5ccccc2c35	9,93	252,09	481,00	488,41	7,41	489,09	8,09
Cc1cccc2ccccc12	5,38	142,08	245,00	247,28	2,28	243,61	1,39
c1ccc2cc3c4cccc5cccc(c3cc2c1)c45	9,92	252,09	481,00	487,37	6,37	488,95	7,95

Cc1ccc2cc(C)ccc2c1	5,75	156,09	262,00	266,22	4,22	264,37	2,37
c1ccc2c(c1)c3cccc4ccc5cccc2c5c34	9,93	252,09	496,00	487,93	8,07	486,48	9,52
Cc1ccc2ccc(C)cc2c1	5,75	156,09	262,00	266,22	4,22	263,34	1,34
c1ccc2c(c1)ccc3c2ccc4c5ccccc5ccc34	10,92	278,11	519,00	540,71	21,71	535,23	16,23
Cc1ccc2cccc(C)c2c1	5,77	156,09	263,00	267,19	4,19	267,34	4,34
c1ccc2c(c1)cc3ccc4cccc5ccc2c3c45	9,92	252,09	496,00	486,93	9,07	486,51	9,49
Cc1cc(C)c2ccccc2c1	5,77	156,09	265,00	267,07	2,07	268,09	3,09
c1cc2cccc3c4cccc5cccc(c(c1)c23)c45	9,93	252,09	497,00	487,90	9,10	496,10	0,90
Cc1ccc2c(C)cccc2c1	5,77	156,09	266,00	267,01	1,01	267,66	1,66
c1ccc2c(c1)ccc3cc4c(ccc5ccccc45)cc23	10,90	278,11	535,00	538,87	3,87	531,58	3,42
Cc1cc2ccccc2cc1C	5,77	156,09	268,00	266,89	1,11	264,89	3,11
c1cc2ccc3ccc4ccc5cccc6c(c1)c2c3c4c56	10,92	276,09	542,00	540,04	1,96	533,25	8,75
Cc1ccc(C)c2ccccc12	5,79	156,09	268,00	267,94	0,06	267,99	0,01
c1cc2ccc3ccc4ccc5ccc6ccc1c7c2c3c4c5c67	11,90	300,09	590,00	592,23	2,23	594,07	4,07
Cc1cccc2c(C)cccc12	5,79	156,09	269,00	267,89	1,11	267,36	1,64
c1ccc2c(c1)cc3c4ccccc4c5cccc6ccc2c3c56	11,92	302,11	592,00	592,66	0,66	593,19	1,19
Cc1cc(C)c2ccccc2c1	5,77	156,09	271,00	266,72	4,28	264,89	6,11
c1ccc2c(c1)cc3ccc4cc5ccccc5c6ccc2c3c46	11,90	302,11	594,00	591,42	2,58	589,01	4,99
Cc1cc2cccc(C)c2cc1C	6,18	170,11	285,00	288,22	3,22	285,01	0,01
c1ccc2c(c1)cc3ccc4cccc5c6ccccc6c2c3c45	11,92	302,11	595,00	592,42	2,58	586,59	8,41
Cc1ccc2cc(C)c(C)cc2c1	6,17	170,11	286,00	286,92	0,92	291,03	5,03
c1ccc2c(c1)cc3ccc4c5ccccc5cc6ccc2c3c46	11,90	302,11	596,00	591,26	4,74	592,84	3,16
Cc1ccc2c(Cc3ccccc23)c1	6,84	180,09	318,00	324,27	6,27	339,44	21,44
C1c2ccccc2c3c1ccc4ccccc34	8,43	216,09	406,00	408,83	2,83	407,48	1,48
Cc1ccc2ccc3ccccc3c2c1	7,34	192,09	352,00	350,54	1,46	359,70	7,70
c1cc2ccc3cc4cccc5ccc6cc(c1)c2c3c6c45	10,90	276,09	547,00	538,78	8,22	533,95	13,05
Cc1ccc2c(ccc3ccccc23)c1	7,34	192,09	355,00	350,47	4,53	357,43	2,43
c1ccc2c(c1)ccc3cc4ccc5ccccc5c4cc23	10,90	278,11	531,00	539,09	8,09	534,00	3,00
c1ccc2c(c1)c3ccc4c5ccccc5c6ccc2c3c46	10,93	276,09	531,00	541,62	10,62	536,51	5,51
Cc1ccc2cc3ccccc3cc2c1	7,33	192,09	359,00	349,35	9,65	357,07	1,93
c1ccc2c(c1)c3ccc4ccc5cccc6cc2c3c4c56	10,92	276,09	534,00	540,42	6,42	535,21	1,21
Cc1cccc2c1ccc3ccccc23	7,36	192,09	359,00	351,41	7,59	358,60	0,40

c1ccc2cc3c4ccccc4c5ccccc5c3cc2c1	10,92	278,11	535,00	539,90	4,90	532,03	2,97
Cc1cccc2cc3ccccc3cc12	7,34	192,09	363,00	350,29	12,71	358,71	4,30
c1ccc2c(c1)ccc3c4cccc5cccc(c45)c23	9,93	252,09	480,00	488,44	8,44	489,27	9,27
Cc1ccc2ccc3ccc(C)cc3c2c1	7,74	206,11	363,00	371,12	8,12	376,39	13,39
c1cc2ccc3ccc4C=Cc5cc(c1)c2c3c45	8,92	226,08	439,00	434,53	4,47	440,17	1,17
Cc1cc2cccc3ccc4cccc1c4c23	8,34	216,09	410,00	403,39	6,61	405,95	4,05
c1ccc2ccccc2c1	4,97	128,06	218,00	227,91	9,91	243,99	25,99
Cc1cc2ccc3cccc4ccc(c1)c2c34	8,33	216,09	410,00	402,36	7,64	405,91	4,09
c1cc2cccc3C=Cc(c1)c23	5,95	152,06	270,00	279,44	9,44	282,29	12,29
Cc1ccc2ccc3cccc4ccc1c2c34	8,34	216,09	410,00	403,39	6,61	405,30	4,70
C1Cc2cccc3cccc1c23	5,95	154,08	279,00	277,71	1,29	267,83	11,17
C1c2ccccc2c3cc4ccccc4cc13	8,42	216,09	402,00	407,89	5,89	407,50	5,50
C1c2ccccc2c3ccccc13	6,45	166,08	294,00	305,01	11,01	288,80	5,20
C1c2ccccc2c3ccc4ccccc4c13	8,43	216,09	407,00	408,81	1,81	402,56	4,44
c1ccc2c(c1)ccc3ccccc23	6,95	178,08	338,00	330,29	7,71	330,98	7,02
c1cc2ccc3ccc4cccc5c(c1)c2c3c45	8,93	226,08	432,00	435,87	3,87	440,54	8,54
c1ccc2cc3ccccc3cc2c1	6,93	178,08	340,00	329,25	10,75	329,42	10,58
c1ccc2cc3c(ccc4ccccc34)cc2c1	8,92	228,09	435,00	434,23	0,77	443,01	8,01
C1c2cccc3ccc4cccc1c4c23	7,43	190,08	359,00	356,15	2,85	357,51	1,49
c1ccc2c(c1)c3ccccc3c4ccccc24	8,95	228,09	439,00	436,10	2,90	442,11	3,11
c1ccc2c(c1)c3cccc4cccc2c34	7,95	202,08	383,00	383,73	0,73	364,67	18,33
c1ccc2c(c1)ccc3c4ccccc4ccc23	8,93	228,09	441,00	435,11	5,89	438,20	2,80
c1ccc2cc3cc4ccccc4cc3cc2c1	8,90	228,09	450,00	432,93	17,07	436,91	13,09

3.3 Uporaba programa Orange za napoved vrelišč alkoholov

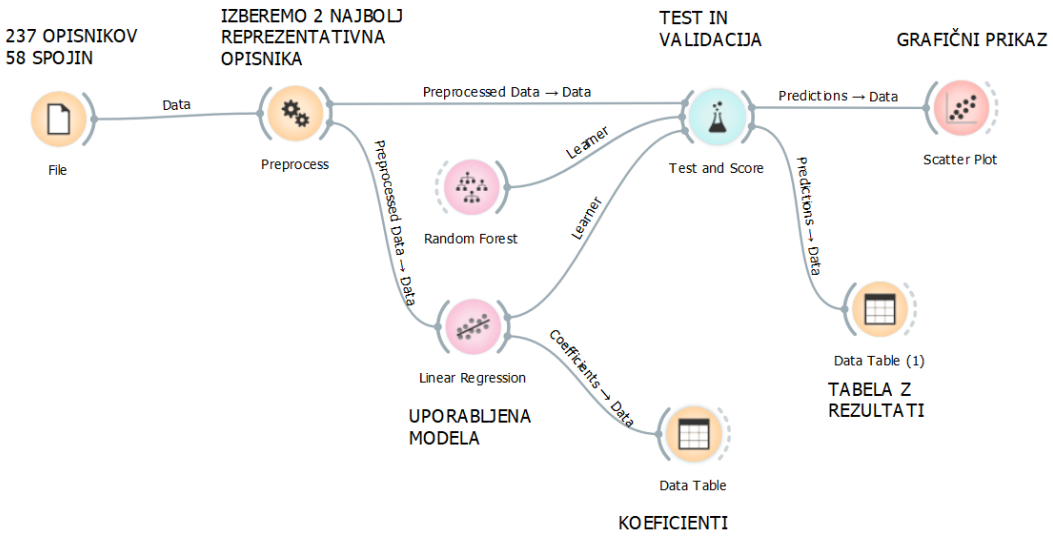
Izdelala sem tudi modela za napoved vrelišč alkoholov. Strukture in vrelišča alkoholov (glej tabelo 6), ki sem jih uporabila za izdelavo modela, so bili pridobljeni iz članka v reviji *Internet Electronic Journal of Molecular Design.* S programom CANVAS (Schrodinger Suite 2018.4) je zunanji mentor izračunal 237 opisnikov za vsako molekulo alkohola. V delovno shemo progama Orange (glej sliko 15) sem vključila gradnik Preprocess, ki je namenjen predhodni obdelavi vhodnih podatkov. V gradniku Preprocess sem postavila zahtevo, da

program izbere dva najbolj reprezentativna opisnika. Izbrana opisnika sem nato uporabila za izgradnjo modelov.

Tabela 6: Vhodni podatki (relativna molekulska masa, izračunani log*P* in vrelišče) izbrane skupine alkoholov za izdelavo modelov za napoved vrelišč ter napovedana vrelišča z metodama linearne regresije in naključnega gozda ter odmiki napovedanih vrednosti od pravih vrednosti

Koda SMILES	Ime spojine	M_r	log*P*	T_v [°C]	LR	Δ_{LR}	RF	Δ_{RF}
CO	METANOL	32,04	-0,36	64,70	56,58	8,12	87,94	23,24
OCC	ETANOL	46,07	-0,01	78,30	69,77	8,53	78,31	0,01
CCCO	PROPAN-1-OL	60,10	0,51	97,20	92,18	5,02	83,13	14,07
CC(C)O	PROPAN-2-OL	60,10	0,37	82,30	85,19	2,89	89,62	7,32
OCCCC	BUTAN-1-OL	74,12	0,97	117,00	110,56	6,44	98,28	18,72
CC(O)CC	BUTAN-2-OL	74,12	0,89	99,60	107,22	7,62	104,87	5,27
CC(C)CO	2-METILPROPAN-1-OL	74,12	0,83	107,90	103,62	4,28	104,33	3,57
CC(C)(C)O	2-METILPROPAN-2-OL	74,12	0,57	82,40	90,87	8,47	94,52	12,12
CCCCCO	PENTAN-1-OL	88,15	1,43	137,80	128,90	8,90	130,97	6,83
CC(O)CCC	PENTAN-2-OL	88,15	1,35	119,00	125,39	6,39	123,98	4,98
CCC(O)CC	PENTAN-3-OL	88,15	1,42	115,30	129,31	14,01	123,90	8,60
OCC(C)CC	2-METILBUTAN-1-OL	88,15	1,29	128,70	121,97	6,73	117,89	10,81
CC(C)CCO	3-METILBUTAN-1-OL	88,15	1,22	131,20	118,29	12,91	114,29	16,91
CC(C)(O)CC	2-METILBUTAN-2-OL	88,15	1,10	102,00	112,51	10,51	112,82	10,82
CC(C)C(C)O	3-METILBUTAN-2-OL	88,15	1,21	111,50	118,23	6,73	122,38	10,88
CC(C)(C)CO	2,2-DIMETILPROPAN-1-OL	88,15	1,11	113,10	112,61	0,49	108,33	4,77
OCCCCCC	HEKSAN-1-OL	102,17	1,88	157,00	147,25	9,75	138,81	18,19
CC(O)CCCC	HEKSAN-2-OL	102,17	1,80	139,90	143,63	3,73	147,40	7,50
CCC(O)CCC	HEKSAN-3-OL	102,17	1,87	135,40	147,51	12,11	148,66	13,26
OCC(C)CCC	2-METILPENTAN-1-OL	102,17	1,75	148,00	140,35	7,65	139,78	8,22
CCC(C)CCO	3-METILPENTAN-1-OL	102,17	1,68	152,40	136,69	15,71	143,96	8,44
CC(C)CCCO	4-METILPENTAN-1-OL	102,17	1,68	151,80	136,70	15,10	139,65	12,15
CC(C)(O)CCC	2-METILPENTAN-2-OL	102,17	1,55	121,40	130,70	9,30	139,05	17,65
CCC(C)C(C)O	3-METILPENTAN-2-OL	102,17	1,67	134,20	136,43	2,23	146,49	12,29
CC(C)CC(C)O	4-METILPENTAN-2-OL	102,17	1,60	131,70	132,90	1,20	133,22	1,52
CC(C)C(O)CC	2-METILPENTAN-3-OL	102,17	1,74	126,60	140,17	13,57	148,73	22,13
CCC(C)(O)CC	3-METILPENTAN-3-OL	102,17	1,62	122,40	134,20	11,80	134,35	11,95
CCC(CC)CO	2-ETILBUTAN-1-OL	102,17	1,75	146,50	140,38	6,12	141,85	4,65
OCC(C)(C)CC	2,2-DIMETILBUTAN-1-OL	102,17	1,56	136,80	130,80	6,00	128,51	8,29
OCC(C)C(C)C	2,3-DIMETILBUTAN-1-OL	102,17	1,54	149,00	129,20	19,80	127,35	21,65
CC(C)(C)CCO	3,3-DIMETILBUTAN-1-OL	102,17	1,43	143,00	122,69	20,31	134,34	8,66
CC(C)(O)C(C)C	2,3-DIMETILBUTAN-2-OL	102,17	1,42	118,60	123,61	5,01	129,86	11,26
CC(C)(C)C(C)O	3,3-DIMETILBUTAN-2-OL	102,17	1,49	120,00	127,19	7,19	143,39	23,39
CCCCCCCO	HEPTAN-1-OL	116,20	2,34	176,30	165,50	10,80	154,62	21,68
CCC(O)CCCC	HEPTAN-3-OL	116,20	2,33	156,80	165,78	8,98	154,63	2,17
CCCC(O)CCC	HEPTAN-4-OL	116,20	2,33	155,00	165,86	10,86	157,23	2,23

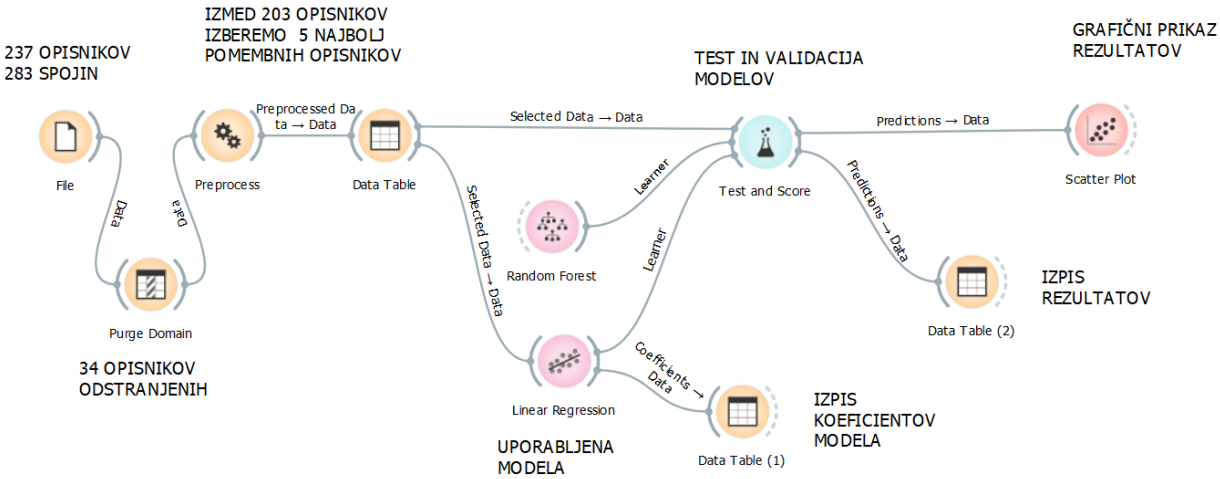
CC(C)(O)CCCC	2-METILHEKSAN-2-OL	116,20	2,01	142,50	148,96	6,46	137,46	5,04
CCC(C)(O)CCC	3-METILHEKSAN-3-OL	116,20	2,08	142,40	152,51	10,11	140,02	2,38
CCC(O)(CC)CC	3-ETILPENTAN-3-OL	116,20	2,14	142,50	156,08	13,58	142,07	0,43
CC(C)(O)C(C)CC	2,3-DIMETILPENTAN-2-OL	116,20	1,87	139,70	141,75	2,05	139,64	0,06
CCC(C)(C)C(C)O	3,3-DIMETILPENTAN-2-OL	116,20	1,94	133,00	145,76	12,76	140,67	7,67
CC(C)(C)C(O)CC	2,2-DIMETILPENTAN-3-OL	116,20	2,01	136,00	149,12	13,12	141,78	5,78
CCC(C)(O)C(C)C	2,3-DIMETILPENTAN-3-OL	116,20	1,94	139,00	145,43	6,43	138,53	0,47
CC(C)C(O)C(C)C	2,4-DIMETILPENTAN-3-OL	116,20	2,05	138,80	151,41	12,61	141,39	2,59
OCCCCCCCC	OKTAN-1-OL	130,23	2,80	195,20	183,64	11,56	183,18	12,02
CC(O)CCCCCC	OkTAN-2-OL	130,23	2,72	179,80	180,24	0,44	187,65	7,85
CCC(CO)CCCC	2-ETILHEKSAN-1-OL	130,23	2,66	184,60	176,92	7,68	182,53	2,07
CC(C)(C)C(C)(O)CC	2,2,3-TRIMETILPENTAN-3-OL	130,23	2,21	152,20	154,19	1,99	155,07	2,87
CCCCCCCCCO	NONAN-1-OL	144,25	3,25	213,10	201,71	11,39	197,07	16,03
CC(O)CCCCCCC	NONAN-2-OL	144,25	3,17	198,50	198,57	0,07	198,48	0,02
CCC(O)CCCCCC	NONAN-3-OL	144,25	3,24	194,70	202,75	8,05	198,74	4,04
CCCC(O)CCCCC	NONAN-4-OL	144,25	3,24	193,00	202,90	9,90	198,34	5,34
CCCCC(O)CCCC	NONAN-5-OL	144,25	3,24	195,10	202,72	7,62	197,94	2,84
CC(C)CCCCCCO	7-METILOKTAN-1-OL	144,25	3,05	206,00	191,22	14,78	190,59	15,41
CC(C)CC(O)CC(C)C	2,6-DIMETILHEPTAN-4-OL	144,25	2,83	178,00	180,98	2,98	189,43	11,43
CCC(C)C(O)C(C)CC	3,5-DIMETILHEPTAN-4-OL	144,25	2,97	187,00	187,85	0,85	196,55	9,55
CC(C)(C)CC(C)CCO	3,5,5-TRIMETILHEKSAN-1-OL	144,25	2,59	193,00	163,34	29,66	186,86	6,14
OCCCCCCCCCC	DEKAN-1-OL	158,28	3,71	230,20	219,76	10,44	207,07	23,13



Slika 15: Shema povezanih gradnikov, ki sem jih uporabila za izdelavo modelov za napoved vrelišč alkoholov

3.4 Uporaba programa Orange za napoved vodne toksičnosti pesticidov za organizem Daphnia magna

V raziskovalni nalogi sem uporabila eksperimentalne podatke o toksičnosti pesticidov (glej prilogo 2) za izdelavo modelov za napoved toksičnosti na osnovi strukture molekul. Kot pri predhodnih primerih sem najprej strukture molekul pesticidov uporabila za izračun molekulskih opisnikov. Tokrat mi je opisnike izračunal zunanji mentor s programom Canvas, ki je del programskega paketa Schrodinger Suite 2018.4. Omenjeni program je za vsako izmed molekul v setu izračunal 237 opisnikov. Opisnike in eksperimentalne podatke o toksičnosti sem shranila v datoteki csv, le-to sem nato uvozila v program Orange. Shema, ki sem jo uporabila za izdelavo modela, je prikazana spodaj na sliki 16.



Slika 16: Shema programa Orange za izdelavo napovedi vodne toksičnosti pesticidov

Pri izdelavi modela sem bila v dilemi, kakšno je najprimernejše število opisnikov in katere opisnike izbrati, da bodo dobljeni modeli imeli smiselno napovedno moč. Sama sem se odločila, da bom izdelala modela (večparametrska regresija, naključni gozd) z uporabo petih deskriptorjev, izbiro opisnikov pa sem prepustila programu Orange (gradnik Preprocess).

4 REZULTATI

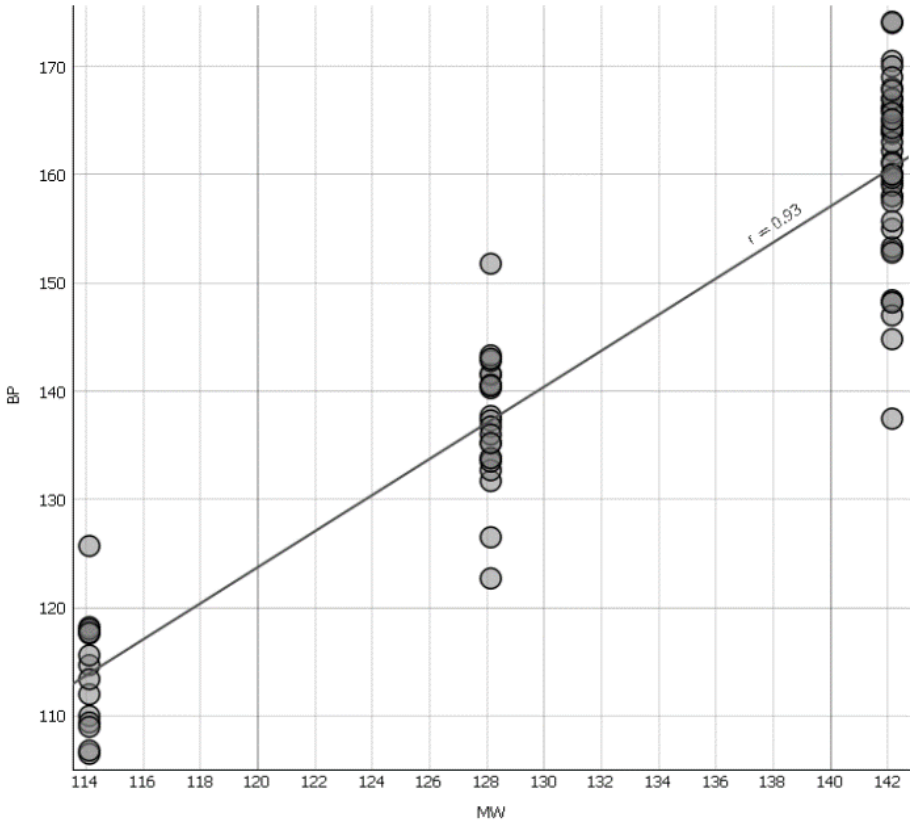
4.1 Uporaba programa Orange za napoved vrelišč alkanov

V prvem delu raziskovalne naloge sem program Orange uporabila za izdelavo modelov za napoved temperatur vrelišč alkanov. Zaradi poznavanja dejstva, da je vrelišče odvisno od razvejenosti molekule, sem se odločila, da pri izdelavi modelov vključim le spojine z največjim številom strukturnih izomerov, to so alkani z molekulskimi formulami C_8H_{18}, C_9H_{20} in $C_{10}H_{22}$ (glej tabelo 1). Izdelala sem različne modele – uporabila sem eno, dve ali pa tri spremenljivke. Rezultati so ločeno predstavljeni v nadaljevanju. V vseh grafih so vrelišča so označena z BP (iz angl. boiling point).

4.1.1 Modeli za napoved vrelišč alkanov z eno spremenljivko

4.1.1.1 Odvisnost vrelišč alkanov od relativne molekulske mase (Mr)

Na sliki 17 je prikazana odvisnost vrelišč razvejenih molekul alkanov od relativne molekulske mase, v tabeli 7 pa so navedeni podatki statistične analize.



Slika 17: Prikaz odvisnosti vrelišč alkanov od njihove molske mase (v analizo so zajeti izomeri spojin z molekulskimi formulami C_8H_{18}, C_9H_{20} in $C_{10}H_{22}$).

Z metodo linearne regresije (v model so bile vključene spojine iz tabele X1) smo dobili enačbo, ki opisuje odvisnost temperature vrelišča T_v od relativne molekulske mase M_r. Ta enačba se glasi:

$T_v = -76{,}33 + 1{,}67 \times M_r$ (T_v v stopinjah Celzija)

Tabela 7: Podatki statistične analize (odvisnost T_v od M_r)

Model	MSE	RMSE	MAE	R2
Linearna regresija	47,832	6,916	5,378	0,868
Naključni gozd	47,890	6,920	5,363	0,868
Linearna regresija (LOO)	50,061	7,075	5,509	0,862
Naključni gozd (LOO)	50,645	7,117	5,561	0,861

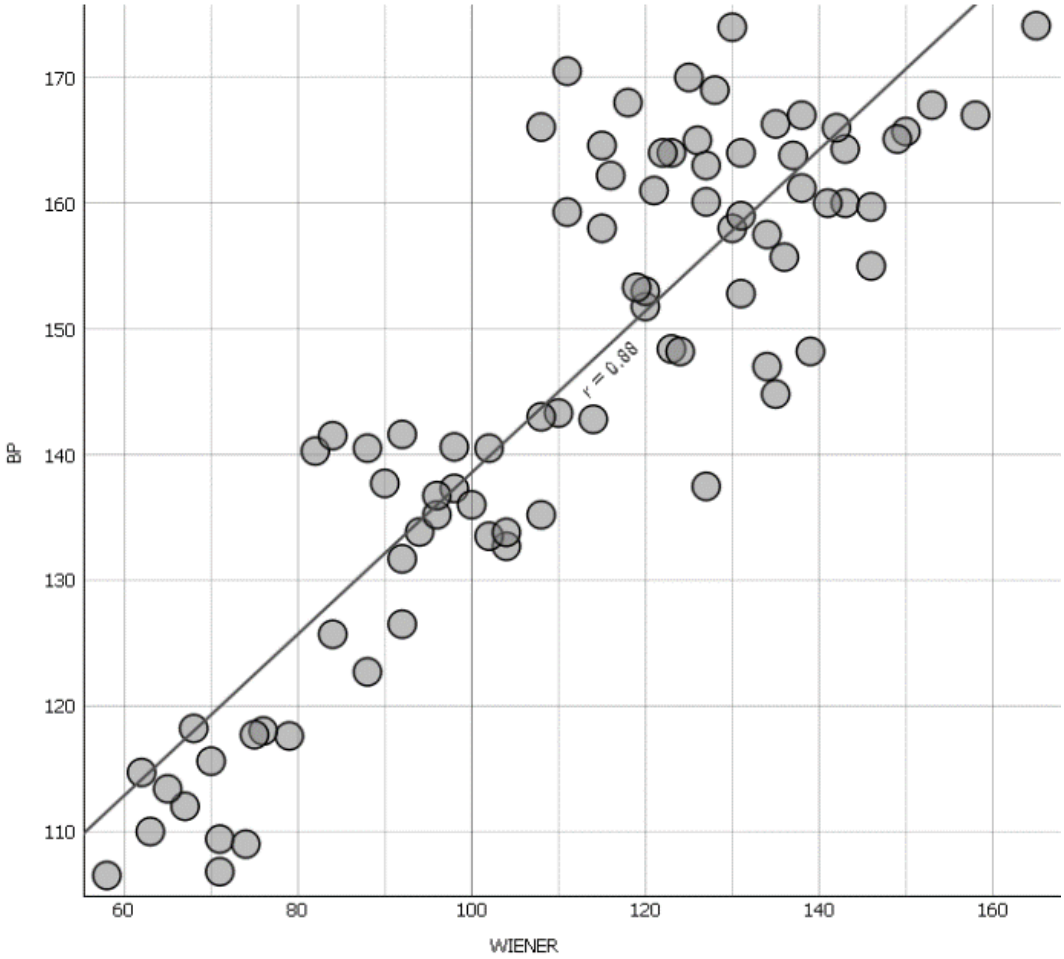
Vrednosti koeficientov determinacije R^2 pri obeh modelih in pri preverjanju napovednih moči modelov znašajo okoli 0,86, zato lahko sklepamo, da modeli, ki za napoved vrelišč upoštevajo le odvisnost od relativne molekulske mase, niso najbolj ustrezni. Odmiki od prave vrednosti vrednoteni z *MAE* znašajo okoli 5,5 °C. Ugotovim lahko, da relativna molekulska masa in vrelišča med seboj korelirajo, a molekulska masa ne pove nič o izomeriji alkanov.

Slabost tega modela je, da ne razlikuje med strukturnimi izomeri alkanov. Po tem modelu imajo vsi strukturni izomeri z enako relativno molekulsko maso, kot so na primer 2,2-dimetilpropan, 2-metilbutan in pentan, enako vrelišče. Dejstvo pa je, da so njihova vrelišča različna.

Za opis razvejenosti molekul so primerni topološki indeksi (glej teoretični del). V mojem delu sem uporabila dva: Wienerjev in Randićev indeks. Na skupini alkanov sem preverila, ali sta ta dva indeksa primerna za izdelavo modelov za napoved vrelišča.

1. Odvisnost vrelišč alkanov od vrednosti Wienerjevega indeksa (*WI*)

Na sliki 18 je prikazana odvisnost vrelišč razvejenih molekul alkanov od vrednosti Wienerjevega indeksa (*LR*), v tabeli 8 pa so navedeni podatki statistične analize.



Slika 18: Prikaz odvisnosti vrelišč alkanov od vrednosti Wienerjevega indeksa (v analizo so zajeti izomeri spojin z molekulskimi formulami C_8H_{18}, C_9H_{20} in $C_{10}H_{22}$)

Z metodo linearne regresije smo dobili enačbo, ki opisuje odvisnost temperature vrelišča T_v od Wienerjevega indeksa *MI*. Ta enačba se glasi:

$T_v = 74{,}31 + 0{,}64 \times WI$ (T_v v stopinjah Celzija)

Tabela 8: Podatki statistične analize (odvisnost T_v od *WI*)

Model	MSE	RMSE	MAE	R2
Linearna regresija	81,079	9,004	7,112	0,776
Naključni gozd	39,293	6,268	4,569	0,892
Linearna regresija (LOO)	84,539	9,195	7,279	0,767
Naključni gozd (LOO)	90,689	9,523	7,493	0,750

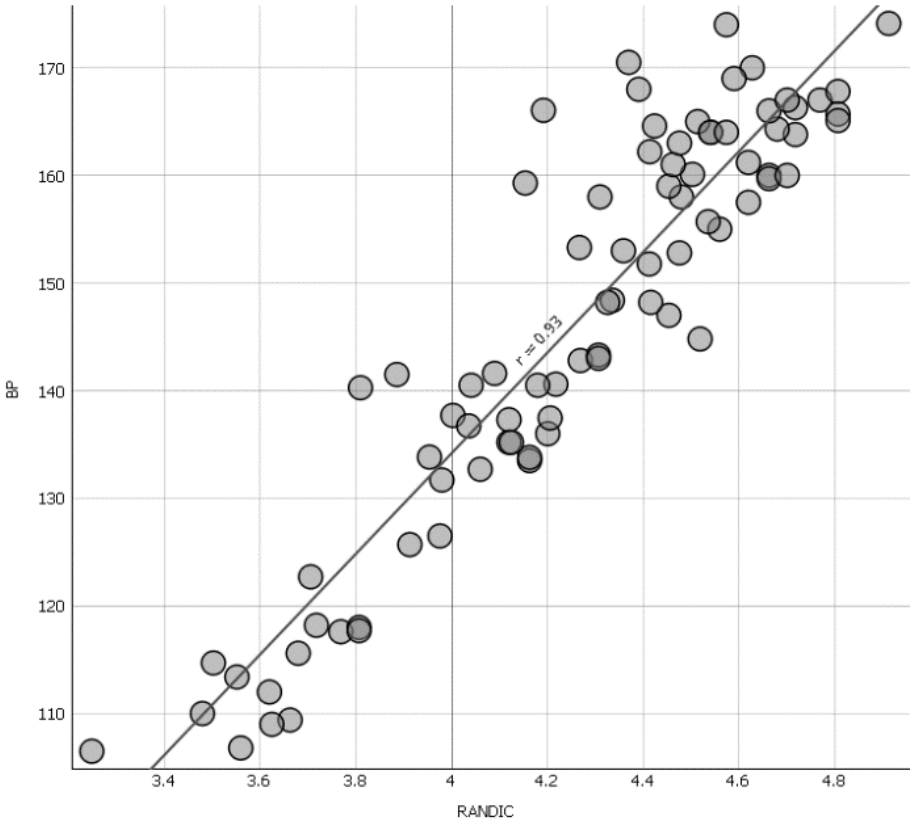
Povprečni absolutni odmik (*MAE*) med pravimi in izračunanimi vrednostmi znaša za model dobljen z metodo naključnega gozda 4,6 °C, za model dobljen z linearno regresijo pa 7,1 °C.

Za modele z enim opisnikom je to kar v redu napoved. Za validacijo modelov smo uporabili metodo »izpusti enega«. Zanimivo je, da pri se linearni regresiji *MAE* praktično ne spremeni, pri metodi naključnega gozda pa se poveča skoraj za 3 °C iz 4,6 °C na 7,5 °C. Podoben trend opazimo tudi pri R^2. Zaključimo lahko, da vrednosti Wienerjevega indeksa in vrelišč alkanov solidno korelirajo.

4.1.1.2 Odvisnost vrelišč alkanov od vrednosti Randićevega indeksa (RI)

Zadnja izdelana enoparametrska modela sta podala odvisnost vrelišča od vrednosti Randićevega indeksa. Rezultat modela z linearno regresijo je predstavljen na sliki 19. Model za napoved vrelišča v odvisnosti od Randićevega indeksa na osnovi linearne regresije opisuje enačba:

$T_v = -52{,}89 + 46{,}75 \times RI$ (T_v v stopinjah Celzija)



Slika 19: Prikaz odvisnosti vrelišč alkanov od vrednosti Randićevega indeksa (v analizo so zajeti izomeri spojin z molekulskimi formulami C_8H_{18}, C_9H_{20} in $C_{10}H_{22}$).

V tabeli 9 so zbrani statistični podatki modelov. Najmanjšo srednjo vrednost absolutnih napak izmed vseh modelov osnovanih na eni spremenljivki opazim pri modelu naključnega gozda (*MAE* = 3,4 °C). Tudi R^2 je za ta model najvišji in znaša 0,94. Validacija tega modela z metodo "izpusti enega" pa razkrije, da je napovedna moč modela osnovana na naključnem gozdu rahlo manjša od tistega osnovanega na uporabi linearne regresije. V splošnem lahko zaključim, da sta vrelišče alkanov in Radnićev indeks soodvisna.

Tabela 9: Podatki statistične analize (odvisnost T_v od *RI*)

Model	MSE	RMSE	MAE	R2
Linearna regresija	50,926	7,136	5,695	0,860
Naključni gozd	20,461	4,523	3,411	0,944
Linearna regresija (LOO)	53,115	7,288	5,829	0,853
Naključni gozd (LOO)	74,143	8,611	6,622	0,795

4.1.2 Model za napoved vrelišč alkanov z dvema spremenljivkama

V nadaljevanju sem pri izdelavi modela za napoved vrelišč uporabila dve spremenljivki: Randićev indeks in relativno molekulsko maso.

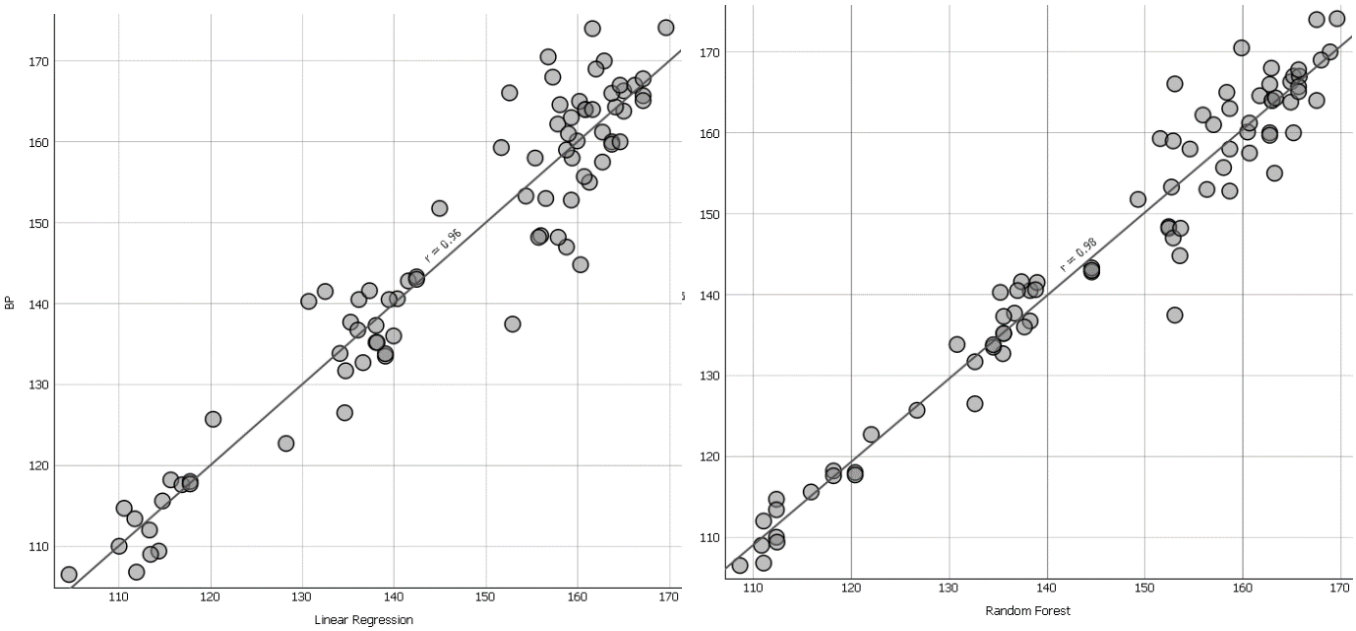
Spodnja enačba, ki povezuje temperaturo vrelišča z vrednostjo Randićevega indeksa in relativne molekulske mase, je dobljena z metodo dvoparametrske linearne regresije:

$T_v = -76{,}87 + 23{,}65 \times RI + 0{,}917 \times M_r$ (T_v v stopinjah Celzija)

Tabela 10: Podatki statistične analize (odvisnost T_v od M_r in *RI*). Modela sta izdelana in validirana na podlagi izomerov spojin z molekulskimi formulami C_8H_{18}, C_9H_{20} in $C_{10}H_{22}$.

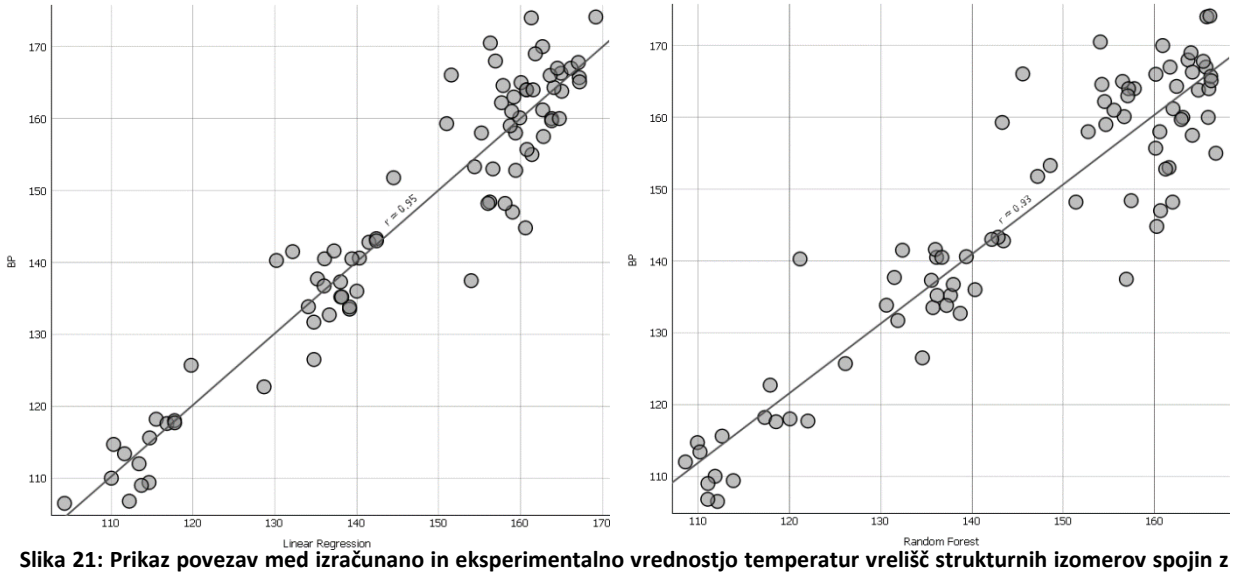
Model	MSE	RMSE	MAE	R2
Linearna regresija	31,863	5,645	3,058	0,912
Naključni gozd	16,024	4,003	3,058	0,956
Linearna regresija (LOO)	34,448	5,869	4,416	0,905
Naključni gozd (LOO)	50,731	7,123	5,433	0,860

Na sliki 20 sta prikazana grafa na osnovi modelov dvoparametrske linearne regresije (levo) in naključnega gozda (desno), ki opisujeta povezavo med eksperimentalno in izračunano vrednostjo vrelišča. V tem primeru se izkaže, da se metoda na osnovi naključnega gozda rahlo bolje odreže pri napovedi vrelišča. V tabeli 10 je namreč koeficient determinacije R^2 večji pri metodi naključnega gozda.



Slika 20: Prikaza povezave med izračunano in eksperimentalno vrednostjo temperatur vrelišč (v analizo so zajeti izomeri spojin z molekulskimi formulami C_8H_{18}, C_9H_{20} in $C_{10}H_{22}$). Diagram na levi predstavlja regresijski model, diagram na desni pa model dobljen z metodo naključnega gozda.

Na sliki 21 sta grafa, ki prikazujeta odvisnost med izračunano in ekperimentalno vrednostjo vrelišča za dvoparameterska modela linearne regresije (levo) in metode naključnega gozda (desno). V tem primeru smo modela validirali z metodo "izpusti enega". Izkaže se, da ima v tem primeru model na osnovi dvoparametrske linearne regresije za malenkost boljšo napovedno moč. To dejstvo lahko razberemo iz tabele 11, saj je koeficient determinacije večji pri metodi linearne regresije.



Slika 21: Prikaz povezav med izračunano in eksperimentalno vrednostjo temperatur vrelišč strukturnih izomerov spojin z molekulskimi formulami C_8H_{18}, C_9H_{20} in $C_{10}H_{22}$. Diagram na levi predstavlja regresijski model, diagram na desni pa model dobljen z metodo naključnega gozda.

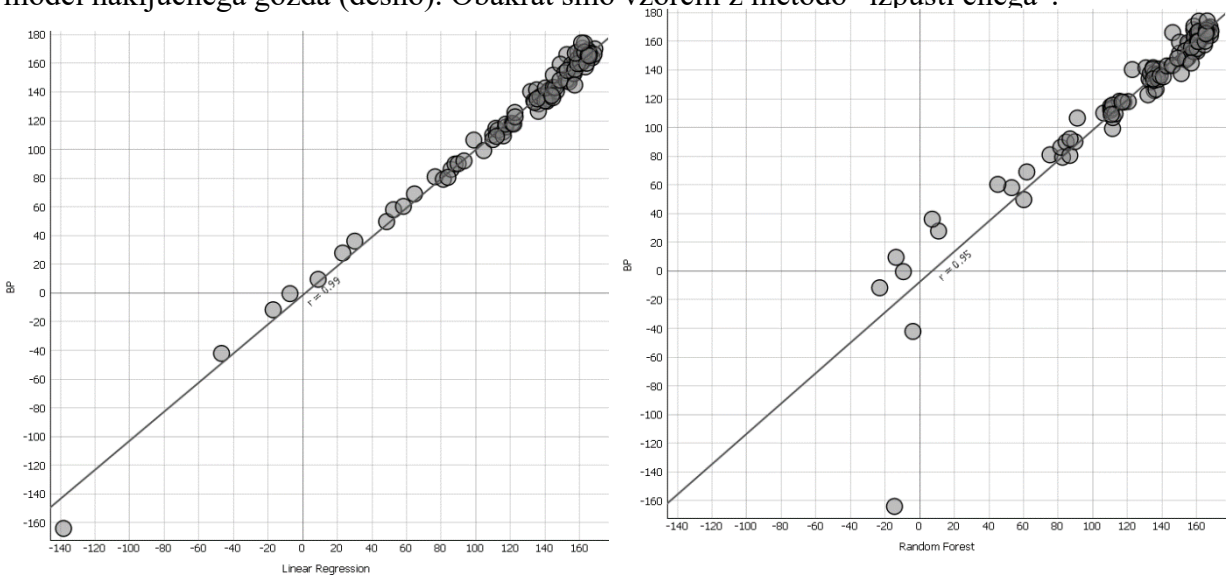
V skladu s pričakovanji je kombinacija molekulske mase in enega izmed topoloških indeksov (Randićev indeks) boljša za izračun modelnih vrednosti vrelišč alkanov, kot katerikoli predstavljen model zgrajen zgolj z enim opisnikom, saj so srednje vrednosti absolutnih napak *MAE* manjše.

4.1.3 Model za napoved vrelišč alkanov s tremi spremenljivkami

Nazadnje sem zgradila model, ki pri napovedi vrelišča upošteva odvisnost od relativne molekulske mase in obeh topoloških indeksov. Pri izdelavi modela sem zajela podatke vseh alkanov iz tabele 3, torej tudi krajše alkane. Triparametrski model z metodo *LR* opisuje enačba:

$T_v = -169{,}51 + 57{,}50 \times RI - 0{,}57 \times WI + 1{,}01 \times M_r$ (T_v v stopinjah Celzija)

Model, ki ga opisuje ta enačba, je prikazan na sliki 22 (levo), medtem ko je na desni prikazan model naključnega gozda (desno). Obakrat smo vzorčili z metodo “izpusti enega”.



Slika 22: Prikaz povezav med izračunano in eksperimentalno vrednostjo temperatur vrelišč alkanov. Diagram na levi predstavlja triparametrski regresijski model, diagram na desni pa model dobljen z metodo naključnega gozda. Za izdelavo modela sem uporabila tri spremenljivke, vzorčila sem z metodo »izpusti enega«.

Iz slike 22 lahko opazim, da metoda naključnega gozda odpove pri napovedi vrelišč nižjih alkanov, tj. tistih, ki imajo pet ali manj ogljikovih atomov v molekuli.

Nato sem ovrednotila kakovost modela tudi z uporabo zunanje testne skupine spojin (iz tabele 4). Rezultati so predstavljeni v tabeli 11.

Tabela 11: Dejanska vrelišča in napovedana vrelišča spojin iz testne skupine z metodama linearne regresije (LR) in naključnega gozda (RF) na osnovi treh parametrov. Z Δ sta označena odmika napovedanega vrelišča od dejanskega vrelišča za obe metodi.

Koda SMILES	Ime spojine	T_v [°C]	LR	Δ_{LR}	RF	Δ_{RF}
CC	ETAN	-88,6	-82,2	6,4	-18,7	69,9
CCC(C)CC	3-METILPENTAN	63,3	61,2	2,1	54,3	9,0
CCCCCCC	HEPTAN	98,4	96,0	2,4	98,1	0,3
CCC(C)CC(C)C	2,4-DIMETHILHEKSAN	115,6	116,0	0,4	110,0	5,6
CC(C)C(C)C(C)(C)C	2,2,3,4-TETRAMETILPENTAN	133,0	132,5	0,5	139,0	6,0
CC(C)CCC(C)(C)C	2,2,5-TRIMETILHEKSAN	124,0	129,3	5,3	131,3	7,3
CCCC(C)CC(C)C	2,4-DIMETILHEPTAN	138,0	141,2	3,2	135,4	2,7
CCCCC(C)CCC	4-METILOKTAN	142,4	146,1	3,7	146,1	3,7
CC(C)C(C)CC(C)(C)C	2,2,4,5-TETRAMETILHEKSAN	155,3	152,2	3,1	150,4	4,9
CC(C)C(C)C(C)C(C)C	2,3,4,5-TETRAMETILHEKSAN	169,4	161,8	7,6	161,1	8,4
CCCC(C)(C)CC(C)C	2,4,4-TRIMETILHEPTAN	153,0	159,2	6,2	160,4	7,4
CCCC(C)C(C)CCC	4,5-DIMETILOKTAN	167,0	168,5	1,5	165,5	1,5
CC(C)CC(C)CC(C)C	2,4,6-TRIMETILHEPTAN	157,0	157,1	0,1	151,0	6,0
CCCC(C)CC(C)CC	3,5-DIMETILOKTAN	162,0	165,8	3,8	165,2	3,2
CC(C)CCCCC(C)C	2,7-DIMETILOKTAN	159,9	154,1	5,8	161,0	1,2

Opazim lahko, da model naključnega gozda zelo slabo napove vrelišče molekule etana, modelna vrednost vrelišča znaša -18,7 °C, kar je skoraj 70 °C več kot pravo vrelišče etana. Če odmislimo etan, obe metodi dostojno napovesta vrelišča ostalih alkanov. Povprečna absolutna napaka pri napovedi vrelišč alkanov je pri modelu na osnovi linearne regresije 3,3 °C, pri modelu naključnega gozda pa 4,8 °C. Vrednosti sem izračunala tako, da sem izračunala povprečno vrednost odmikov napovedanih vrelišč od dejanskih vrelišč spojin v tabeli 11.

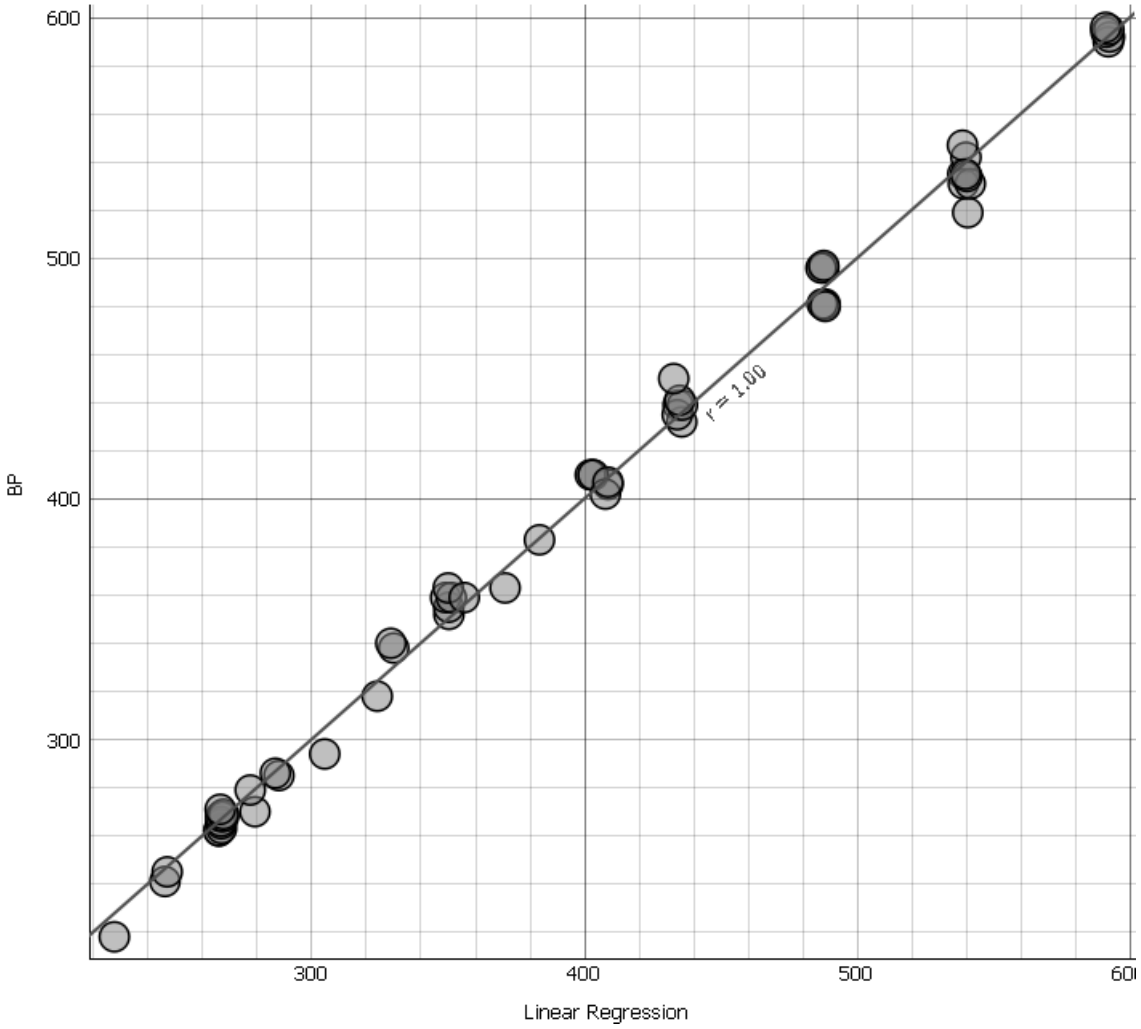
4.2 Uporaba programa Orange za napoved vrelišč aromatskih ogljikovodikov

Izkaže se, da lahko vrelišča aromatskih ogljikovodikov dobro opiše model z dvema spremenljivkama, z relativno molekulsko maso in Randićevim indeksom. Model izdelan z metodo LR poda naslednjo enačbo:

$T_v = -34{,}56 + 58{,}41 \times RI - 0{,}23 \times M_r$ (T_v v stopinjah Celzija)

Izkaže se, da lahko z modelom izdelanim na osnovi omenjenih dveh opisnikov z veliko zanesljivostjo izračunamo vrelišče, kar je razvidno iz slike 23 in tabele 12. Na sliki 23 je

prikazana korelacija med izračunano in pravo vrednostjo vrelišča za model dobljen na osnovi linearne regresije z metodo vzorčenja "izpusti enega" (LOO).



Slika 23: Prikaz povezave med izračunano in eksperimentalno vrednostjo temperature vrelišča za aromatske ogljikovodike. Diagram predstavlja regresijski model z dvema spremenljivkama Randićevim indeksom in relativno molekulsko maso. Vzorčila sem z metodo navzkrižnega preverjanja "izpusti enega" (LOO).

Tabela 12: Podatki statistične analize (odvisnost T_v od M_r in *RI*).

Model	MSE	RMSE	MAE	R2
Linearna regresija	44,148	6,644	5,351	0,996
Naključni gozd	28,646	5,352	3,919	0,998
Linearna regresija (LOO)	48,489	6,963	5,620	0,996
Naključni gozd (LOO)	65,199	8,075	6,045	0,995

Iz tabele 12 lahko razberemo, da oba modela za napoved vrelišč z upoštevanjem le dveh molekulskih opisnikov (eden povezan v molekulsko maso, drugi pa je topološki indeks), precej dobro napovesta vrelišče, saj je koeficient determinacije R^2 pri obeh modelih večji od 0,99. Tudi vrednosti *MAE* so majhne (okoli 5 °C), torej napovedne vrednosti vrelišč malo odstopajo od eksperimentalno določenih vrednosti.

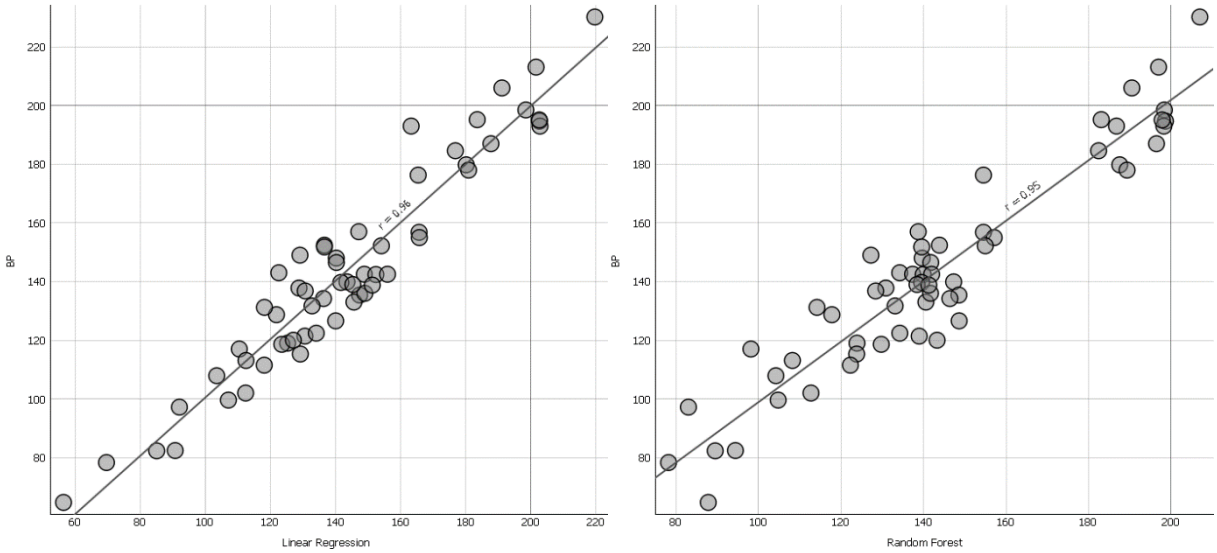
4.3 Uporaba programa Orange za napoved vrelišč alkoholov

V modelih za napoved vrelišč alkoholov smo uporabili dva opisnika: logaritem porazdelitvenega koeficienta *logP* in relativno molekulsko maso M_r. Model izdelan z metodo linearne regresije napove naslednjo odvisnost:

$T_v = 89{,}17 + 52{,}19 \times logP - 0{,}389 \times M_r$ (T_v v stopinjah Celzija)

Iz enačbe je razvidno, da je vrelišče povezano z relativno molekulsko maso in tudi z logaritmom porazdelitvenega koeficienta med nepolarno in polarno fazo. Vrednost *logP* se v praksi uporablja kot merilo za polarnost (oz. nepolarnost molekul). Višja vrednost *logP* pomeni, da je molekula bolj nepolarna.

Na sliki 24 je prikazana korelacija med izračunano in pravo vrednostjo vrelišča za modela dobljena na osnovi linearne regresije in naključnega gozda.



Slika 24: Prikaz povezave med napovedano in eksperimentalno vrednostjo vrelišč alkoholov. Pri obeh modelih sta za napoved vrelišč uporabljena le dva parametra. Na levi strani je model dobljen z linearno regresijo, na desni strani je model osnovan na metodi naključnega gozda. Za validacijo obeh modelov je uporabljena metoda »izpusti enega«.

V splošnem se izkaže, da tako model na osnovi linearne regresije kot model naključnega gozda dobro napovesta vrelišča. Iz tabele 13 je razvidno, da znaša odmik z modelom napovedanih vrelišč od pravih vrednosti merjen s količino MAE za model linearne regresije nekaj več kot 8 °C in za model naključnega gozda nekaj več kot 5 °C. V obeh primerih opazimo linearno odvisnost med izračunanimi in pravimi vrednostmi (R^2 (LR) = 0,93 in R^2 (NG) = 0,97). Modela smo ovrednotili z metodo »izpusti enega« in izkaže se, da statistični podatki za linearno regresijo ostanejo po validaciji praktično enaki, med tem ko se pri naključnem gozdu poveča odmik med izračunano in pravo vrednostjo za več kot 4 °C, zniža pa se tudi vrednost R^2 iz 0,97 na 0,89.

Tabela 13: Podatki statistične analize modelov za napoved vrelišč alkoholov.

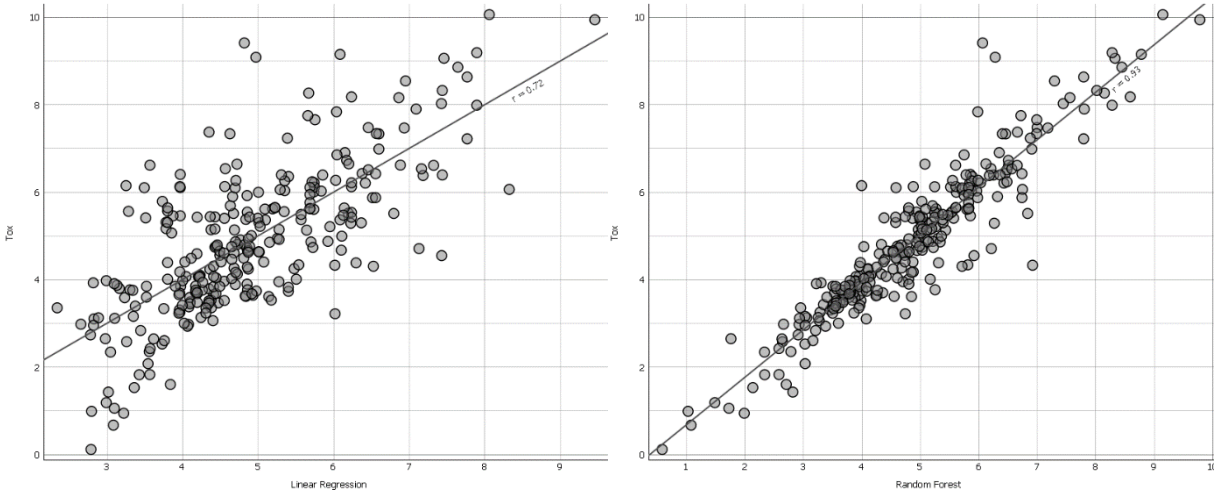
Model	MSE	RMSE	MAE	R2
Linearna regresija	93.61	9,68	8,32	0,93
Naključni gozd	39,98	6,32	5,33	0,97
Linearna regresija (LOO)	106,09	10,30	8,79	0,92
Naključni gozd (LOO)	137,38	11,72	9,76	0,89

4.4 Uporaba programa Orange za napoved vodne toksičnosti pesticidov za organizem Daphnia magna

Program Orange je iz nabora molekulskih opisnikov izbral le fizikalno-kemijske opisnike, kot so relativna molekulska masa (M_r), porazdelitveni koeficient med nepolarno in polarno fazo (*logP*), število atomov v molekuli, ki so donorji (*HBD*) in akceptorji (*HBA*) vodikove vezi in število vrtljivih vezi (*RB*), v končnem modelu ni nobenega topološkega opisnika.
Končni večparametrski regresijski model opisuje enačba:

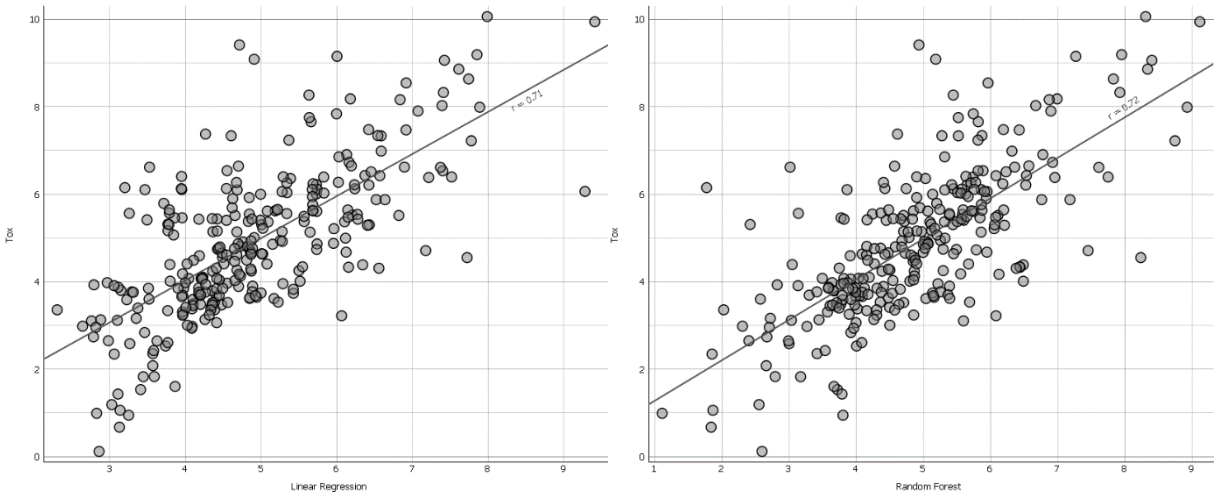
$$pLC_{50} = 2{,}65 + 0{,}0058 \times M_r + 0{,}39 \times logP + 0{,}20 \times HBA - 0{,}094 \times HBD - 0{,}076 \times RB$$

Večparametrski regresijski model in model naključnega gozda za napoved vodne toksičnosti pesticidov na organizem vodne bolhe prikazuje slika 25. Iz tabele 14 je razvidno, da napovedane vrednosti toksičnosti dobljene z metodo naključnega gozda (R^2 = 0,87) precej dobro korelirajo z eksperimentalnimi, vsekakor boljše kot tiste dobljene z večparametrskim regresijskim modelom (R^2 = 0,52).



Slika 25: Prikaz modelov za napoved toksičnosti (Tox = pLC_{50}) spojin za organizem Daphnia magna (vodna bolha). Na levi je model večparametrske regresije, na desni je model na osnovi naključnega gozda (random forest).

Z uporabo metode »izpusti enega« sem preizkusila napovedno moč obeh modelov (glej sliko 26). Izkaže se (glej tabelo 14), da napovedane vrednosti dobljene z metodo naključnega gozda (R^2 = 0,51) bolje korelirajo s pravimi kot tiste napovedane z večparametrskim regresijskim modelom (R^2 = 0,50), vendar je ta razlika majhna. Povprečni absolutni odmik modelnih vrednosti od pravih znaša pri obeh metodah okoli 0,9 (glej prilogo 2).



Slika 26: Prikaz modelov za napoved toksičnosti (Tox = pLC_{50}) spojin za organizem Daphnia magna (vodna bolha). Na levi je model večparametrske regresije, na desni je model na osnovi naključnega gozda (random forest). Vrednotenje sem izvedla po principu »izpusti enega«.

Tabela 14: Podatki statistične analize (odvisnost pLC_{50} od petih parametrov)

Model	MSE	RMSE	MAE	R2
Linearna regresija	1,371	1,171	0,895	0,523
Naključni gozd	0,365	0,604	0,427	0,873
Linearna regresija (LOO)	1,448	1,203	0,917	0,497
Naključni gozd (LOO)	1,409	1,187	0,874	0,510

Dobljeni modeli za vodno toksičnost imajo v skladu s pričakovanjem slabšo napovedno moč kot predhodno obravnavani modeli za napoved vrelišč alkanov, aromatskih ogljikovodikov in alkoholov.

5 RAZPRAVA

V svoji raziskovalni nalogi sem pokazala uporabnost programa Orange, ki so ga razvili na Fakulteti za računalništvo in informatiko Univerze v Ljubljani za obdelavo kemijskih podatkov. Program sem uporabila za napoved vrelišč alkanov, aromatskih spojin in alkoholov ter za napoved toksičnosti pesticidov za organizem *Daphnia magna* (vodna bolha). Ugotovila sem, da je delo s programom Orange dokaj enostavno, saj sem osnovne operacije za delo z omenjenim programom s pomočjo zunanjega mentorja osvojila dokaj hitro. V okviru svoje raziskovalne naloge sem se seznanila z možnostjo napovedi nekaterih lastnosti molekul na osnovi njihove strukture. Znanje in veščine potrebne za izdelavo modela, s katerim lahko na osnovi strukture molekul napovemo določeno lastnost, so interdisciplinarne, saj združujejo tematike več ved, kot so kemija (koncept kemijske strukture, vrelišča), matematika (izračun opisnikov), statistika (izdelava in vrednotenje modelov) in biologija s toksikologijo. Pri izdelavi naloge sem se poleg dela s programom Orange naučila tudi, kako izračunati nekatere izmed opisnikov, ki sem jih uporabila za izdelavo modelov (Randićev in Wienerjev indeks). Modeli, ki sem jih naredila, so uporabni za napoved lastnosti, vendar pa se potrebno zavedati tudi določenih dejstev:

- Modele lahko uporabimo le za molekule, ki so podobne tistim, ki so vključene v izgradnjo modela.
- Kakovost modela je v veliki meri odvisna od zanesljivosti vhodnih podatkov.

- V splošnem velja, da lahko fizikalno-kemijske podatke, med katere sodi npr. vrelišče, izmerimo z večjo zanesljivostjo npr. kot biološke podatke, kot je na primer meritev toksičnosti pesticidov za organizem *Daphnia magna*.

5.1 Modeli za napoved vrelišč alkanov

Na temperaturo vrelišča alkanov poleg velikosti molekul (relativne molske mase) vpliva tudi njihova razvejanost. Z analizo rezultatov sem pokazala, da se z večanjem relativne molekulske alkanov povečuje temperatura vrelišča. Relativna molekulska masa kot edini opisnik za napoved temperature vrelišč alkanov ne zadostuje, saj ne upošteva vpliva razvejenosti molekul alkanov. Razvejenost molekul lahko opišemo s topološkimi indeksi, kot sta na primer Randićev in Wienerjev index. Vrednost obeh indeksov se z naraščanjem dolžine molekul alkanov povečuje in se zmanjšuje z večanjem razvejenosti molekul. Izkaže se, da lahko s kombinacijo treh opisnikov, tj. relativne molekulske mase, Randićevega in Wienerjevega indeksa, izdelamo najbolj ustrezen model za napoved vrelišč alkanov.

5.2 Modeli za napoved vrelišč aromatskih ogljikovodikov

Ugotovila sem, da lahko za napoved vrelišč aromatskih spojin uporabim model z le dvema opisnikoma, to sta Randićev indeks (topološki opisnik) in relativna molekulska masa (fizikalno-kemijski opisnik). Dobljeni modeli imajo izmed vseh v raziskovalni nalogi obravnavanih modelov največjo napovedno moč.

5.3 Modeli za napoved vrelišč alkoholov

Pri napovedi vrelišč alkoholov se izkažeta kot uporabna tista modela, ki združujeta dva fizikalno-kemijska deskriptorja: *logP* (negativni logaritem porazdelitvenega koeficienta med oktan-1-olom in vodo) in M_r (relativna molekulska masa). Izdelana modela sta primerna za oceno vrelišč alkoholov.

5.4 Model za napoved toksičnosti pesticidov za vodne bolhe

V primeru napovedi toksičnosti gre za kompleksen problem, saj mnogokrat mehanizem delovanja posameznega strupa na organizem ni v celoti pojasnjen. Ker strup deluje na vsak organizem drugače, so meritve toksičnosti manj zanesljive. Medtem ko je vrelišče relativno enostavno merljiva fizikalna količina, ki je odvisna predvsem od molske mase, razvejanosti molekul in prisotnosti določenih funkcionalnih skupin. Iz navedenih razlogov sta modela za napoved toksičnosti pesticidov za vodne bolhe manj verodostojna (imata manjšo napovedno moč).

6 ZAKLJUČEK

S programom Orange je mogoče na osnovi eksperimentalnih podatkov izdelati modele za napoved lastnosti strukturno podobnih spojin. Program sem uporabila za napoved vrelišč treh različnih skupin organskih spojin.

Največ različnih modelov (z eno, dvema ali s tremi spremenljivkami) sem izdelala v skupini alkanov. Z modeli sem potrdila dejstvo, da so vrelišča odvisna od velikosti molekul in njihove razvejenosti, saj model, pri katerem je bila npr. edina spremenljivka relativna molekulska masa, ne napove vrelišč dovolj točno, precej bolje vrelišča napove model z vsaj dvema spremenljivkama, tj. relativno molekulsko maso in enim topološkim indeksom.

Ugotovila sem, da sta tudi za napoved policikličnih aromatskih ogljikovodikov dovolj le dve spremenljivki: relativna molekulska masa in en topološki indeks.

Pri izdelavi modela z dvema spremenljivkama za napoved vrelišč alkoholov program Orange poleg relativne molekulske mase ni predlagal uporabe topološkega indeksa, ampak je izbral drug opisnik, tj. logaritem porazdelitvenega koeficienta. Sklepam, da so vrelišča alkoholov odvisna od interakcij med hidrofobnim in hidrofilnim delom, ki jih ta opisnik opredeljuje.

Izdelali smo tudi model za napoved vodne toksičnosti (vrednosti pLC_{50}) pesticidov za vodne bolhe s petimi spremenljivkami. Vodna toksičnost je kot kaže precej odvisna od tega, kako dobro se pesticid raztopi v vodi, saj sta med računalniško izbranimi opisniki tudi število atomov, ki so donorji vodikove vezi in število atomov, ki so akceptorji vodikove vezi.

7 ZAHVALE

Zahvaljujem se svojima mentorjema za pomoč pri raziskovalni nalogi. Mag. Mojci Podlipnik, da mi je razložila snov organske kemije, o kateri še nisem imela dovolj znanja. Pomagala pa mi je tudi pri oblikovanju raziskovalne naloge. Dr. Črtomir Podlipnik pa mi je pojasnil delovanje programa Orange in mi pomagal pri zbiranju podatkov, ki sem jih potrebovala za eksperimentalni del raziskovalne naloge.

8 VIRI IN LITERATURA

VIRI

https://pubchem.ncbi.nlm.nih.gov/edit3/index.html (15. januar 2020; 20.15)

https://www.youtube.com/watch?v=KzHJXdFJSIQ (15. januar 2020; 14.45)

https://studentski.net/gradivo/upr_fhs_ge1_kmg_sno_regresija_in_korelacija_01 (15. januar 2020; 21.30)

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&ved=2ahUKEwj (15. januar 2020; 20.30)

https://en.wikipedia.org/wiki/Polycyclic_aromatic_hydrocarbon (14. januar 2020; 21.05)

https://www.youtube.com/watch?v=J4Wdy0Wc_xQ (27. januar 2020; 15.10)

https://sl.wikipedia.org/wiki/SMILES (14. januar 2020; 19.50)

https://commons.wikimedia.org/wiki/File:Daphnia_magna_asexual.jpg (27. januar 2020; 15.20)

LITERATURA

Basant, N., Gupta, S.in Singh, K. P. (2015). Modeling the toxicity of chemical pesticides in multiple test species using local and global QSTR approaches. Toxicology research, Vol. 5, str. 340–353. Pridobljeno s https://doi.org/10.1039/c5tx00321k

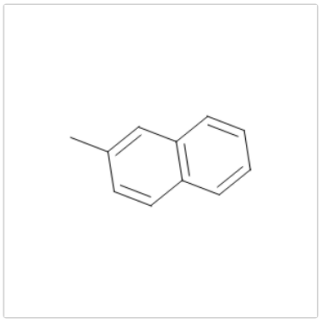
Bouarra, N., Kherouf, S., Bouakkadia, A. in Messadi, D. (november 2017). QSPR Application on Modeling of Boiling Point of Polycyclic Aromatic Hydrocarbons. *Research Journal of Pharmaceutical, Biological and Chemical Sciences (RJPBCS)*, Vol. 8, str. 19-28

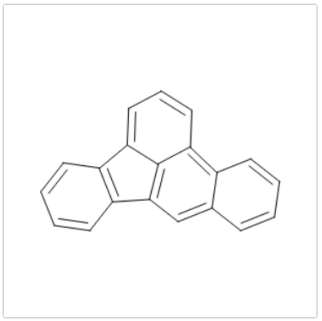
Cherqaoui, D. in Villemin, D. (1994). Use of neural network to determine the boiling point of alkanes. J. Chem. Soc. Faraday Trans., Vol. 90, str. 97-102

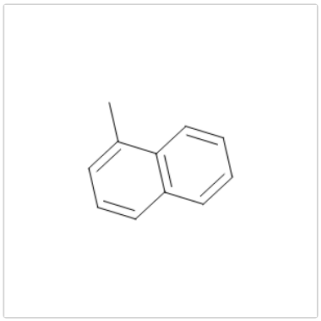
Janežič, D., Lučić, B., Nikolić, S., Miličević, A. in Trinajstić, N. (2006). Boiling Points of Alcohols – A Comparative QSPR Study. *Internet Electronic Journal of Molecular Design*, Vol. 5, str. 192–200

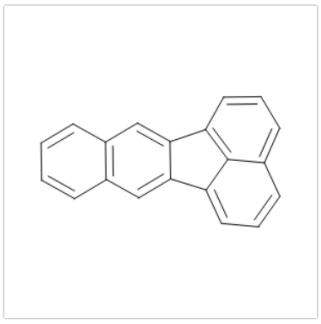
9 PRILOGE

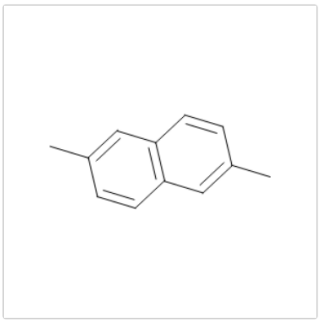
Priloga 1: Skeletne formule in kode SMILES obravnavanih policikličnih aromatskih ogljikovodikov


Cc(c1)ccc(c12)cccc2

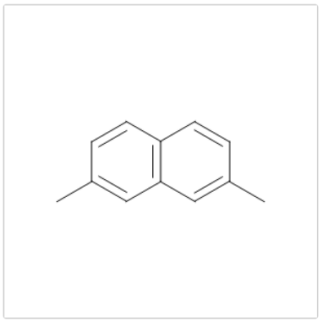

c1cccc(c1c2c34)c3cccc4c5c(c2)cccc5

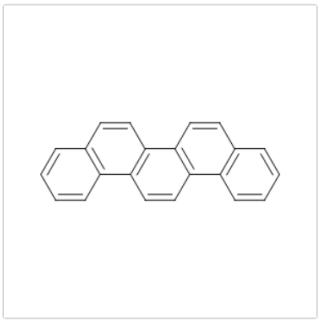

Cc1cccc(c12)cccc2

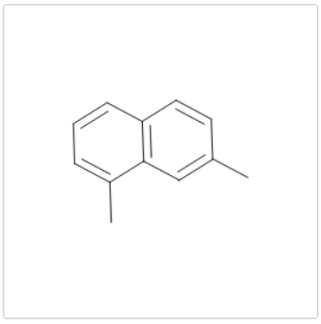

c1cccc(c2)c1cc(c2c3c45)c4cccc5ccc3

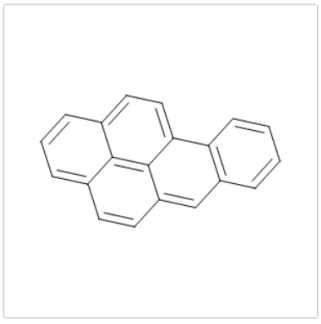

Cc(c1)ccc(c12)cc(C)cc2

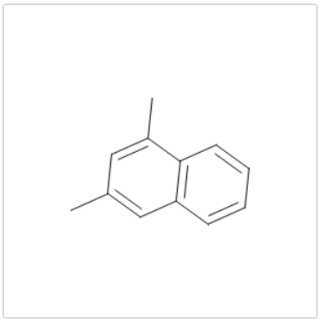

c12c3c4c5c(cccc5)c1cccc2ccc3ccc4

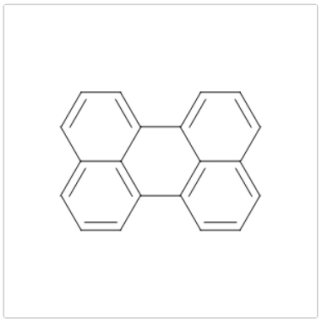

Cc(c1)ccc(c12)ccc(C)c2

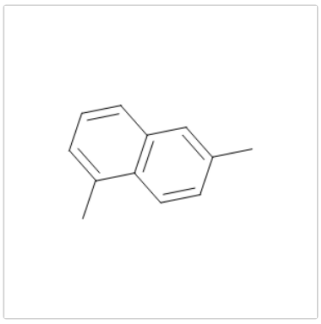

c1cccc(cc2)c1c(c2c34)ccc3c5c(cc4)cccc5

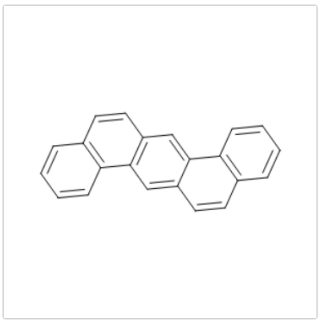

Cc1cccc(c12)ccc(C)c2

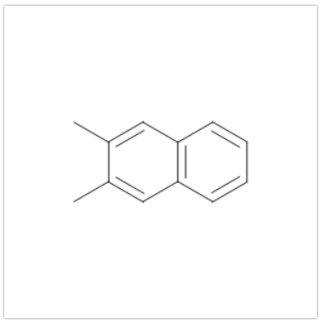

c12c3c4ccc1c5c(cccc5)cc2ccc3ccc4

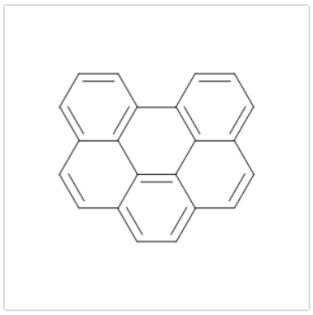

Cc1cc(C)cc(c12)cccc2

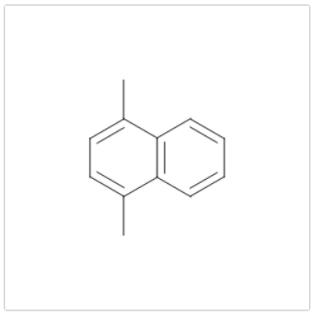

c1ccc2cccc(c2c1c3c45)c4cccc5ccc3

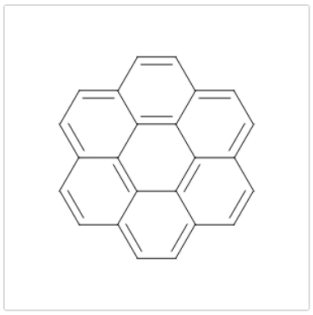

Cc1cccc(c12)cc(C)cc2

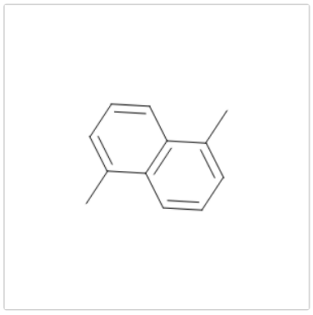

c1cccc(c1c23)ccc2cc4c5c(ccc4c3)cccc5

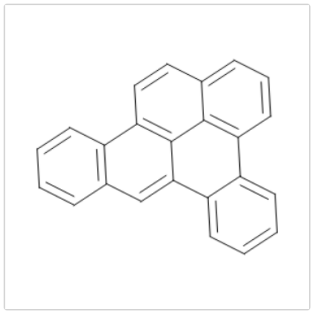

Cc(c1)c(C)cc(c12)cccc2

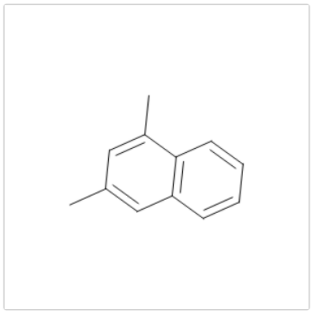

c12c3c4c5c6c1c(ccc6)ccc2ccc3ccc4ccc5

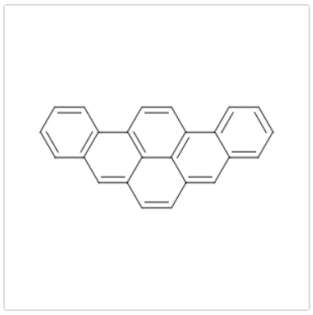

Cc1ccc(C)c(c12)cccc2

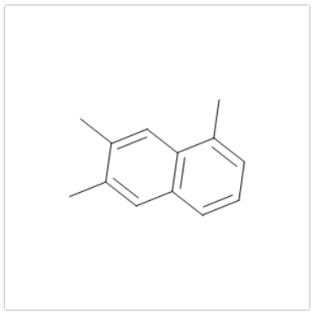

c12c3c4c5c6c1c7ccc2ccc3ccc4ccc5ccc6cc7

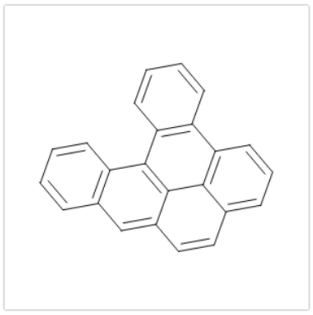

Cc1cccc(c12)c(C)ccc2

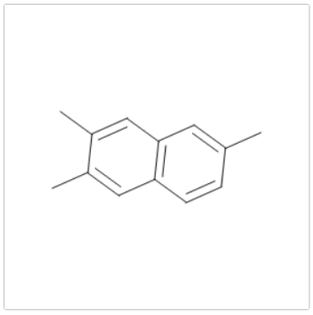

c12c3c4c5c(cccc5)c1cc6c(cccc6)c2ccc3ccc4


Cc1cc(C)cc(c12)cccc2

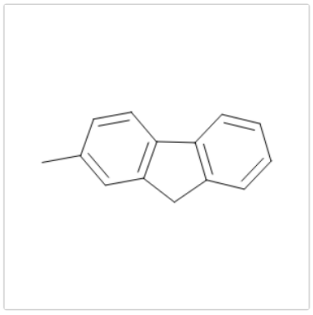

c12c3c4c5c(cccc5)cc3ccc2cc6c(c1cc4)cccc6

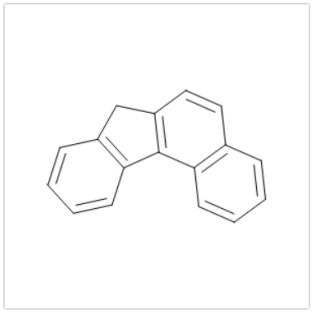

Cc(c1)c(C)cc(c12)cccc2C

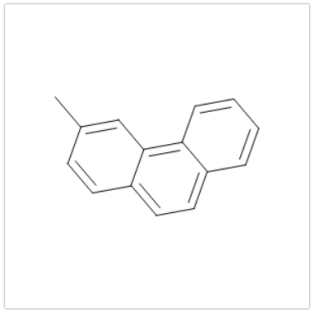

c1cccc(c1c2c34)c5c3c(ccc5)ccc4cc6c2cccc6

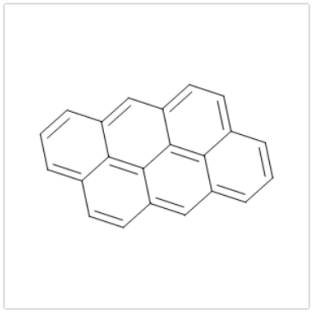

Cc(c1)c(C)cc(c12)ccc(C)c2

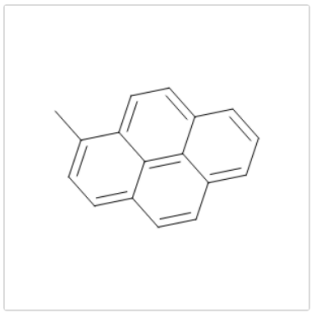

c12c3c4c5c(cccc5)cc3ccc1c6c(cc2cc4)cccc6

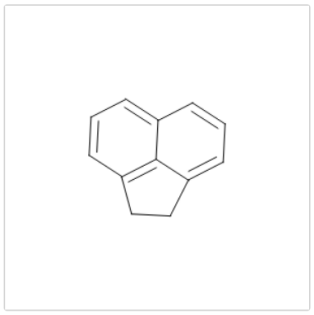

c1cc(C)cc2Cc(c3c12)cccc3

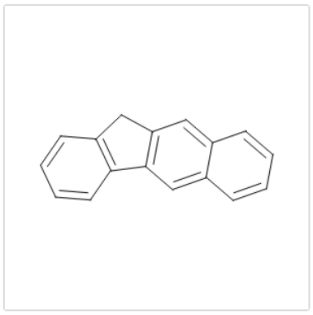

c1cccc(c1c23)Cc2ccc4c3cccc4

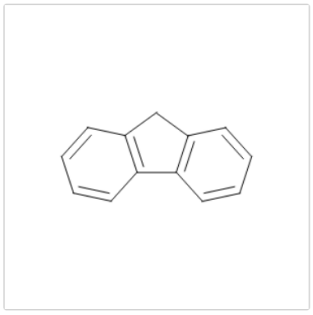

Cc(c1)ccc2ccc(c3c12)cccc3

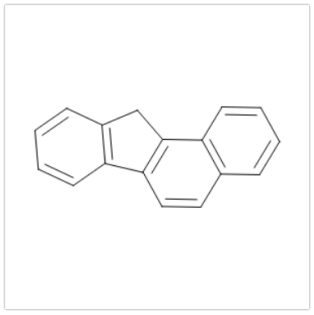

c12c3c4c5cc2ccc6c1c(ccc6)cc3ccc4ccc5


c12c3c4ccc1c(C)ccc2ccc3ccc4

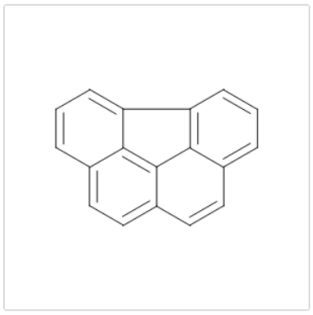

c12c3cccc2cccc1CC3

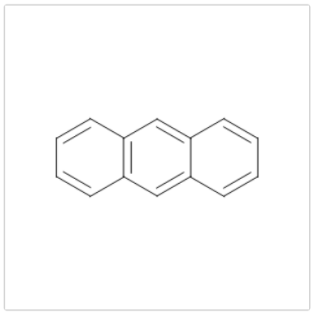

c1cccc(c1c23)Cc2cc4c(c3)cccc4

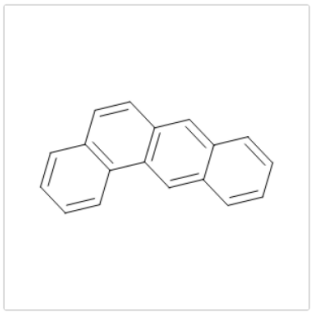

c1cccc(c1c23)Cc2cccc3

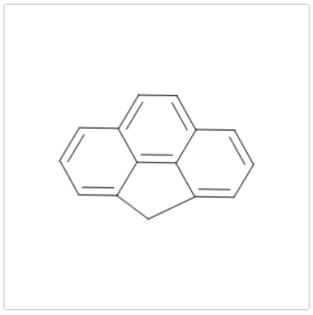

c1cccc(c1c23)Cc2c4c(cc3)cccc4

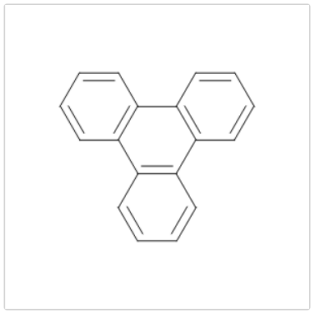

c1cccc(c1c23)ccc2cccc3

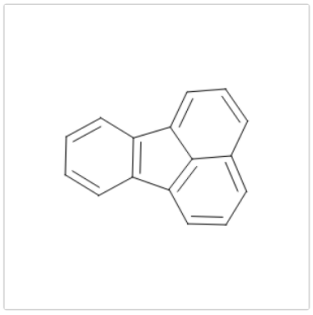

c12c3c4c5c1c(ccc5)ccc2ccc3ccc4

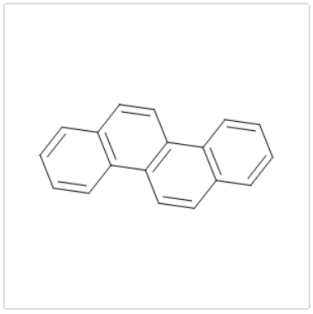

c1cccc(c12)cc3c(c2)cccc3

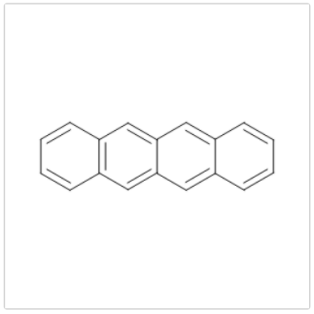

c1cccc(c1c23)ccc2cc4c(c3)cccc4

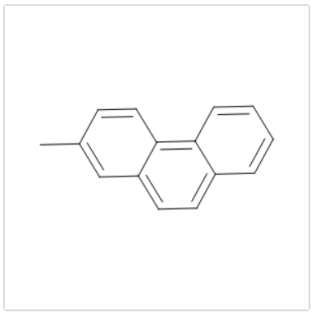

c12c3c4cccc3ccc2cccc1C4

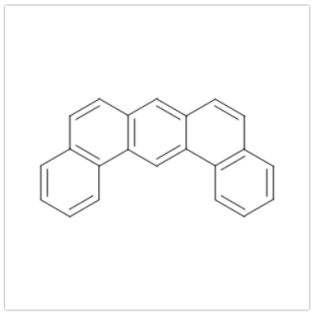

c1cccc(c1c23)c4c(cccc4)c2cccc3

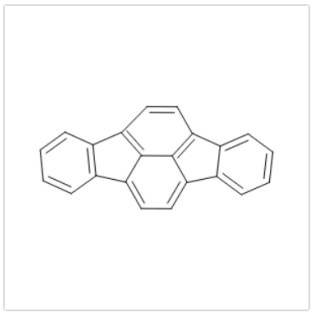

c1cccc(c1c2c34)c3cccc4ccc2

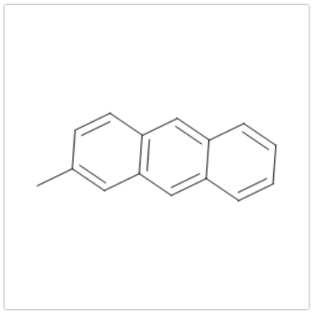

c1cccc(cc2)c1c(c23)ccc4c3cccc4

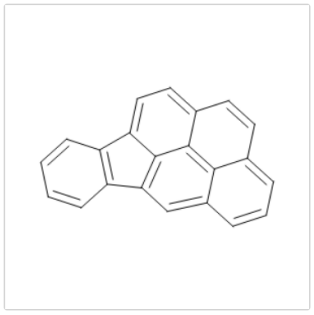

c1cccc(c2)c1cc(c23)cc4c(c3)cccc4

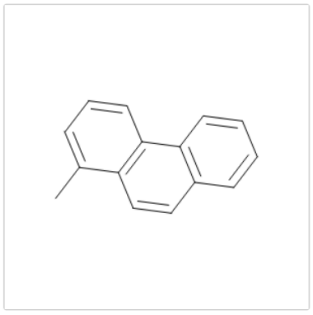

c1cc(C)cc2ccc(c3c12)cccc3

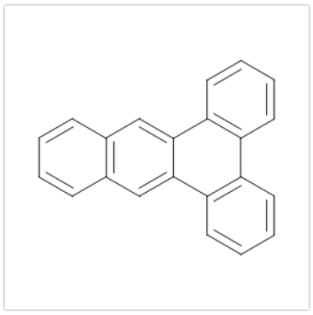

c1cccc(c1c23)ccc2cc4c(c3)c5c(cc4)cccc5

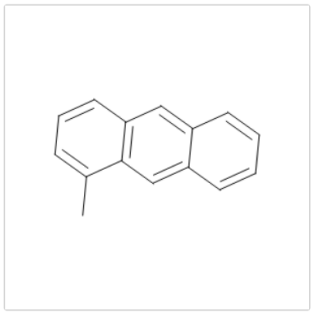

c12c3c4c5c(cccc5)c3ccc2c6c(c1cc4)cccc6

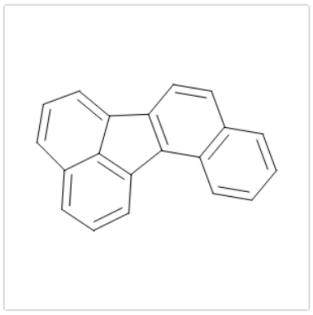

Cc(c1)ccc(c12)cc3c(c2)cccc3

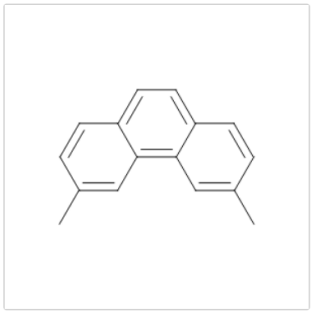

c1cccc(c1c2c34)c3ccc5c4c6c(c2)cccc6cc5

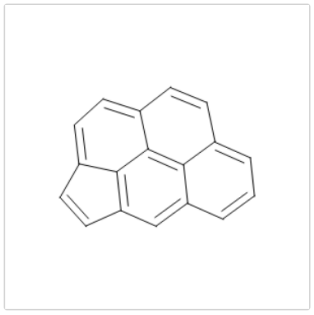

Cc1cccc(c12)c3c(cc2)cccc3

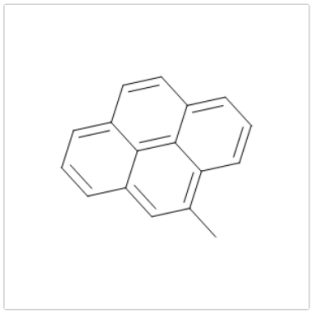

c1cccc(c2)c1cc(c23)c4c(cccc4)c5c3cccc5

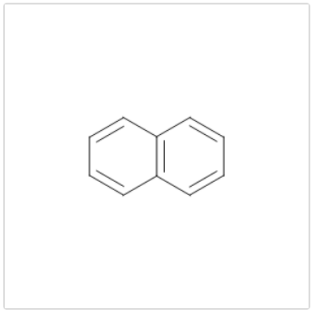

Cc1cccc(c12)cc3c(c2)cccc3

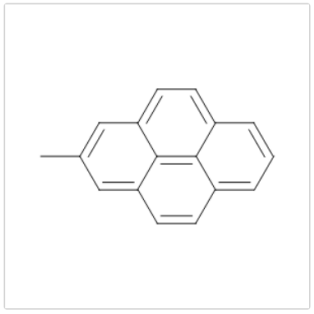

c1ccc2cccc(c2c1c34)c3ccc5c4cccc5

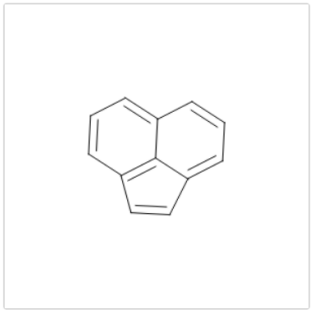

Cc(c1)ccc2ccc(c3c12)ccc(C)c3


c12c3ccc1ccc4c2c5c(c3)cccc5cc4


c12c3c4cc(C)c1cccc2ccc3ccc4


c1cccc(c12)cccc2


c12c3c4ccc1cc(C)cc2ccc3ccc4


c12c3cccc2cccc1C=C3

Priloga 2: Podatki o toksičnosti pesticidov (pLC_{50}) in ostali opisniki ter napovedane vrednosti z metodama *LR* in *RF* (preverjanje napovedne moči je bilo izvršeno z načinom *LOO*). V tabeli so zapisane številke CAS pesticidov. Kode SMILES teh spojin so precej dolge, zato smo podali številke CAS. CAS (Chemical Abstracts Service) je oddelek Ameriškega kemijskega društva, ki je pooblaščen, da vsaki kemikaliji, ki je bila kdaj opisana v literaturi, dodeli enoznačni številčni identifikator.

Št. CAS	M_r	*logP*	*HBA*	*HBD*	*RB*	pLC_{50}	LR	Δ_{LR}	RF	Δ_{RF}
56-38-2	291,26	3,28	3	0	7	8,27	5,67	2,60	5,87	2,39
58-14-0	248,71	2,75	1	2	2	4,63	5,04	0,40	4,79	0,16
78-59-1	138,21	1,95	1	0	0	3,06	4,39	1,33	4,08	1,02
80-56-8	136,23	2,87	0	0	0	3,52	4,56	1,03	4,67	1,14
83-79-4	394,42	3,93	6	0	3	8,03	7,40	0,63	6,64	1,39
88-89-1	229,10	1,25	1	1	3	3,43	4,34	0,91	5,08	1,65
90-05-1	124,14	1,55	2	1	1	3,68	4,17	0,49	4,04	0,36
95-48-7	108,14	2,05	1	1	0	3,86	4,16	0,31	3,87	0,02
96-45-7	102,16	0,48	0	2	0	3,59	3,23	0,35	3,30	0,29
97-74-5	208,37	3,61	1	0	4	4,86	5,16	0,30	4,50	0,36
101-84-8	170,21	3,39	1	0	2	5,41	5,00	0,41	5,40	0,01
107-07-3	80,51	0,31	1	1	1	2,58	3,24	0,66	2,30	0,28
120-72-9	117,15	1,61	0	1	0	5,07	3,86	1,21	4,87	0,20
135-19-3	144,17	2,47	1	1	0	4,61	4,54	0,07	4,42	0,19
139-13-9	191,14	-1,31	0	0	6	2,74	2,78	0,05	2,77	0,03
140-66-9	206,32	4,33	1	1	3	6,36	5,42	0,94	5,18	1,18
205-39-0	218,25	4,15	1	0	0	4,74	5,73	1,00	5,86	1,12
554-00-7	162,02	2,41	0	1	0	5,43	4,44	0,99	4,80	0,63
959-98-8	406,93	2,14	2	0	0	5,54	6,25	0,72	5,58	0,05
1570-64-5	142,58	2,71	1	1	0	5,69	4,63	1,06	4,96	0,73
1570-65-6	177,03	3,38	1	1	0	5,62	5,09	0,52	5,22	0,40
2668-24-8	227,47	3,54	2	1	1	5,38	5,57	0,19	4,96	0,41
13408-56-5	464,63	2,24	6	5	5	6,42	6,56	0,14	5,88	0,54
15263-53-3	237,34	0,66	4	2	7	7,38	4,33	3,05	5,64	1,73
23564-05-8	342,39	4,05	4	4	8	4,33	6,03	1,70	6,75	2,42
52645-53-1	391,29	5,44	3	0	7	7,90	7,11	0,79	7,71	0,19
57157-80-9	381,72	10,24	0	0	22	6,54	7,25	0,71	6,58	0,04
68085-85-8	449,85	5,79	4	0	7	7,22	7,79	0,56	8,68	1,46
70630-17-0	279,33	2,06	4	0	6	3,82	5,38	1,55	4,18	0,36
56-53-1	268,35	5,09	2	2	4	5,37	6,05	0,68	5,61	0,24
67-72-1	236,74	3,38	0	0	1	4,94	5,32	0,38	5,37	0,43
71-55-6	133,40	2,03	0	0	0	4,08	4,27	0,19	3,86	0,21
75-25-2	252,73	2,23	0	0	0	3,74	5,07	1,33	4,60	0,86
83-42-1	171,58	2,88	0	0	1	4,61	4,74	0,13	4,32	0,29
95-15-8	134,20	2,46	0	0	0	3,36	4,44	1,08	4,20	0,84
95-82-9	162,02	2,41	0	1	0	4,74	4,46	0,29	4,88	0,13
98-82-8	120,19	3,02	0	0	1	3,68	4,48	0,80	4,61	0,93

99-65-0	168,11	1,62	0	0	2	3,59	4,17	0,58	3,75	0,15
100-61-8	107,15	1,64	0	1	1	5,79	3,75	2,04	4,59	1,20
101-55-3	249,10	4,14	1	0	2	6,12	5,80	0,32	5,83	0,29
107-06-2	98,96	1,51	0	0	1	2,53	3,79	1,26	4,37	1,84
110-86-1	79,10	0,68	1	0	0	1,83	3,63	1,80	2,92	1,10
111-70-6	116,20	2,34	1	1	5	3,23	3,95	0,71	3,94	0,71
121-73-3	157,55	2,39	0	0	1	3,84	4,47	0,63	3,94	0,10
122-66-7	184,24	3,07	0	2	3	4,65	4,47	0,18	4,97	0,31
156-60-5	96,94	1,06	0	0	0	2,64	3,69	1,04	3,87	1,23
503-87-7	116,14	-0,07	1	2	0	3,77	3,32	0,45	3,47	0,29
532-55-8	163,20	2,46	2	0	1	4,93	4,92	0,01	4,42	0,52
622-78-6	149,21	2,60	1	0	2	6,54	4,61	1,93	5,85	0,69
636-30-6	196,46	3,08	0	1	0	4,76	4,92	0,15	5,18	0,41
693-21-0	196,12	-0,04	3	0	8	3,34	3,83	0,49	4,00	0,66
935-95-5	231,89	4,22	1	1	0	5,61	5,75	0,15	5,91	0,30
1024-57-3	389,32	3,39	1	0	0	6,21	6,54	0,33	6,72	0,51
1982-47-4	290,74	4,37	2	1	3	4,99	6,13	1,13	5,87	0,88
14315-14-1	148,22	2,95	0	0	0	4,03	4,70	0,68	4,54	0,52
33213-65-9	406,93	2,14	2	0	0	5,43	6,37	0,95	5,67	0,24
57057-83-7	227,47	3,54	2	1	1	5,49	5,59	0,10	5,33	0,16
57837-19-1	279,33	2,06	4	0	6	3,73	5,47	1,74	4,61	0,88
58-27-5	172,18	2,10	2	0	0	5,55	4,83	0,71	4,08	1,46
62-53-3	93,13	1,08	0	1	0	5,41	3,50	1,92	4,74	0,68
62-56-6	76,12	-0,21	0	2	0	3,93	2,80	1,13	2,30	1,63
63-25-2	201,22	2,70	2	1	2	5,30	5,00	0,30	5,08	0,22
67-56-1	32,04	-0,36	1	1	0	0,99	2,78	1,79	1,37	0,38
72-20-8	380,91	3,41	1	0	0	6,43	6,32	0,10	6,48	0,05
79-00-5	133,40	1,96	0	0	1	3,49	4,09	0,60	3,71	0,22
95-47-6	106,17	2,80	0	0	0	3,79	4,34	0,55	3,54	0,25
99-99-0	137,14	2,21	0	0	1	4,06	4,21	0,15	4,08	0,02
100-02-7	139,11	1,46	1	1	1	4,07	4,03	0,04	4,04	0,03
106-41-2	173,01	2,31	1	1	0	4,46	4,63	0,17	4,26	0,20
107-15-3	60,10	-1,48	0	0	1	3,36	2,32	1,04	2,36	0,99
107-92-6	88,11	0,89	0	0	2	3,16	3,34	0,18	3,66	0,50
108-18-9	101,19	1,24	0	0	2	2,35	3,55	1,20	2,89	0,53
108-42-9	127,57	1,75	0	1	0	6,11	3,95	2,15	5,77	0,34
110-02-1	84,14	1,13	0	0	0	2,42	3,55	1,13	3,30	0,87
118-79-6	330,80	3,81	1	1	0	6,91	6,11	0,80	6,36	0,54
120-83-2	163,00	2,89	1	1	0	4,80	4,80	0,01	4,51	0,29
149-30-4	167,25	3,31	0	1	0	4,49	4,79	0,30	5,16	0,67
534-52-1	198,13	1,84	1	1	2	4,79	4,45	0,35	4,18	0,61
542-75-6	110,97	1,52	0	0	1	4,39	3,79	0,60	2,78	1,61
542-85-8	87,14	1,36	1	0	1	5,31	3,79	1,51	3,15	2,15
586-62-9	136,23	3,64	0	0	0	4,73	4,84	0,11	5,14	0,41

709-98-8	218,08	2,95	1	1	2	4,64	5,00	0,36	4,96	0,32
2741-06-2	180,27	2,86	0	2	4	3,35	4,32	0,97	4,58	1,23
6972-05-0	104,17	0,76	0	1	1	3,40	3,36	0,03	3,87	0,47
33813-20-6	176,28	2,54	1	0	0	5,92	4,83	1,09	4,52	1,41
76738-62-0	293,79	2,66	3	1	5	4,01	5,50	1,49	7,15	3,14
91465-08-6	449,85	5,79	4	0	7	8,64	7,77	0,87	8,36	0,28
52-68-6	257,44	1,13	4	1	4	9,09	4,83	4,26	5,32	3,77
76-44-8	373,32	4,24	0	0	0	6,51	6,38	0,14	6,40	0,12
79-01-6	131,39	1,74	0	0	0	3,38	4,09	0,71	3,36	0,01
88-85-7	240,21	3,00	1	1	4	6,00	4,97	1,03	5,08	0,92
89-59-8	171,58	2,88	0	0	1	4,27	4,68	0,41	4,62	0,35
95-50-1	147,00	3,16	0	0	0	4,81	4,74	0,06	4,95	0,14
97-00-7	202,55	2,28	0	0	2	5,40	4,52	0,88	4,14	1,27
97-77-8	296,54	5,60	0	0	9	6,39	5,78	0,62	5,86	0,54
103-85-5	152,22	2,08	0	2	2	3,41	4,09	0,68	4,59	1,17
106-46-7	147,00	3,16	0	0	0	4,13	4,74	0,61	4,95	0,82
107-21-1	62,07	-0,90	2	2	1	0,12	2,84	2,72	2,82	2,71
108-38-3	106,17	2,80	0	0	0	3,47	4,38	0,92	3,69	0,22
108-39-4	108,14	2,05	1	1	0	3,76	4,22	0,46	3,93	0,17
108-88-3	92,14	2,32	0	0	0	2,98	4,11	1,14	4,14	1,17
121-75-5	330,36	2,16	6	0	11	7,66	5,44	2,22	5,22	2,43
124-18-5	142,28	4,93	0	0	7	3,90	4,86	0,96	5,01	1,11
206-44-0	202,25	3,95	0	0	0	6,26	5,35	0,91	5,77	0,49
589-16-2	121,18	2,03	0	1	1	6,13	4,03	2,10	4,66	1,47
592-82-5	115,20	2,34	1	0	3	5,43	4,17	1,26	3,97	1,45
634-67-3	196,46	3,08	0	1	0	5,43	4,93	0,50	4,67	0,76
1241-94-7	362,40	6,38	4	0	11	6,38	6,96	0,57	7,69	1,30
1836-77-7	318,54	5,28	1	0	3	5,88	6,43	0,55	6,83	0,95
2051-61-8	188,65	4,01	0	0	1	5,64	5,22	0,42	5,64	0,01
29232-93-7	305,33	3,11	5	0	7	9,15	5,84	3,31	7,54	1,61
51630-58-1	419,90	6,56	4	0	7	7,99	7,67	0,32	9,16	1,17
52315-07-8	416,30	5,30	4	0	6	9,06	7,22	1,84	8,27	0,79
52918-63-5	505,20	5,52	4	0	6	10,06	7,78	2,28	8,08	1,98
960003-91-2	283,35	2,73	4	0	6	5,95	5,48	0,46	5,79	0,16
58-90-2	231,89	4,22	1	1	0	6,01	5,70	0,31	5,78	0,23
75-08-1	62,13	0,96	0	1	0	5,56	3,25	2,31	4,32	1,25
75-21-8	44,05	-0,13	1	0	0	2,34	3,03	0,69	2,48	0,13
78-87-5	112,99	1,89	0	0	1	3,31	3,96	0,65	3,65	0,34
92-83-1	182,22	3,53	1	0	0	3,95	5,26	1,30	5,48	1,53
95-95-4	197,45	3,56	1	1	0	4,92	5,23	0,31	5,27	0,35
100-00-5	157,55	2,39	0	0	1	4,31	4,42	0,11	3,97	0,34
103-72-0	135,19	2,59	1	0	1	6,13	4,53	1,60	4,55	1,58
108-44-1	107,15	1,57	0	1	0	5,17	3,76	1,41	4,98	0,19
108-85-0	163,06	2,85	0	0	0	3,89	4,71	0,82	4,62	0,73

111-42-2	105,14	-1,29	2	2	4	2,98	2,60	0,38	2,70	0,28
115-29-7	406,93	2,14	2	0	0	6,13	6,32	0,19	5,44	0,69
121-29-9	372,45	3,74	5	0	9	7,34	6,51	0,83	7,03	0,32
129-00-0	202,25	3,95	0	0	0	6,04	5,36	0,68	5,64	0,40
132-65-0	184,26	3,80	0	0	0	5,62	5,19	0,43	4,93	0,69
138-86-3	136,23	3,50	0	0	1	6,64	4,70	1,94	4,43	2,22
148-01-6	225,16	1,08	1	1	3	3,14	4,25	1,11	4,97	1,84
333-41-5	304,35	3,51	5	0	7	8,18	6,16	2,02	7,04	1,14
630-20-6	167,85	2,44	0	0	1	3,85	4,51	0,66	3,63	0,21
632-22-4	116,16	0,83	1	0	0	1,60	3,84	2,24	3,79	2,18
1071-83-6	169,07	-1,59	2	0	4	3,91	3,13	0,78	2,77	1,14
1455-18-1	148,22	2,95	0	0	0	4,76	4,65	0,11	4,23	0,53
1646-88-4	222,26	0,35	5	1	4	5,90	4,59	1,31	5,89	0,01
3209-22-1	192,00	3,05	0	0	1	4,63	4,88	0,26	4,62	0,01
3483-12-3	154,25	0,02	2	4	3	3,76	3,22	0,53	5,21	1,45
35367-38-5	310,68	3,94	2	2	2	7,84	5,98	1,86	5,41	2,43
98886-44-3	283,35	2,73	4	0	6	5,63	5,64	0,01	5,91	0,28
138261-41-3	255,66	1,05	2	0	3	4,17	4,73	0,56	5,50	1,33
50-29-3	354,49	6,33	0	0	3	7,47	6,99	0,48	6,45	1,02
58-89-9	290,83	4,16	0	0	0	5,21	5,95	0,74	5,65	0,44
72-43-5	345,65	4,97	2	0	5	7,34	6,65	0,69	5,82	1,52
78-99-9	112,99	1,90	0	0	1	3,69	3,95	0,26	3,42	0,27
80-05-7	228,29	3,73	2	2	2	4,25	5,52	1,27	4,36	0,11
86-30-6	198,22	3,88	0	0	3	4,41	5,11	0,70	5,22	0,81
90-13-1	162,62	3,40	0	0	0	5,01	4,89	0,11	5,06	0,06
103-69-5	121,18	1,99	0	1	2	5,46	3,90	1,56	4,95	0,51
104-94-9	123,15	1,07	1	1	1	5,57	3,79	1,78	4,72	0,86
105-67-9	122,16	2,54	1	1	0	4,77	4,43	0,33	4,41	0,35
117-81-7	390,56	7,57	4	0	16	4,55	7,64	3,09	7,38	2,83
120-93-4	86,09	-0,42	1	2	0	1,19	2,98	1,80	2,13	0,94
122-14-5	277,23	3,06	3	0	5	7,75	5,68	2,08	5,47	2,28
126-73-8	266,31	4,19	4	0	12	4,86	5,81	0,95	5,34	0,48
534-13-4	104,17	0,66	0	2	2	3,80	3,21	0,59	3,47	0,33
576-26-1	122,16	2,54	1	1	0	4,04	4,43	0,39	4,41	0,37
626-43-7	162,02	2,41	0	1	0	5,16	4,43	0,73	5,02	0,14
1031-07-8	422,92	2,24	2	0	0	5,30	6,33	1,03	5,71	0,41
1897-45-6	265,91	4,25	2	0	0	6,22	6,20	0,02	5,88	0,33
o1918-02-1	241,46	2,17	1	1	1	3,68	4,93	1,25	4,79	1,11
2257-09-2	163,24	2,92	1	0	3	6,10	4,70	1,39	5,06	1,03
2303-17-5	304,66	4,48	2	0	5	6,73	6,22	0,51	6,92	0,19
2437-79-8	291,99	6,01	0	0	1	6,99	6,62	0,37	6,39	0,60
4104-75-0	166,24	2,43	0	1	2	3,24	4,34	1,11	4,18	0,95
25875-51-8	334,20	4,04	3	1	6	6,65	6,27	0,38	6,43	0,22
66230-04-4	419,90	6,56	4	0	7	9,19	7,96	1,23	8,06	1,14

112410-23-8	352,47	4,92	2	1	5	4,31	6,60	2,30	5,95	1,64
960003-90-1	283,35	2,73	4	0	6	6,11	5,70	0,41	5,75	0,36
56-23-5	153,82	3,58	0	0	0	3,64	4,96	1,31	4,95	1,30
71-23-8	60,10	0,51	1	1	1	0,94	3,24	2,30	3,93	2,98
75-05-8	41,05	0,04	1	0	0	1,06	3,16	2,10	3,46	2,40
77-73-6	132,20	2,11	0	0	0	4,10	4,25	0,15	3,74	0,36
79-34-5	167,85	2,40	0	0	1	3,47	4,48	1,01	4,05	0,58
83-41-0	151,16	2,70	0	0	1	4,56	4,51	0,05	4,69	0,13
91-22-5	129,16	2,02	1	0	0	3,53	4,42	0,89	3,84	0,31
92-69-3	170,21	3,08	1	1	1	4,67	4,87	0,20	4,62	0,05
95-51-2	127,57	1,75	0	1	0	5,46	3,96	1,50	5,58	0,12
108-95-2	94,11	1,56	1	1	0	3,85	3,93	0,08	5,03	1,18
109-89-7	73,14	0,48	0	0	2	3,12	3,11	0,00	3,06	0,05
111-90-0	134,17	-0,27	3	1	6	1,53	3,38	1,85	2,17	0,64
116-29-0	356,05	5,59	2	0	2	4,71	7,15	2,44	6,99	2,27
118-96-7	227,13	2,00	0	0	3	4,41	4,48	0,07	4,39	0,02
131-11-3	194,18	1,54	4	0	4	3,77	4,94	1,17	6,44	2,67
137-26-8	240,43	4,20	0	0	5	6,06	5,27	0,79	5,28	0,78
142-96-1	130,23	2,71	1	0	6	3,70	4,21	0,51	5,34	1,64
536-90-3	123,15	1,07	1	1	1	5,64	3,81	1,83	4,39	1,25
578-54-1	121,18	2,03	0	1	1	4,18	3,95	0,22	5,13	0,95
625-53-6	104,17	0,57	0	2	2	3,87	3,08	0,79	3,65	0,23
825-44-5	166,20	1,04	2	0	0	4,07	4,47	0,39	3,31	0,77
2051-62-9	188,65	4,01	0	0	1	5,65	5,24	0,42	5,62	0,03
2764-72-9	184,24	2,50	0	0	0	5,14	4,69	0,44	5,08	0,06
4044-65-9	192,26	3,35	2	0	2	6,40	5,36	1,04	5,01	1,38
4180-23-8	148,20	2,77	1	0	2	7,34	4,66	2,67	5,24	2,10
10265-92-6	141,13	-0,12	2	1	2	6,62	3,58	3,04	3,29	3,33
15862-07-4	257,54	5,34	0	0	1	5,64	6,14	0,50	6,18	0,54
82657-04-3	422,87	6,37	2	0	7	8,33	7,41	0,91	7,84	0,49
51-28-5	184,11	1,35	1	1	2	4,59	4,25	0,34	3,94	0,65
55-63-0	227,09	0,08	3	0	8	3,85	4,07	0,22	5,38	1,53
60-51-5	229,26	0,73	3	1	5	5,14	4,45	0,69	5,69	0,55
78-83-1	74,12	0,83	1	1	1	1,83	3,46	1,63	3,24	1,42
79-06-1	71,08	-0,27	1	1	1	2,65	3,03	0,38	2,34	0,31
83-32-9	154,21	3,34	0	0	0	5,38	4,84	0,54	4,99	0,39
84-66-2	222,24	2,24	4	0	6	3,61	5,17	1,56	5,24	1,63
85-01-8	178,23	3,65	0	0	0	5,37	5,10	0,26	4,52	0,85
88-06-2	197,45	3,56	1	1	0	5,15	5,29	0,14	5,32	0,18
98-95-3	123,11	1,72	0	0	1	3,66	3,98	0,32	3,86	0,20
99-08-1	137,14	2,21	0	0	1	4,08	4,25	0,17	3,94	0,14
102-08-9	228,31	4,37	0	2	4	3,53	5,23	1,70	5,62	2,09
106-44-5	108,14	2,05	1	1	0	3,71	4,19	0,48	4,06	0,35
106-47-8	127,57	1,75	0	1	0	6,41	4,01	2,40	5,64	0,77

115-86-6	326,28	4,94	4	0	6	5,51	6,81	1,30	5,96	0,45
121-87-9	172,57	1,64	0	1	1	4,49	4,17	0,32	4,66	0,17
142-28-9	112,99	1,57	0	0	2	2,61	3,79	1,19	3,74	1,13
143-50-0	490,64	4,50	1	0	0	6,39	7,51	1,12	7,47	1,08
611-06-3	192,00	3,05	0	0	1	4,66	4,90	0,24	4,53	0,13
634-90-2	215,89	4,49	0	0	0	5,13	5,65	0,51	5,96	0,82
680-31-9	179,20	-1,12	0	0	3	1,43	3,13	1,70	3,58	2,15
877-43-0	157,21	2,79	1	0	0	3,62	4,84	1,22	5,27	1,65
1014-70-6	213,30	2,37	3	2	5	3,63	4,88	1,25	4,91	1,28
1582-09-8	335,28	4,47	0	0	8	6,24	5,80	0,44	5,72	0,52
2921-88-2	350,59	4,98	4	0	6	8,55	6,97	1,57	6,48	2,07
5289-74-7	480,63	0,89	7	6	5	5,29	6,38	1,08	5,89	0,60
10605-21-7	191,19	1,79	3	2	2	6,27	4,74	1,53	4,16	2,11
51218-45-2	283,79	3,36	2	0	6	4,34	5,59	1,25	6,10	1,77
56-55-3	228,29	4,56	0	0	0	6,22	5,69	0,53	5,20	1,02
57-74-9	409,78	4,79	0	0	0	6,62	6,86	0,24	5,95	0,67
60-00-4	292,24	-1,72	0	0	11	3,10	2,61	0,50	6,35	3,24
62-73-7	220,98	1,03	4	0	4	9,42	4,95	4,46	4,17	5,25
87-86-5	266,34	4,88	1	1	0	5,52	6,24	0,72	6,02	0,49
88-73-3	157,55	2,39	0	0	1	3,66	4,35	0,68	4,13	0,47
90-04-0	123,15	1,07	1	1	1	4,01	3,83	0,18	5,43	1,42
92-52-4	154,21	3,35	0	0	1	4,91	4,69	0,21	5,36	0,45
95-57-8	128,56	2,23	1	1	0	4,24	4,40	0,17	4,49	0,25
99-86-5	136,23	3,45	0	0	1	4,87	4,63	0,24	5,70	0,83
100-42-5	104,15	2,38	0	0	1	3,45	4,03	0,58	3,67	0,22
100-52-7	106,12	1,59	1	0	1	3,98	3,99	0,02	4,05	0,08
101-20-2	315,58	5,46	1	2	2	7,48	6,50	0,98	6,84	0,64
107-11-9	57,09	-0,03	0	0	1	3,13	2,82	0,31	1,67	1,46
111-91-1	173,04	1,23	2	0	6	2,94	4,03	1,10	3,89	0,96
115-20-8	149,40	1,24	1	1	1	3,00	4,05	1,05	4,09	1,08
116-06-3	190,26	1,36	4	1	4	5,51	4,83	0,68	4,52	0,99
150-19-6	124,14	1,55	2	1	1	3,48	4,28	0,80	4,33	0,85
298-00-0	263,21	2,58	3	0	5	7,24	5,44	1,80	5,53	1,70
634-83-3	230,91	3,74	0	1	0	5,56	5,33	0,22	5,47	0,09
2539-17-5	261,92	4,20	2	1	1	6,27	6,12	0,16	6,02	0,25
11141-17-6	720,71	-1,64	16	3	10	6,06	9,23	3,17	7,48	1,41
22431-62-5	364,48	5,25	3	0	7	8,16	6,87	1,29	6,49	1,67
35723-83-2	464,63	2,24	6	5	5	5,88	6,96	1,09	6,63	0,75
51235-04-2	252,31	3,77	3	0	2	3,22	6,12	2,90	6,18	2,96
66841-25-6	665,01	6,93	4	0	7	9,95	9,56	0,39	9,12	0,82
131860-33-8	403,39	3,69	8	0	9	6,61	7,61	0,99	6,97	0,36
960003-92-3	283,35	2,73	4	0	6	5,76	5,78	0,02	6,00	0,24
67-64-1	58,08	-0,24	1	0	0	0,67	3,09	2,42	1,74	1,07
67-66-3	119,38	1,62	0	0	0	3,25	3,96	0,71	3,49	0,24

84-74-2	278,34	4,20	4	0	10	4,88	6,05	1,18	6,04	1,16
85-68-7	312,36	4,45	4	0	9	4,39	6,41	2,03	7,09	2,70
89-61-2	192,00	3,05	0	0	1	4,24	4,87	0,62	4,76	0,52
95-53-4	107,15	1,57	0	1	0	5,31	3,75	1,56	5,02	0,30
100-41-4	106,17	2,77	0	0	1	3,76	4,28	0,52	3,57	0,19
105-55-5	132,23	1,36	0	2	4	2,84	3,39	0,56	4,18	1,35
106-42-3	106,17	2,80	0	0	0	3,52	4,37	0,85	3,65	0,14
106-89-8	92,52	0,56	1	0	1	3,60	3,54	0,06	2,90	0,70
107-03-9	76,16	1,49	0	1	1	6,10	3,48	2,63	3,40	2,71
108-90-7	112,56	2,49	0	0	0	3,94	4,28	0,34	3,62	0,32
114-26-1	209,24	2,50	3	1	4	5,22	5,07	0,15	5,57	0,35
124-40-3	45,08	-0,22	0	0	0	2,96	2,81	0,14	2,92	0,03
127-18-4	165,83	2,43	0	0	0	3,96	4,54	0,58	3,99	0,03
141-78-6	88,11	0,37	2	0	2	2,08	3,59	1,51	3,64	1,57
205-43-6	234,32	4,71	0	0	0	6,03	5,84	0,19	5,85	0,17
506-77-4	61,47	0,13	1	0	0	6,15	3,27	2,88	1,82	4,33
598-52-7	90,15	0,23	0	2	1	3,98	2,92	1,06	4,14	0,17
602-01-7	182,13	2,11	0	0	2	5,44	4,36	1,08	4,70	0,74
608-93-5	250,34	5,15	0	0	0	4,67	6,11	1,43	5,64	0,97
618-62-2	192,00	0,90	0	0	1	4,41	4,00	0,41	4,14	0,26
1516-32-1	132,23	1,55	0	2	4	3,85	3,48	0,37	4,22	0,37
3547-04-4	251,15	5,39	0	0	2	6,86	6,06	0,79	5,22	1,64
30125-65-6	213,30	2,04	3	2	3	4,29	4,85	0,57	5,30	1,02
40596-69-8	310,47	4,95	3	0	11	5,48	6,24	0,76	7,01	1,53
68359-37-5	434,29	5,51	4	0	6	8,86	7,72	1,14	8,19	0,67