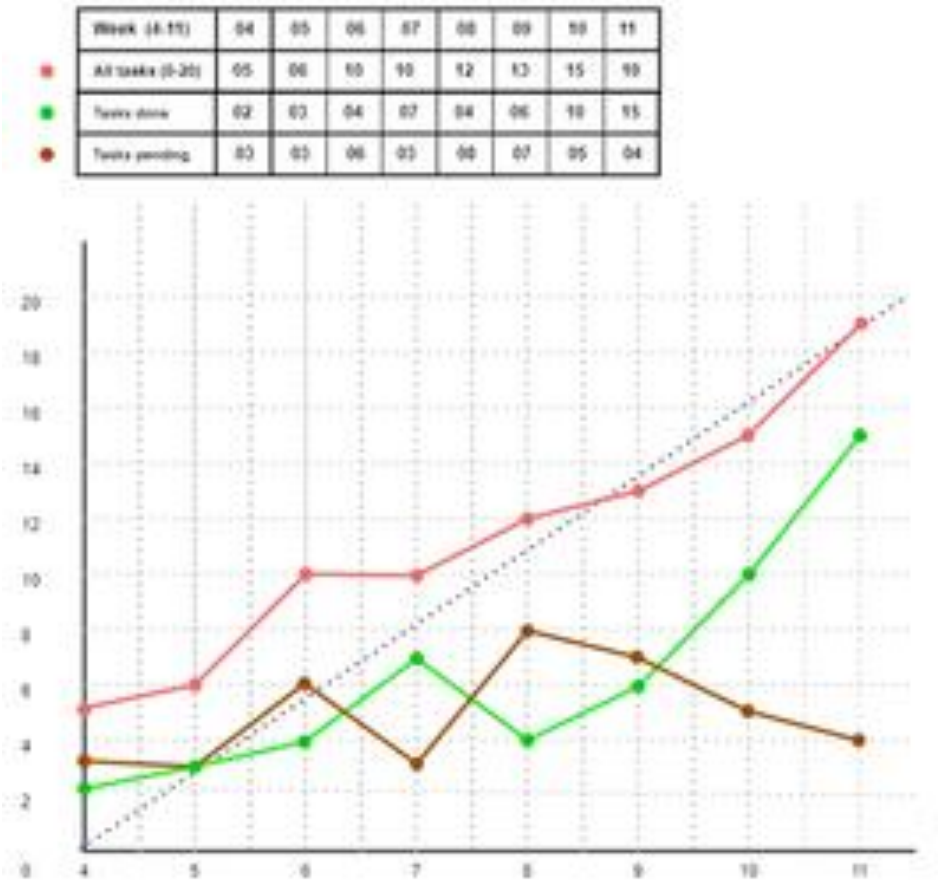
Šprinty

Celkový pokrok práce a vývoj úloh možno vidieť na nasledujúcom grafe:



https://drive.google.com/file/d/1hxJ-sRTlOnNS75TZ7NfIGWV-iGpGFfZh/view?usp=sharing

Šprint číslo 1:

Dátum - October 11th, 2023 - October 25th, 2023

Hlavný cieľ šprintu - spoznať Jiru a zorientovať sa v problematike.

Výsledky šprintu

V tomto šprinte sme mali nasledovné úlohy a výsledky:

Všetky úlohy (6)

- Naštudovať si OECD guidelines
- Vytvoriť Jiru
- Rozdeliť si role - kto sa bude venovať datasetu a kto QSARu
- Vytvoriť vyrovnaný dataset
- Nájsť čo najviac molekulárnych deskriptorov - RDKit / CDKit
- Vytvoriť webovú stránku pre tímak

Dokončené úlohy (3)

- Naštudovať si OECD guidelines
- Vytvoriť Jiru
- Nájsť čo najviac molekulárnych deskriptorov - RDKit / CDKit

Rozrobené úlohy (3)

- Rozdeliť si role - kto sa bude venovať datasetu a kto QSARu
- Vytvoriť vyrovnaný dataset
- Vytvoriť webovú stránku pre tímak

Poznámky k šprintu

Výzvy

- Najväčší problém bolo nájsť in vivo datasety alebo údaje na vytvorenie in vivo datasetu.
- Výber deskriptorov a identifikácia ich významu podľa ich názvov
- Niektoré dáta mali viacero mien, čo robilo problémy pri vytváraní SMILES kódov
- Malá dostupnosť negatívnych výsledkov pre Amesov test

Poznatky

- Zistili sme, ako z kódov v datasete vytvárať SMILES kódy
- Práca s datasetmi

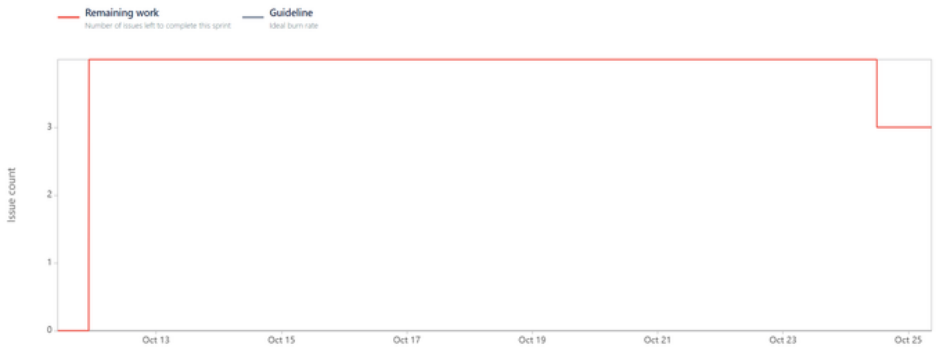
Zhrnutie

V období od 11. októbra 2023 do 25. októbra 2023 sme si stanovili konkrétne úlohy. Medzi naše hlavné ciele patrilo štúdium smerníc OECD, vytvorenie projektu v Jira, rozdelenie rolí v tíme pre spracovanie dátových sád a prácu s QSAR (kvantitatívna štruktúrovaná-vzťahová analýza). Rovnako sme si kladli za úlohu vytvoriť vyváženú sadu dát a identifikovať čo najviac molekulárnych deskriptorov pomocou nástrojov RDKit a CDKit. Okrem toho sme sa venovali aj vytváraniu webovej stránky pre efektívnu komunikáciu v tíme.

Počas tohto šprintu sme úspešne dokončili niekoľko kľúčových úloh, vrátane štúdia smerníc OECD, vytvorenia projektu v Jira a identifikácie molekulárnych deskriptorov pomocou RDKit a CDKit. Napriek tomu sme čelili niekoľkým výzvam, ako napríklad hľadanie in vivo dátových sád a identifikácia deskriptorov na základe ich názvov. Niektoré údaje mali viacero názvov, čo komplikovalo proces vytvárania SMILES kódov. Taktiež sme sa stretli s obmedzenou dostupnosťou negatívnych výsledkov pre Amesov test.

Napriek tomu sme získali cenné poznatky, najmä v oblasti tvorby SMILES kódov z kódov v dátových sadách a práce s rôznymi typmi datasetov. Tento sprint nám poskytol pevné základy pre budúcu prácu a zdokonalenie našich metód v oblasti chemoinformatiky a kvantitatívnej analýzy štruktúry.

Burndown graf



Šprint číslo 2:

Dátum - October 25th, 2023 - November 7th, 2023

Hlavný cieľ šprintu - hľadanie dát

Výsledky šprintu

V tomto šprinte sme mali nasledovné ciele a výsledky:

Nové úlohy (7)

- Dať preč 3D a 4D deskriptory
- Nájsť články a dáta, ktoré majú priložené na konci ako supplement data + pozieť vedecké články o QSARoch, kde bude genotoxicita - aspoň 3
- K datasetu - pozrieť, ktoré parametre má RDkit a nemá CDkit
- Zistiť korelácie medzi hodnotami - genotoxicita vs deskriptor - treba spraviť korelácie / regriesie a nájsť najvyššie R (spraviť: support vector machine, lineárne regresia, random forest, gradient boosting machine, knn, shapley value).
- Pozrieť článok o LamPS na Githube - či nemá 2D deskriptory
- Založiť GitHub repo
- Pridať biogénne látky ako proteín a aminokyseliny, cukry, vodu atd

Úlohy z predchádzajúceho šprintu (3)

- Rozdeliť si role - kto sa bude venovať datasetu a kto QSARu
- Vytvoriť vyrovnaný dataset
- Vytvoriť webovú stránku pre tímak

Dokončené úlohy (7)

- Rozdeliť si role - kto sa bude venovať datasetu a kto QSARu
- Dať preč 3D a 4D deskriptory
- Nájsť články a dáta, ktoré majú priložené na konci ako supplement data + pozieť vedecké články o QSARoch, kde bude genotoxicita - aspoň 3
- K datasetu - pozrieť, ktoré parametre má RDkit a nemá CDkit
- Zistiť korelácie medzi hodnotami - genotoxicita vs deskriptor - treba spraviť korelácie / regriesie a nájsť najvyššie R (spraviť: support vector machine, lineárne regresia, random forest, gradient boosting machine, knn, shapley value).
    - Korelácie deskriptory rdkit, cdkit (1d,2d) vs. Ames
    - SVM, LR, RF, GBM, KNN a shapley value (treba tuning + oversampling, výsledky modelov zatiaľ nedostačujúce)
    - Feature selection, oversampling, evaluation metrics
- Pozrieť článok o LamPS na Githube - či nemá 2D deskriptory

- Založiť GitHub repo

Rozrobené úlohy (3)

- Vytvoriť webovú stránku pre tímak
- Vytvoriť vyrovnaný dataset
- Pridať biogénne látky ako proteín a aminokyseliny, cukry, vodu atd

Poznámky k šprintu

Výzvy

- Hodnoty pre in vivo dataset nie sú v dostatočnej miere dostupné na internete

Poznatky

- Mutagenita = genotoxicita
- In vivo dataset je možné poskladať z biogénnych látok
- Robiť modely má zmysel až s dostatočným množstvom dát
- Vytvorili sme si základ pre dataset v Excel tabuľke
- 3D a 4D deskriptory pre náš účel nie sú vhodné
- Našli sme užitočné zdroje:
    - https://www.ebi.ac.uk/chembl/
    - https://onlinelibrary.wiley.com/doi/epdf/10.1002/%28SICI%291098-2280%282000%2935%3A3%3C206%3A%3AAID-EM8%3E3.0.CO%3B2-J
    - PaDEL descriptor software - https://github.com/ecrl/padelpy#padelpy-a-python-wrapper-for-padel-descriptor-software
    - https://docs.lammps.org/dimension.html

Zhrnutie

Šprint číslo 2, ktorý prebiehal od 25. októbra 2023 do 7. novembra 2023, bol zameraný na dôležité hľadanie dát a vývoj v projekte. Hlavným cieľom bolo rozšírenie našich poznatkov o QSAR (kvantitatívna štruktúrovaná-vzťahová analýza) s dôrazom na genotoxicitu. Vytýčili sme si 7 nových úloh, vrátane odstránenia 3D a 4D deskriptorov (ktoré sme zistili, že pre náš účel nie sú vhodné), vyhľadávania vedeckých článkov o QSAR s genotoxickým obsahom, a skúmania parametrov RDKit a CDKit v datasete. Dôležitou úlohou bolo aj hľadanie korelácií medzi genotoxicitou a deskriptormi, vrátane implementácie rôznych modelov ako support vector machine, lineárna regresia, random forest a ďalšie.

V priebehu šprintu sme úspešne dokončili 7 úloh vrátane rozdelenia rolí v tíme, odstránenia 3D a 4D deskriptorov a skúmania vedeckých článkov. Zároveň sme čelili výzvam, ako je nedostatok hodnôt pre in vivo dataset dostupných na internete. Získali sme ďalšie poznatky o genotoxicite a zistili sme, že je možné zostaviť in vivo dataset z biogénnych látok.

Vytvorili sme základy pre dataset v Excel tabuľke a zhromaždili užitočné zdroje informácií vrátane PaDEL descriptor software. Napriek nedostatočným výsledkom modelov sme si uvedomili, že vytváranie efektívnych modelov vyžaduje dostatočné množstvo dát a že tieto výsledku sú ešte len v počiatočných štádiach. Zároveň sme rozpracovali úlohy, ktoré sa týkali vytvorenia webovej stránky pre tím, vytvorenia vyváženého datasetu a pridania biogénnych látok.

Celkovo sme dosiahli podstatný pokrok v smerovaní projektu a rozvíjali sme náš vhľad do oblasti QSAR s dôrazom na genotoxicitu. Šprint nám priniesol nové výzvy a poznatky, ktoré budú dôležité pre ďalší rozvoj a zdokonalenie našej práce.

Burndown graf

Šprint číslo 3:

Dátum - November 8th, 2023 - November 21st, 2023

Hlavný cieľ šprintu - dokončenie in vivo datasetu

Výsledky šprintu

V tomto šprinte sme mali nasledovné ciele a výsledky:

Nové úlohy (10)

- Homolumo deskriptor - zistiť ako sa počíta a spraviť kód pre výpočet deskriptora
- Webstranka pre tím - vytvoriť nejaký nápad / návrh
- Pozrieť isomeric / caonical smiles (čo znamenajú zavináče v smiles kódoch)
- Integrácia cdkitu a padelu do kódu, ktorý generuje descriptory
- Pridať biogénne prvky - Sacharóza a Erythritol
- Zistiť, či sú sladidlá negenotoxické a ak nie tak pridať do datasetu
- Pridať roly členom tímu na stránke
- Pridať fotky členov tímu na stránku
- Skúsiť zistiť čo znamenajú Padel descriptory
- Začať trénovať model

Úlohy z predchádzajúceho šprintu (3)

- Vytvoriť webovú stránku pre tímak

- Vytvoriť vyrovnaný dataset
- Pridať biogénne látky ako proteín a aminokyseliny, cukry, vodu atd

Dokončené úlohy (6)

- Vytvoriť webovú stránku pre tímak
- Pridať biogénne látky ako proteín a aminokyseliny, cukry, vodu atd
- Pozrieť isomeric / caonical smiles (zavináče v smiles)
- Pridať biogénne prvky - Sacharóza a Erythritol
- Zistiť, či sú sladidlá negenotoxické a ak nie tak pridať do datasetu
- Pridať roly členom tímu na stránke

Rozrobené úlohy (7)

- Vytvoriť vyrovnaný dataset
- Homolumo deskriptor - zistiť ako sa počíta a spraviť kód pre výpočet deskriptora
- Webstranka pre tím - vytvoriť nejaký nápad / návrh

- Integrácia cdkitu a padelu do kódu, ktorý generuje descriptory
- Pridať fotky členov tímu na stránku
- Skúsiť zistiť čo znamenajú Padel descriptory
- Začať trénovať model

Poznámky k šprintu

Výzvy

- TimeoutError v PaDELi pri “Isometric” SMILES kódoch
- Význam PaDEL deskriptorov nie je úplne jasný
- Náročné vyrovnávanie datasetu - pozitívnych záznamov je dostupných výrozne viac, ako negatívnych - náš model stále nie je vyvážený
- Chýbalo nám asi 400 negenotoxických látok pre vyrovnaný dataset

Poznatky

- Namiesto vytvorenia funkcie pre výpočet HomoLumo deskriptora je možné nájsť iný parameter, ktorý súvisí so vzdialenosťami medzi elektrónmi
- Zistili sme, aký je rozdieľ medzi CanonicalSMILES a IsometricSMILES
- Pri zmene SMILES kódov z “Isometric” na “Canonical” PaDEL funguje bez TimeoutError-u
- Našli sme nové biogénne prvky do negatívneho datasetu - postupne sa blížime k dostatočnému množstvu vzoriek pre vyvážený dataset
- Korelácie atd nie je vhodné robiť funkciami, ktoré programovali chemici
- Optická otáčavosť štruktúry chemickej látky má vplyv na toxicitu

Zhrnutie

Šprint číslo 3, ktorý prebiehal od 8. novembra 2023 do 21. novembra 2023, bol zameraný na dokončenie in vivo datasetu a pokračovanie v ďalšom vývoji projektu. Hlavným cieľom bolo dosiahnuť pokrok v pridaní nových prvkov do datasetu a začatie trénovania modelu. Medzi nové úlohy patrilo napríklad zistenie spôsobu výpočtu a implementácia deskriptora Homolumo, integrácia nástrojov cdkit a PaDEL do kódu na generovanie deskriptorov, pridanie biogénnych prvkov ako sacharóza a erythritol, a skúmanie, či sú sladidlá genotoxické a ich prípadné pridanie do datasetu.

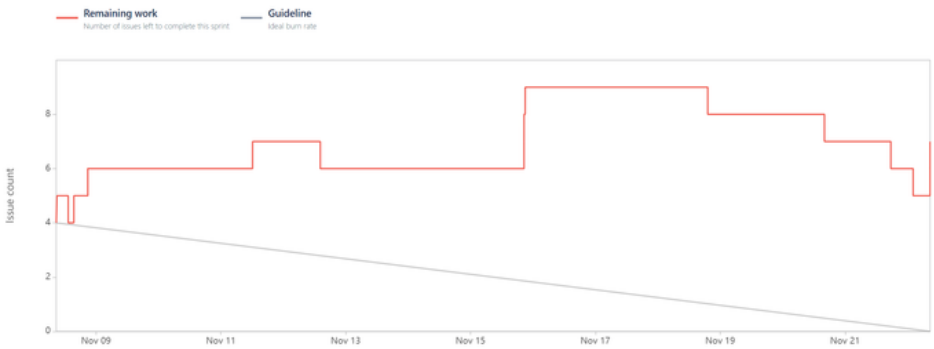
V priebehu šprintu sme úspešne dokončili 6 úloh, vrátane vytvorenia webovej stránky pre tím, pridania biogénnych prvkov a skúmania rôznosti medzi isomeric a canonical SMILES. Rozpracovali sme aj ďalších 7 úloh, medzi ktorými bolo vytvorenie návrhu webovej stránky pre tím, začatie trénovania modelu a skúmanie významu PaDEL deskriptorov.

S konkrétnymi výzvami sme sa stretli v podobe TimeoutError v PaDEL pri "Isometric" SMILES kódoch a nejasného významu PaDEL deskriptorov. Vyrovnanie datasetu bolo náročné, pretože pozitívnych záznamov bolo výrazne viac ako negatívnych. Napriek tomu sme získali poznatky o

výpočte Homolumo deskriptora a zistili sme rozdiely medzi CanonicalSMILES a IsometricSMILES, čo nám umožnilo vyhnúť sa TimeoutError-u v PaDEL.

S postupným pridávaním nových biogénnych prvkov do negatívneho datasetu sme sa približovali k vytvoreniu vyváženého datasetu s dostatočným množstvom vzoriek.

Burndown graf



Šprint číslo 4:

Dátum - November 22nd, 2023 - December 5th, 2023

Hlavný cieľ šprintu - pridať prvky pre Amesov test a práca na HomoLumo deskriptore

Výsledky šprintu

V tomto šprinte sme mali nasledovné ciele a výsledky:

Nové úlohy (12)

- Pozrieť ako sa v minulých rokoch robili zápisy zo šprintov
- Nájsť pozitívne látky pre In Vivo dataset
- Dať preč deskriptory, ktoré sú konštanty / rovnaké pre každý záznam / nedávajú zmysel
- Pridať negatívne látky k Amesovmu datasetu
- Pridať všetky doterajšie zápisy zo stretnutí na stránku
- Pridať na stránku krátky opis k zápisniciam zo stretnutí
- Spísať reporty zo šprintov
- Pridať reporty zo šprintov na tímovú stránku

- Skontrolovat CAS čísla / pridať chýbajúce
- Prihláška na TP cup a ITsrc
- Zmeniť Isomeric smiles kódy za Canonical smiles kódy
- Zistiť postup na deploy stránky na školský server

Úlohy z predchádzajúceho šprintu (7)

- Vytvoriť vyrovnaný dataset
- Homolumo deskriptor - zistiť ako sa počíta a spraviť kód pre výpočet deskriptora
- Webstranka pre tím - vytvoriť nejaký nápad / návrh
- Integrácia cdkitu a padelu do kódu, ktorý generuje descriptory
- Pridať fotky členov tímu na stránku
- Skúsiť zistiť čo znamenajú Padel descriptory
- Začať trénovať model

Dokončené úlohy (15)

- Nájsť pozitívne látky pre In Vivo dataset
    - Pridanych 138 pozitívnych latok, dorovnané negatívnymi.
- Vytvoriť vyrovnaný dataset
- Webstranka pre tím - vytvoriť nejaký nápad / návrh
- Integrácia cdkitu a padelu do kódu, ktorý generuje descriptory
- Pridať fotky členov tímu na stránku

- Skúsiť zistiť čo znamenajú Padel descriptory
- Začať trénovať model
- Pozrieť ako sa v minulých rokoch robili zápisy zo šprintov
- Nájsť pozitívne látky pre In Vivo dataset
- Dať preč deskriptory, ktoré sú konštanty / rovnaké pre každý záznam / nedávajú zmysel
- Pridať negatívne látky k Amesovmu datasetu
- Pridať všetky doterajšie zápisy zo stretnutí na stránku
- Spísať reporty zo šprintov

- Skontrolovať CAS čísla / pridať chýbajúce
- Prihláška na TP cup a ITsrc
- Zmeniť Isomeric smiles kódy za Canonical smiles kódy

Rozrobené úlohy (4)

- Zistiť postup na deploy stránky na školský server
- Pridať reporty zo šprintov na tímovú stránku
- Pridať na stránku krátky opis k zápisniciam zo stretnutí
- Homolumo deskriptor - zistiť ako sa počíta a spraviť kód pre výpočet deskriptora

Poznámky k šprintu

Výzvy

- HomoLumo knižnica simulovala kvantové počítanie a pri pokuse o spustenie nezbehol ani prvý prvok - nedostatok výpočtovej sily.
- Hľadať nové negatívne dáta už je ťažké
- Nie všetky deskriptory majú výpovednú hodnotu

Poznatky

- HomoLumo - podla výsledkov testov zatiaľ vyzerá, že bude dôlezite ho mať (všetky najvýznamnejšie deskriptory nám hovoria o elektrónoch)
- Smiles kódy je možné získavať z PUGchemu podľa CAS čísiel
- Zistili sme, ako sa integruje Padel
- Vo finálnej verzií bude dôležitá “vysvetliteľnosť - explainability”

Zhrnutie

Šprint číslo 4, ktorý prebehol od 22. novembra 2023 do 5. decembra 2023, sa venoval cieľu rozšíriť náš in vivo dataset pridaním prvkov pre Amesov test a práci na Homolumo deskriptore.

V rámci tohto šprintu sme si stanovili 12 nových úloh, ktoré zahŕňali preskúmanie minuloročných zápisov zo šprintov na získanie perspektívy, hľadanie pozitívnych látok pre In Vivo dataset, elimináciu nezmyselných deskriptorov a doplnenie negatívnych látok do

Amesovho datasetu. Ďalej sme pracovali na integrácii nástrojov CDKit a PaDEL do nášho kódu na generovanie deskriptorov a snažili sme sa lepšie pochopiť význam PaDEL deskriptorov.

V rámci šprintu sme úspešne dokončili 7 úloh, vrátane nájdenia pozitívnych látok pre In Vivo dataset, vytvorenia vyrovnaného datasetu a ďalších aktivít smerujúcich k rozvoju tímovej webovej stránky a získavaniu vysvetlení pre deskriptory. Okrem toho sme rozrobili 4 úlohy, medzi ktoré patrilo zisťovanie postupu na nasadenie stránky na školský server a integrácia Homolumo deskriptora do projektu.

Napriek dosiahnutým úspechom sme sa stretli s niekoľkými výzvami, vrátane nedostatku výpočtovej sily pre simuláciu Homolumo knižnice a náročnosti hľadania nových negatívnych dát. Taktiež sme si uvedomili, že nie všetky deskriptory majú výpovednú hodnotu.

Získali sme významné poznatky o Homolumo deskriptore a jeho význame v kontexte elektrónovej štruktúry. Tiež sme sa naučili, ako integrovať nástroje PaDEL do projektu a zdôraznili dôležitosť vysvetliteľnosti v konečnej verzii projektu. Celkovo nám šprint číslo 4 poskytol pevný základ pre ďalší rozvoj projektu, a to nielen v oblasti rozširovania datasetu, ale aj v zdokonaľovaní nástrojov a procesov v našom tímovom prostredí.

Burndown graf