

Uczenie maszynowe w projektowaniu leków – tematy projektów zaliczeniowych

**Dr hab. Sabina Podlewska¹
Dr Tomasz Danel^{2,3}**

¹**Zakład Chemii Leków, Instytut Farmakologii im. Jerzego Maja
Polskiej Akademii Nauk**

²**Wydział Chemiczny, Uniwersytet Jagielloński**

³**Wydział Matematyki i Informatyki, Uniwersytet Jagielloński**

04.03.2024

Informacje podstawowe

- Realizacja projektów w grupach (3-4 osoby)*

**Praca w mniejszych grupach lub indywidualna możliwa w przypadku udziału w dłuższych projektach badawczych realizowanych w GMUM i kończących się publikacją lub pracą magisterską.*

- Projekt oceniany będzie na podstawie 3 kamieni milowych:
 - *raport wstępny*
 - *prezentacja śródsemestralna*
 - *prezentacja semestralna*
- Ocena wyliczana na podstawie standardowej punktacji:

[0, 50]	2
(50, 60]	3
(60, 70]	3,5
(70, 80]	4
(80, 90]	4,5
(90, 100]	5
- Szczegółowa punktacja podana na dalszych slajdach

Kod projektu

- Efektem projektu powinno być powstanie repozytorium kodu, który umożliwia odtworzenie wyników projektu. Poniższa lista kontrolna wskazuje na czynniki brane pod uwagę przy ocenie kodu.
- Repozytorium jest udostępnione prowadzącym (mokosaur) i posiada licencję umożliwiającą podzielenie się wynikami projektu w przyszłości. [10 ]:

Przykład: Publiczne repozytorium GitHub z licencją MIT.

Przykład: Prywatne repozytorium GitLab zawierające plik LICENSE ze zgodą na upublicznienie kodu i wyników w celach edukacyjnych.

- Repozytorium zawiera dokumentację projektu w postaci pliku README, który zawiera między innymi: krótki opis projektu, instrukcję uruchomienia kodu wraz z listą zależności, podsumowanie wyników. Możliwe są również odniesienia do wygenerowanych plików PDF i iPython notebooków z uzupełnieniem dokumentacji, pod warunkiem że znajdują się również w repozytorium. [10 ]
- Repozytorium zawiera wszystkie dane lub odniesienia do źródeł danych [5 ]
- Kod umożliwia odtworzenie kluczowych wyników projektu [5 ]

Uwaga: przedmiot nie jest kursem programowania, dlatego styl i jakość kodu nie będą dokładnie sprawdzane. Niemniej jednak taki projekt może być okazją na rozbudowanie swojego CV, więc polecamy estetyczne formatowanie kodu (np. przy użyciu narzędzi black oraz isort załączonych w środowisku) oraz dodawanie docstringów i testów do implementowanych klas i funkcji

Raport wstępny [10]

- objętość: max. 1 strona A4
- termin: koniec marca (31.03.2024)
 - Repozytorium kodu zostało założone i udostępnione prowadzącym (*link w raporcie*)
 - Wszystkie dane i narzędzia potrzebne do realizacji projektu zostały znalezione (*wstępna lista źródeł danych, narzędzi, dodatkowych paczek Pythona*)
 - Zostały zidentyfikowane wszelkie problemy związane z realizacją projektu (*pytania załączone w raporcie*)

Prezentacja śródsemestralna

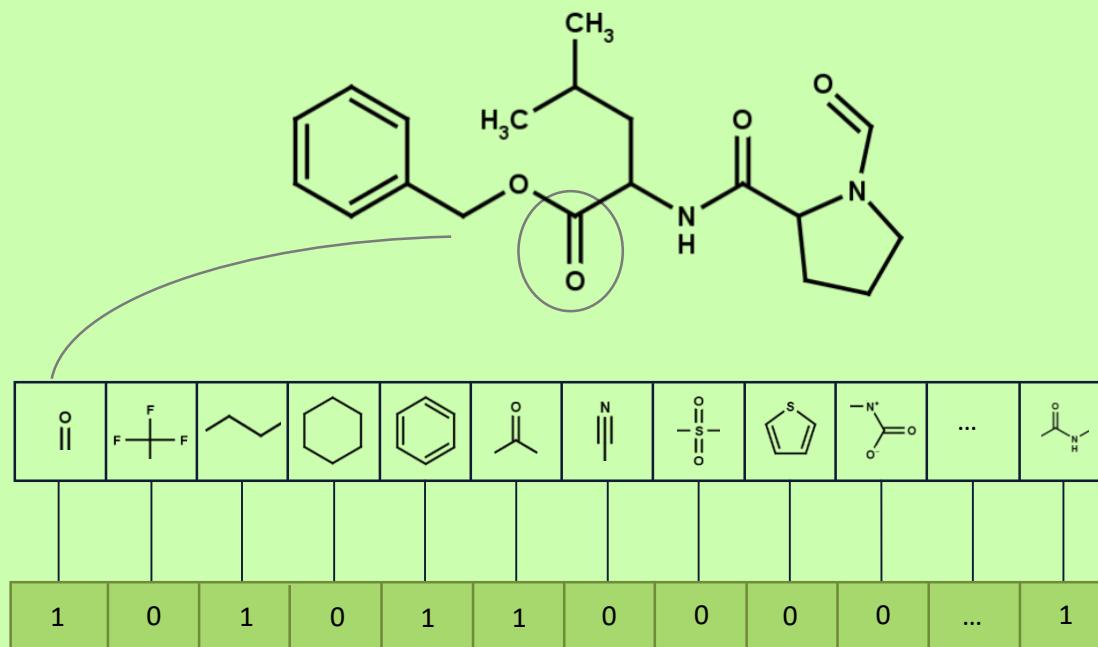
- Czas trwania: 15 minut
- Termin: 06.05.2024
 - Zrozumienie tematu [5 ]
(wytlumaczyć i zaciekawić pozostałych uczestników kursu tematyką projektu)
 - Co jest celem projektu?
 - Jakie jest znaczenie biologiczne/chemiczne projektu?
 - Co jest potencjalną trudnością w wykonaniu projektu?
 - Jaki jest spodziewany efekt projektu?
 - Zrozumienie danych [5 ]
(mile widziane wyniki eksploracyjnej analizy danych: wizualizacja danych i wnioski)
 - Jakie dane wejściowe będą użyte w projekcie?
 - Skąd można pozyskać więcej danych do projektu?
 - Jakie problemy w dostępnych danych są widoczne?
 - Jakie dodatkowe informacje (metadane) są dostępne?
 - Planowana implementacja [5 ]
(mile widziane schematy architektury sieci lub algorytmu, wyniki prostych modeli odniesienia [baselines] lub najnowocześniejszych modeli [SOTA] z literatury)
 - Czy ten temat był już poruszany w literaturze? Jeśli tak, to jakie narzędzia są dostępne?
 - Jakie metody uczenia maszynowego będą wykorzystane w projekcie?
 - Jak zdefiniowane będzie wejście i wyjście modelu?
 - Jakie miary będą zastosowane do zmierzenia skuteczności modelu?
 - Jaki stos technologiczny będzie użyty do wykonania projektu?

Prezentacja semestralna

- Klarowna prezentacja tematyki projektu [5 ]
- Odpowiedź na każdą z zadanych hipotez badawczych [3x10 ] (~3 hipotezy na projekt)
- Własne hipotezy i dodatkowe wyniki [10 ]

1. Generowanie związków w oparciu o optymalny zestaw podstruktur

Fingerprint podstrukturalny



Generowanie związków w oparciu o optymalny zestaw podstruktur

Fingerprint podstrukturalny:

- Jeden z najpopularniejszych sposobów reprezentacji struktur chemicznych
- Informuje o obecności lub braku określonych podstruktur w organizmie
- Dla związków wykazujących aktywność wobec danego typu receptora często jest charakterystyczne występowanie określonych fragmentów strukturalnych (przy jednoczesnej nieobecności innych).
- Analiza częstości występowania określonych podstruktur pozwala na identyfikację tzw. fragmentów uprzywilejowanych, które zwiększą prawdopodobieństwo posiadania przez związek pożądanego profilu aktywności

Generowanie związków w oparciu o optymalny zestaw podstruktur

- Celem projektu jest identyfikacja optymalnych fragmentów strukturalnych dla danego profilu aktywności oraz wygenerowanie propozycji nowych ligandów na podstawie tych „optymalnych” fragmentów
- Podobne podejście zastosowano w pracy **Podlewska, S.; Czarnecki, W.; Kafel, R.; Bojarski, A.J.** Creating the New from the Old: Combinatorial Libraries Generation with Machine-Learning-Based Compound Structure Optimization. *J. Chem. Inf. Model.* **2017**, 57, 133-147
- Celem bieżącego projektu jest zastosowanie innej metody selekcji optymalnych podstruktur, ewentualne łączenie różnych reprezentacji podstrukturalnych ze sobą, a także automatyczne generowanie związków z powstałych fragmentów. Ponadto, można rozważyć wykorzystanie metod regresyjnych, tj. w cytowanej publikacji wykorzystano jedynie informację o tym czy związek jest aktywny czy nieaktywny, natomiast w tym przypadku można wykorzystać konkretne wartości parametrów aktywnościowych

Generowanie związków w oparciu o optymalny zestaw podstruktur

- Konieczne wprowadzenie ograniczenia na masę molową powstających związków, ew. razu optymalizacja np. pod kątem reguł Lipińskiego (lub innych)
- Przykładowe dane:

Wartość parametru aktywności				...	
20	1	0	0	...	1
1000	0	1	1	...	1
5	0	1	1	...	1
...	0	0	0	...	0
540	1	1	0	...	0

Generowanie związków w oparciu o optymalny zestaw podstruktur

Plan działania:

- Identyfikacja pożądanych podstruktur
- Wygenerowanie fragmentów z istniejących ligandów
- Łączenie „optymalnych” fragmentów w nowe cząsteczki

Wymagania:

- Zapoznanie się z konceptem fingerprintu podstrukturalnego oraz możliwościami związanymi z jego wykorzystaniem w procesie projektowania nowych leków

Generowanie związków w oparciu o optymalny zestaw podstruktur

Hipotezy badawcze

- Weryfikacja czy zestawy optymalnych fragmentów są podobne dla różnych celów biologicznych
- Sprawdzenie podobieństwa zestawów wygenerowanych związków dla różnych celów biologicznych
- Weryfikacja czy związki powstałe z fragmentów istniejących ligandów mają własności drug-like (np. czy spełniają reguły Lipińskiego)

Generowanie związków w oparciu o optymalny zestaw podstruktur

Kamienie milowe

- **Raport wstępny**
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- **Prezentacja śródsemestralna**
 - Opracowanie metody selekcji optymalnych fragmentów na podstawie fingerprintów
 - Przygotowanie charakterystyki zbiorów fragmentów wygenerowanych dla poszczególnych targetów (np. liczność, rozkład masy molowej)
 - Prezentacja przykładów związków, które mogą powstać z wygenerowanych fragmentów, podanie propozycji rozwiązania problemu z ograniczeniem masy molowej nowo formowanych struktur (np. tworzenie podzbiorów z powstały zestawów fragmentów)
 - przedstawienie propozycji algorytmu łączenia powstałych fragmentów w propozycje nowych cząsteczek chemicznych

Generowanie związków w oparciu o optymalny zestaw podstruktur

Kamienie milowe

■ Prezentacja semestralna

- wygenerowanie bibliotek nowych propozycji ligandów na podstawie optymalnych fragmentów istniejących ligandów
- charakterystyka powstałych związków (np. pod kątem ich własności drug-like)
- analiza - czy optymalne fragmenty wygenerowane dla danej reprezentacji (rodzaju fingerprintu) pokrywają się w przypadku stworzenia reprezentacji hybrydowej powstałej ze sklejenia innych reprezentacji fingerprintowych.

2. Przewidywanie własności ADMET metodami structure-based

- Przewidywanie własności fizykochemicznych i farmakokinetycznych/ADMET są równie istotne jak zapewnienie odpowiedniego powinowactwa do określonego celu biologicznego (związek nie może być lekiem jeśli nie jest w stanie dotrzeć do dedykowanego)
- Niektóre z tych cech ADMET są bezpośrednio związane z oddziaływaniem z określonym białkiem - np. kardiotoksyczność jest bezpośrednio powiązana z blokowaniem kanałów potasowych hERG, stabilność metaboliczna - rozkład związku jest związany przede wszystkim z oddziaływaniem z różnymi podtypami cytochromu P450, które są odpowiedzialne za metabolizm wszystkich substancji obcych dostających się do organizmu, z kolei biodostępność leku może być modulowana przez wiązanie leku z białkami osocza, głównie z albuminami

Przewidywanie własności ADMET metodami structure-based

- Celem projektu jest konstrukcja modeli predykcyjnych do przewidywania wybranych własności ADMET związków na podstawie przewidywanego sposobu oddziaływania tychże związków z wybranymi białkami.
- **Schemat postępowania:**
 - zadokowanie związków o znanym powinowactwie do rozpatrywanych celów biologicznych do odpowiednich białek (istniejące narzędzia)
 - zakodowanie uzyskanych kompleksów ligand-receptor do postaci zero-jedynkowej za pomocą tzw. fingerprintu oddziaływań strukturalnych (istniejące narzędzia)
 - stworzenie modelu predykcyjnego, który będzie przewidywać siłę potencjalnego oddziaływania rozpatrywanego związku z danym białkiem na podstawie kontaktów ligand-białko zadokowanych w postaci zero-jedynkowej

Przewidywanie własności ADMET metodami structure-based

Hipotezy badawcze:

- weryfikacja "trudności" zadania predykcyjnego w zależności od różnych parametrów, np. rodzaju białka, liczności zestawu uczącego, dystrybucji aktywności w obrębie zestawu uczącego
- zbadanie wpływu zastosowanego modelu/kryształu

Przewidywanie własności ADMET metodami structure-based

Kamienie milowe

- **Raport wstępny**
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- **Prezentacja śródsemestralna**
 - podpięcie dockera i zadokowanie związków
 - wygenerowanie reprezentacji zero-jedynkowej dla uzyskanych kompleksów ligand-receptor
- **Prezentacja semestralna**
 - przygotowanie modeli predykcyjnych różnego typu przewidujących aktywność wobec rozpatrywanych celów biologicznych na podstawie oddziaływań w kompleksie ligand-receptor
 - porównanie skuteczności różnych metod
 - detekcja oddziaływań istotnych dla danego typu aktywności

3. Generowanie nowych ligandów w podejściu structure-based

- Mamy zestaw ligandów, zadokowanych z docking-scorem i generujemy nowe ligandy -> dwie ścieżki optymalizacyjne:
- Generujemy ligandy optymalizując ich docking score
- Generujemy ligandy optymalizując parametr aktywności
- Porównujemy uzyskane zestawy ligandów
- Stosujemy istniejące metody generowania (CVAE, GVAE, REINVENT)

Generowanie nowych ligandów w podejściu structure-based

Hipotezy badawcze:

- Weryfikacja poprawności chemicznej związków generowanych przez poszczególne metody
- Analiza związków wygenerowanych różnymi metodami pod kątem własności drug-like
- Analiza/porównanie przestrzeni chemicznej związków wygenerowanych dla różnych celów biologicznych

Generowanie nowych ligandów w podejściu structure-based

Kamienie milowe

- **Raport wstępny**
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- **Prezentacja śródsemestralna**
 - podpięcie dockera i zadokowanie związków
 - wstępne użycie wybranych metod do generowania nowych związków
- **Prezentacja semestralna**
 - wygenerowanie zestawów związków dla wszystkich rozpatrywanych celów biologicznych
 - przeprowadzenie analizy statystycznej własności związków (masa molowa oraz inne cechy wchodzące w skład tzw. reguł Lipińskiego pozwalające na wstępne określenie czy związek może w przyszłości stać się lekiem) biorąc pod uwagę różne metody i targety

4. Analiza wyników dynamiki molekularnej

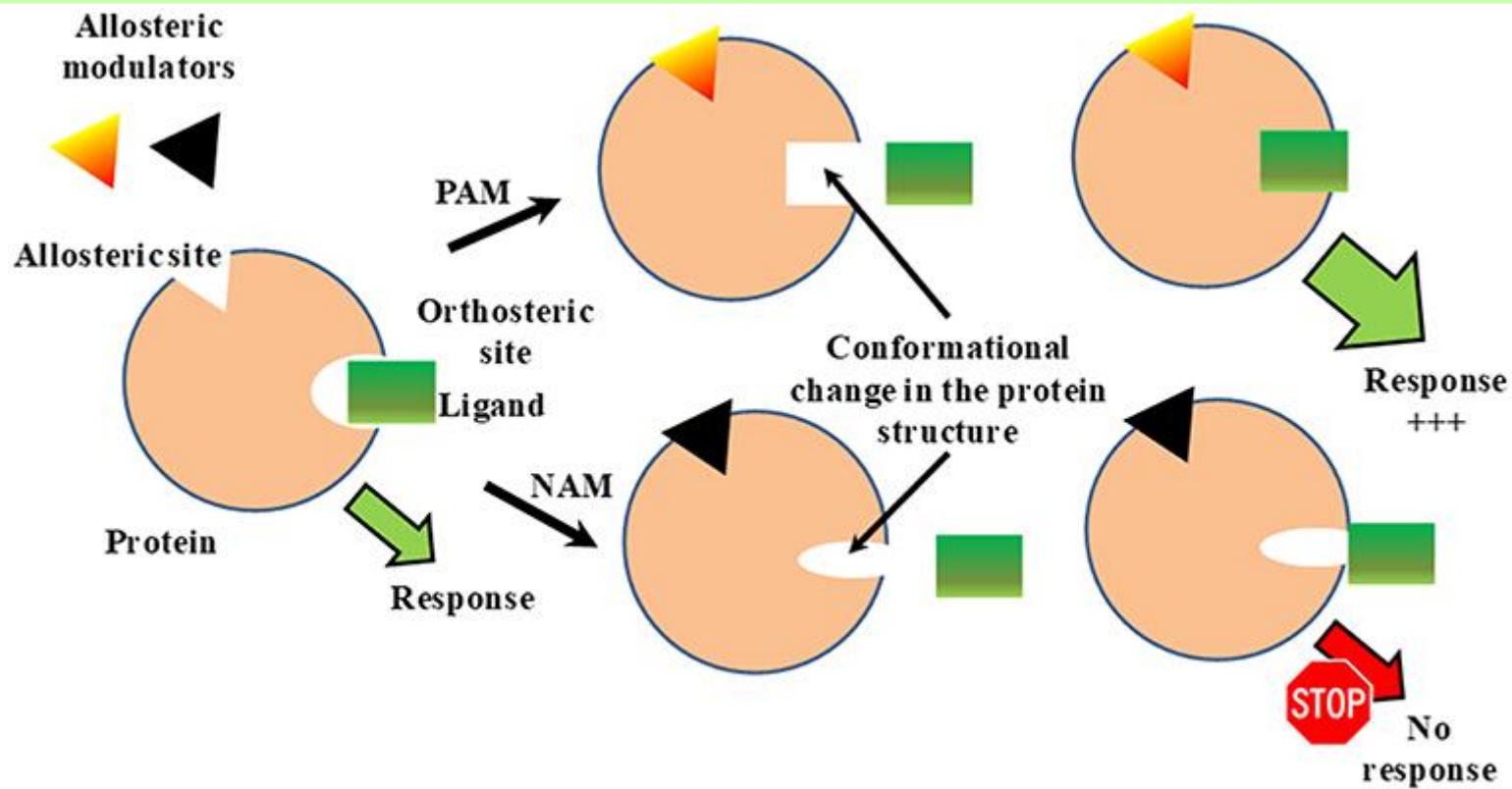
- Umożliwia badanie wybranego układu w funkcji czasu (na poziomie atomowym), którego zachowanie jest opisywane przez odpowiednie równania ruchu (np. równania Newtona)
- Podstawowym problemem w symulacjach dynamiki molekularnej jest wyznaczenie sił działających na każdy z atomów układu. Wykorzystuje się do tego celu funkcję, która reprezentuje energię potencjalną danego atomu:

$$\vec{F}_i = -\nabla_{r_j} V(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_n)$$

- Opisany wyżej potencjał jest zdefiniowany przez pole siłowe, które uwzględnia różne przyczynki energii potencjalnej

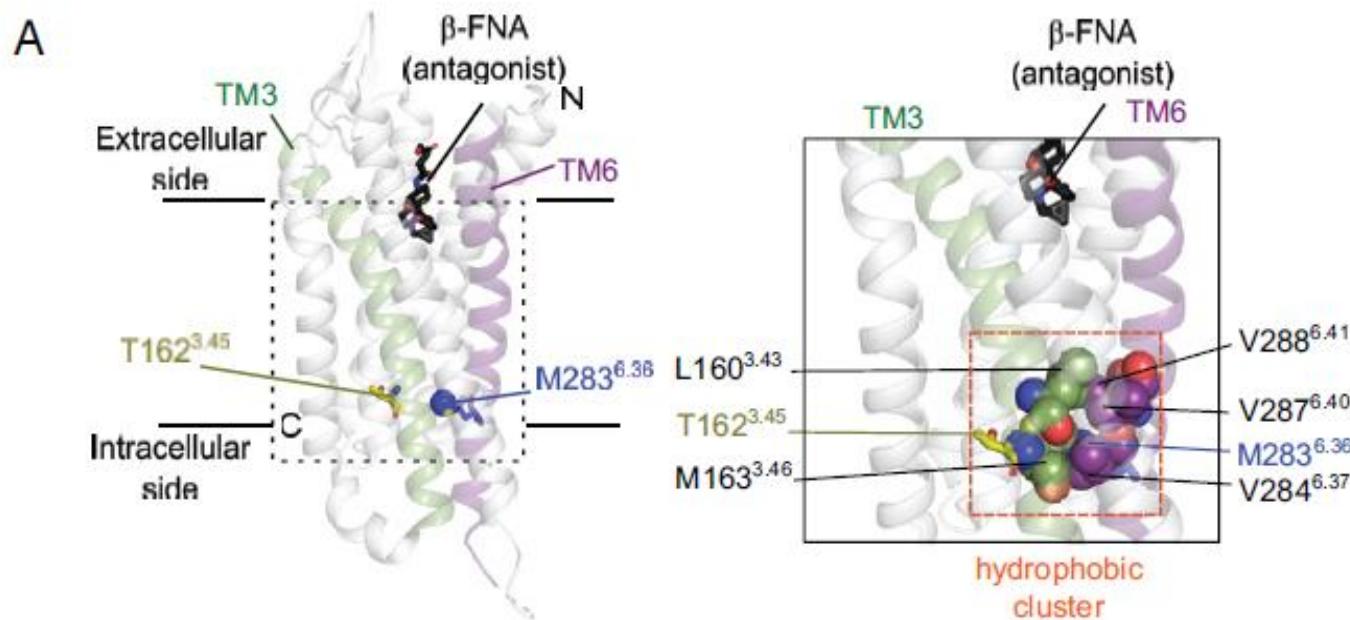
Analiza wyników dynamiki molekularnej

Problem: modulacja allosteryczna receptora mu opioidowego



Analiza wyników dynamiki molekularnej

- Cel projektu: identyfikacja determinantów molekularnych koniecznych do wykazywania przez związek modulacji allosterycznej
- Dane:
 - Wyniki symulacji dynamiki molekularnej dla serii ligandów badanych pod kątem wykazywania zdolności do modulacji allosterycznej
 - Dwa potencjalne miejsca allosteryczne

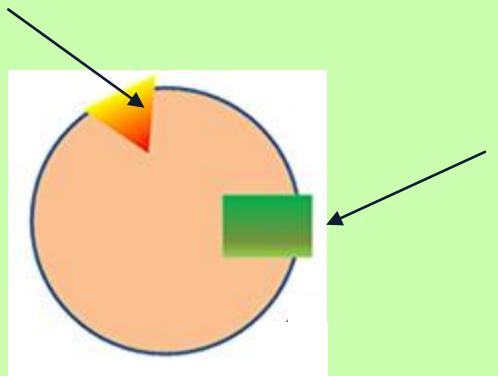


Analiza wyników dynamiki molekularnej

- Badane związki:

Miejsce allosteryczne:

- SR-11501
- Sufentanil
- Fentanyl
- Olicerydyna
- SR-15098
- SR-15099
- SR-17018
- SR-14698



Miejsce ortosteryczne:

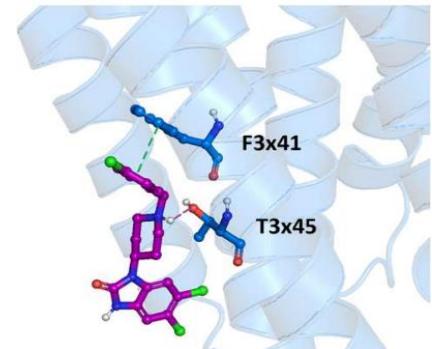
- nalokson

Analiza wyników dynamiki molekularnej

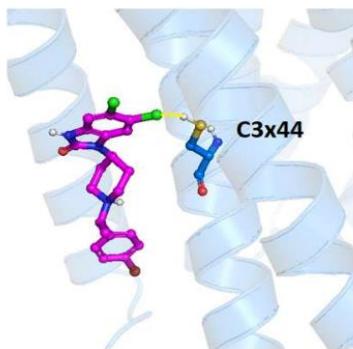
Format danych

1. Kompleksy ligand-receptor dla poszczególnych chwil czasu
2. Lista oddziaływań ligand-receptor w poszczególnych chwilach czasowych

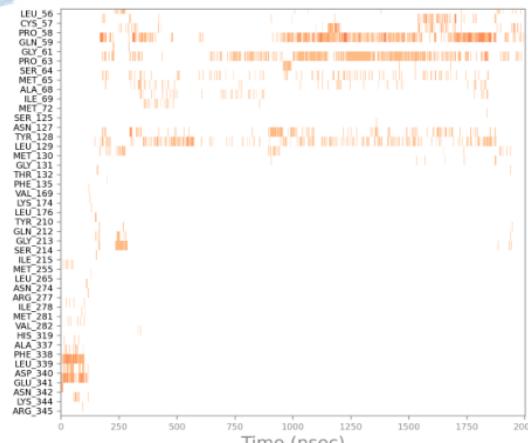
SR-17018



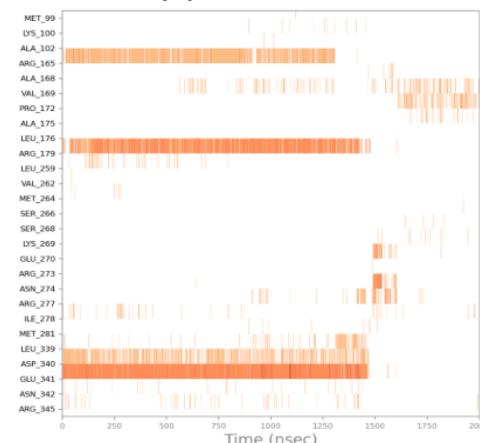
(R)-SR-14968



SR17018



(S)-SR14698



Analiza wyników dynamiki molekularnej

Hipotezy badawcze:

- Zbadanie korelacji pomiędzy wynikami symulacji dynamiki molekularnej a wynikami badań *in vitro* dla dwóch proponowanych miejsc allosterycznych; wskazanie bardziej prawdopodobnego miejsca wiązania modulatora
- Zbadanie wpływu/korelacji oddziaływań z poszczególnymi aminokwasami z wynikami badań *in vitro* (zbadanie zachowania liganda ortosterycznego, naloksonu, oraz modulatora allosterycznego)
- Zbadanie korelacji odległości pomiędzy wybranymi atomami/kątami torsyjnymi z wynikami badań *in vitro*

Analiza wyników dynamiki molekularnej

Kamienie milowe

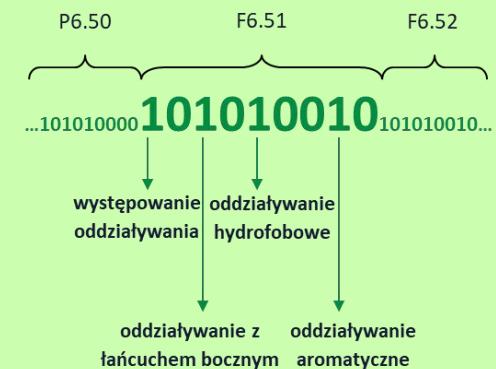
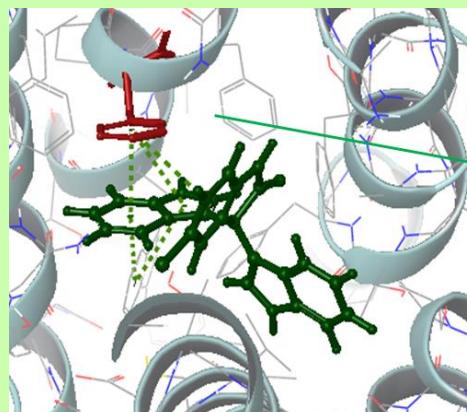
- **Raport wstępny**
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- **Prezentacja śródsemestralna**
 - pre-processing danych, wstępna analiza statystyczna oddziaływań
 - zdefiniowanie badanych odległości atomowych i kątów
 - przygotowanie propozycji własności ligandów do zbadania podczas przebiegu symulacji dynamiki molekularnej
- **Prezentacja semestralna**
 - przygotowanie analiz korelacyjnych dla oddziaływań ligandów z aminokwasami vs. wyniki badań *in vitro*
 - zbadanie analiz korelacyjnych dla wybranych odległości atomowych i kątów vs. wyniki badań *in vitro*
 - zbadanie analiz korelacyjnych dla wybranych własności ligandów vs. wyniki badań *in vitro*

5. Explainable AI - metody wyjaśnialności w analizie kompleksów ligand-receptor

- Analiza kompleksów ligand-receptor reprezentowanych w postaci zero-jedynkowej (fingerprinty oddziaływań strukturalnych) przez metody uczenia maszynowego - korelacja konkretnych zestawów oddziaływań z aktywnością biologiczną
- Wykorzystanie metod wyjaśnialności do analizy konkretnych predykcji
- Porównanie wyników uzyskiwanych przez metody wyjaśnialności (najczęściej wskazywane aminokwasy jako te, które są istotne dla danego profilu aktywności) z prostą analizą statystyczną oddziaływań dla grup związków aktywnych i nieaktywnych

Explainable AI - metody wyjaśnialności w analizie kompleksów ligand-receptor

■ Fingerprints oddziaływań strukturalnych



Explainable AI - metody wyjaśnialności w analizie kompleksów ligand-receptor

Hipotezy badawcze:

- Porównanie aminokwasów wskazywanych jako kluczowe dla różnych metod wyjaśnialności
- Porównanie aminokwasów wskazywanych jako kluczowe przez metody wyjaśnialności z prostą analizą statystyczną
- Porównanie aminokwasów wskazywanych jako kluczowe na odpowiadających sobie pozycjach w sekwencji spokrewnionych receptorów

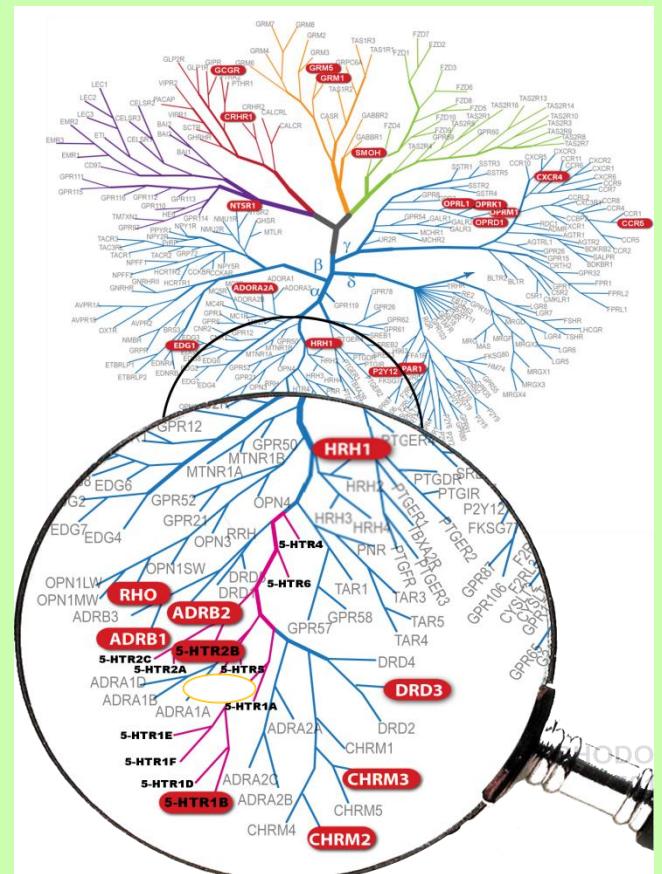
Explainable AI - metody wyjaśnialności w analizie kompleksów ligand-receptor

Kamienie milowe

- **Raport wstępny**
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- **Prezentacja śródsemestralna**
 - pre-processing danych, wstępna analiza statystyczna oddziaływań
 - zdefiniowanie metod uczenia maszynowego, które będą używane w badaniach
- **Prezentacja semestralna**
 - przygotowanie analiz korelacyjnych dla oddziaływań ligandów z aminokwasami względem ich aktywności biologicznej
 - identyfikacja aminokwasów kluczowych dla danej aktywności biologicznej (ML + prosta statystyka)
 - Zbadanie stopnia pokrywania się istotnych aminokwasów dla odpowiadających sobie pozycji w sekwencji dla spokrewnionych receptorów.

6. „Drug repurposing” w obrębie aminergicznych receptorów typu GPCR

- Celem projektu jest zbadanie przydatności ligandów jednego receptora do modulacji aktywności innego, pokrewnego białka, bez potrzeby wykonywania procedury dokowania, na podstawie badania kompatybilności sekwencji i struktury chemicznej związku.



„Drug repurposing” w obrębie aminergicznych receptorów typu GPCR

Hipotezy badawcze:

- Porównanie wskazywanych jako różnych receptorów aminokwasów kluczowe dla
- Mapowanie określonych chemicznych na konkretne aminokwasy
- Sprawdzanie przydatności konkretnych ugrupowań chemicznych również w receptorach pokrewnych; ocena przydatności całego liganda

„Drug repurposing” w obrębie aminergicznych receptorów typu GPCR

Kamienie milowe

- **Raport wstępny**
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- **Prezentacja śródsemestralna**
 - Podpięcie dockera, przeprowadzenie procedury dokowania
 - Reprezentacja kompleksów ligand-receptor w postaci fingerprintów oddziaływań
- **Prezentacja semestralna**
 - Analiza kompleksów ligand-receptor pod kątem występowania określonych wzorców aktywności (ugrupowanie chemiczne-aminokwas)
 - Identyfikacja aminokwasów kluczowych dla danej aktywności biologicznej
 - Przypisanie badanym ligandom listy receptorów wobec których mogą wykazywać potencjalną aktywność biologiczną

7. Generowanie sztucznych zbiorów związków

- Dostępność syntetyczna, czyli łatwość syntetyzowania określonego związku, jest jedną z najważniejszych cech związku, na jakie trzeba zwrócić uwagę na etapie projektowania leku. Nie każda kombinacja atomów jest możliwa do utworzenia w laboratorium. W związku z tym ważne jest śledzenie syntetyzowalności na etapie projektowania. Czasem do projektowania nowych pochodnych używa się jedynie znanych reakcji, by móc kontrolować dostępność syntetyczną na każdym kroku.
- Projekt polega na zaimplementowaniu narzędzia (najlepiej z prostym interfejsem graficznym), które potrafić będzie generować sztuczne zbiory danych od zera lub modyfikować podane związki na podstawie zestawu dopuszczalnych zasad. Zasady te mogą być określone przy pomocy SMARTS-ów, które można porównać do wyrażeń regularnych na strukturach związku. W sieci można znaleźć gotowe zestawy SMARTS-ów dla istniejących reakcji. Do przetestowania użyteczności takich zbiorów można użyć dowolnego istniejącego modelu generatywnego wytrenowanego na tym syntetycznym zbiorze. Czy model jest w stanie generować podobne związki?

Generowanie sztucznych zbiorów związków

Wymagania

- Przyda się znajomość narzędzi do projektowania prostych interfejsów graficznych, np. **streamlit**. Projekt może wymagać zapoznania się z podstawami **chemii organicznej** (jak przebiegają reakcje) oraz uważnego przyjrzenia się funkcjom związanym z modyfikacjami związków w bibliotece **RDKit**.

Generowanie sztucznych zbiorów związków

Hipotezy badawcze

- Program jest w stanie wygenerować różnorodne pochodne związku
- Model generatywny jest w stanie wygenerować nowe związki (spoza zbioru uczącego), ale podobne strukturalnie do tych syntetycznych
- Generator związków jest parametryzowalny, tzn. można zmodyfikować sposób generowania zbioru

Generowanie sztucznych zbiorów związków

Kamienie milowe

- **Raport wstępny**
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- **Prezentacja śródsemestralna**
 - Zaproponowanie zestawu reguł, które posłużą do generowania nowych związków.
 - Wybranie modelu generatywnego, który będzie służył do przetestowania danych.
- **Prezentacja semestralna**
 - Implementacja algorytmu generującego sztuczne zbiory związków.
 - Wytrenowanie modelu generatywnego i obliczenie wybranego zestawu metryk.

8. Neuronowy 3D fingerprint

- *pretrenig · uczenie reprezentacji · sieci neuronowe*
- Fingerprinty strukturalne są wyliczane na podstawie informacji o podstrukturach, które zawierają się w cząsteczce. Takie fingerprinty rzadko kodują informację o ułożeniu przestrzennym atomów (co najwyżej pozwalają zapisywać w atomach konfigurację absolutną R lub S). W niektórych zadaniach informacja o kształcie (konformacji) cząsteczki może być niezwykle ważna.
- Z drugiej strony fingerprinty są coraz częściej wypierane przez sieci grafowe, które można porównać do trenowalnych fingerprintfów. Sieci te agregują informację o sąsiedztwie atomów i używają ją do przewidywania własności cząsteczek. Coraz większą popularnością cieszą się sieci grafowe dla grafów zorientowanych w przestrzeni (gdzie wierzchołki mają określone pozycje).
- Celem projektu jest zaproponowanie nowego fingerprintu trenowalnego, który będzie uwzględniał informację o konformacji związku. Dodatkowo, najlepiej byłoby taki fingerprint nauczyć w sposób nienadzorowany (samonadzorowany), czyli bez udziału etykiet. Taki fingerprint będzie można później użyć do dowolnego zadania, podobnie jak klasyczne fingerprinty ECFP.
- Uczenie samonadzorowane polega na uczeniu sieci na pomocniczym zadaniu skonstruowanym na bazie zbioru treningowego. Przykładem jest uczenie kontrastywne, gdzie podobne związki chemiczne łączy się w pary, a zadaniem sieci jest ujednolicenie ich reprezentacji (wyjście z sieci powinno być jak najbardziej podobne dla podobnych związków).

Neuronowy 3D fingerprint

Wymagania

- Znajomość pakietu PyTorch. Przydatna może się okazać znajomość technik **uczenia samonadzorowanego** dla danych obrazowych albo technik **pretreningu** w danych tekstowych.

Hipotezy badawcze

- Zaproponowany fingerprint zostanie porównany z innymi fingerprintami na kilku zadaniach predykcyjnych.
- Uczenie kontrastywne znajdujące podobieństwo cech 3D związku daje lepsze rezultaty niż bazowanie na reprezentacji 2D.
- Czy uczenie kontrastywne na poziomie atomów (sąsiedztwa atomów) daje lepsze wyniki niż uczenie na całych związkach?

Neuronowy 3D fingerprint

Kamienie milowe

- **Raport wstępny**
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- **Prezentacja śródsemestralna**
 - Zaproponowanie funkcji kosztu używanej do wytrenowania fingerprintu.
 - Przygotowanie zbioru danych (pre)treningowych.
 - Zaproponowanie zbiorów ewaluacyjnych (najlepiej takich, gdzie konformacje są istotne)
- **Prezentacja semestralna**
 - Implementacja i wytrenowanie modelu.
 - Wytrenowanie modeli bazowych na zwykłych fingerprintach i porównanie wyników.

9. Klastrowanie publicznych baz związków

- Publiczne bazy danych zawierają ogromną liczbę związków chemicznych. Niestety często jakość tych danych pozostawia wiele do życzenia. Często publikowane są jedynie pozytywne wyniki, pomijając cały ogrom stworzonych związków, które nie sprawdziły się w testach laboratoryjnych. Różnice między sprzętem stosowanym w różnych ośrodkach naukowych także wpływają na niespójność danych.
- Celem projektu jest znalezienie ciekawych zależności w dużych publicznych zbiorach danych. Odbywać się to może przy pomocy narzędzi klastrowania i zmniejszenia wymiarowości danych w celu wizualizacji. Przykładowo związki można zakodować w postaci fingerprintów Morgana i użyć algorytmu klastrowania k-means, aby podzielić dane na grupy. Następnie można użyć algorytmu t-SNE, aby zwizualizować dane na dwuwymiarowej przestrzeni. Jakie wnioski można wyciągnąć z tego eksperymentu? Co daje nam zmiana rodzaju fingerprintu lub metody klastrowania?

Klastrowanie publicznych baz związków

Wymagania

- W projekcie może pomóc znajomość narzędzi klastrowania, np. pakietu **scikit-learn**, oraz umiejętność **wizualizacji danych**.

Hipotezy badawcze

- Co można powiedzieć o przestrzeni chemicznej? Czy związki są rozłożone na niej równomiernie?
- Czy wewnątrz grup można dostrzec wyraźne elementy wspólne wszystkich związków? Czy grupy korelują ze źródłem danych?
- Czy klastrowanie uwzględniające wartości aktywności pokazuje niespójność pomiędzy różnymi testami biologicznymi (assays)?

Klastrowanie publicznych baz związków

Kamienie milowe

- **Raport wstępny**
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- **Prezentacja śródsemestralna**
 - Wybór reprezentacji związków chemicznych i algorytmów klastrowania.
 - Pobranie danych i przeprowadzenie wstępnej analizy danych.
- **Prezentacja semestralna**
 - Poklastrowanie danych i przygotowanie odpowiednich wizualizacji.
 - Interpretacja osiągniętych wyników klastrowania.

10. Platforma do oceny jakości związku

- W procesie projektowania leków związki chemiczne są sprawdzane pod kątem wielu własności. Przebadanie eksperymentalnie dużych zbiorów związków pod kątem każdej własności jest bardzo kosztowne, stąd wielką popularnością cieszą się platformy do przewidywania własności ADMET (absorpcja, dystrybucja, metabolizm, eliminacja, toksyczność).
- Celem projektu jest stworzenie prostej platformy dającej możliwość wgrania zbioru związków chemicznych (np. w formacie smiles) i otrzymanie listy predykcji modeli ML dla wybranych własności. Można też wykonać proste wizualizacje wejściowych zbiorów danych, np. rozkłady przewidywanych własności w raporcie zbiorczym.
- Przykład: <http://www.swissadme.ch/>

Platforma do oceny jakości związku

Wymagania

- Aplikację można stworzyć w dowolnym pakiecie, ale dla osób bez doświadczenia w tym zakresie polecam **streamlit**. Do zbudowania podstawowych modeli ML można użyć **scikit-learn** albo **PyTorch**. Przyda się znajomość narzędzi do wizualizacji danych, np. **matplotlib**, **plotly**, **seaborn** lub **bokeh**.

Hipotezy badawcze

- Interfejs graficzny platformy jest intuicyjny i daje możliwość zapisu wyników.
- Zastosowane modele ADMET osiągają wysoką skuteczność na zbiorze testowym.
- Raport zbiorczy daje wgląd w ciekawe własności całego zbioru związków.

Platforma do oceny jakości związku

Kamienie milowe

- **Raport wstępny**
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- **Prezentacja śródsemestralna**
 - Wybranie frameworka, który będzie użyty do stworzenia interfejsu graficznego.
 - Zaproponowanie własności związków, na które będą wytrenowane modele.
 - Zaproponowanie architektury modelu predykcyjnego.
- **Prezentacja semestralna**
 - Stworzenie platformy do oceny związków.
 - Podpięcie modeli predykcyjnych pod platformę.
 - Stworzenie widoku z raportem zbiorczym wejściowych danych.

11. Wyjaśnialność modeli QSAR

- Większość modeli uczenia maszynowego nie tłumaczy w zrozumiałym dla człowieka sposób procesu myślowego wiodącego do danej predykcji. W wielu przypadkach takie wytlumaczenie mogłoby pomóc podjąć kolejne kroki. Przykładem mogą być modele QSAR (quantitative structure-activity relationship), które na podstawie struktury chemicznej przewidują aktywność biologiczną związku. Jeśli zrozumiemy jakie ugrupowania atomów przyczyniają się pozytywnie bądź negatywnie do predykcji, będziemy w stanie lepiej projektować związki aktywne.
- W projekcie przeprowadzona powinna zostać analiza predykcji dowolnego modelu uczenia maszynowego dla problemu QSAR. Mogą to być zarówno modele klasyczne, np. random forest albo SVM, jak również sieci neuronowe. W przypadku sieci neuronowych mamy narzędzia bazujące na gradientach takie jak saliency maps, natomiast dla pozostałych modeli można zastosować metody typu LIME, które są odpowiednie nawet dla modeli czarnoskrzynkowych. Idealnym wynikiem projektu byłoby wskazanie elementów związków aktywnych, które przyczyniają się najbardziej do aktywności.

Wyjaśnialność modeli QSAR

Wymagania

- Przydatna może okazać się znajomość biblioteki **scikit-learn** lub **PyTorch**. Poszerzenie wiedzy dotyczącej **interakcji ligand-białko** w trakcie projektu będzie plusem.

Hipotezy badawcze

- Wyjaśnienia sieci pokrywają się z obecną wiedzą o aktywności tych związków, tj. wskazują na znane w literaturze interakcje lub cechy farmakoforowe.
- Użytkownicy zgadzają się z wyjaśnieniami modelu (możemy poprosić ekspertów o opinię).
- Wyjaśnienia sieci są istotnie różne od modeli bazowych (losowych), co można potwierdzić konkretnymi miarami (można znaleźć m.in. typu "rzadkość/sparsity" wyjaśnień).

Wyjaśnialność modeli QSAR

Kamienie milowe

- Raport wstępny
 - założenie repozytorium kodu
 - przygotowanie listy narzędzi/bibliotek, które będą używane do realizacji projektu
 - przygotowanie listy potencjalnych problemów, na które można się natknąć podczas realizacji projektu
- Prezentacja śródsemestralna
 - Przegląd literatury dotyczącej interpretowalnych modeli lub wyjaśnialności modeli.
 - Wybranie modelu do tworzenia wyjaśnień.
 - Znalezienie zbiorów aktywności biologicznej i odszukanie informacji na temat aktywnych związków (co jest istotne dla aktywności wobec danego białka).
- Prezentacja semestralna
 - Implementacja wyjaśnialności modeli.
 - Zaprojektowanie i przeprowadzenie eksperymentów uzasadniających poprawność modelu.

12. Duże modele językowe w przewidywaniu aktywności

Opis

- Duże modele językowe (ang. Large Language Models, LLMs) zyskały ostatnio ogromną popularność ze względu na pojawienie się narzędzi takich jak ChatGPT. Zaczęły pojawiać się zastosowania modeli językowych w chemii. Z jednej strony możliwe jest użycie podobnych architektur sieci neuronowych do przetwarzania reprezentacji takich jak SMILES albo sekwencja aminokwasów, z drugiej strony modele wytrenowane na tekstach w języku naturalnym posiadają do pewnego stopnia zrozumienie tematyki chemii.
- Projekt polegałby na sprawdzeniu, czy narzędzia takie jak ChatGPT są w stanie przewidywać aktywność związków chemicznych. Czy jesteśmy w stanie przy ich pomocy tłumaczyć aktywność związków? Do tego zadania moglibyśmy wziąć na przykład kilka zbiorów danych aktywnościowych i przetestować dokładność przewidywań modelu na zbiorze testowym. Czy zapytanie modelu o tok rozumowania dla jego predykcji da nam wyjaśnienie aktywności?

Wymagania

- Przydatna może okazać się znajomość biblioteki **PyTorch** w przypadku użycia modeli open source.

Duże modele językowe w przewidywaniu aktywności

Hipotezy badawcze

- Model jest w stanie przewidywać aktywność wobec różnych celów biologicznych na zadowalającym poziomie.
- Jesteśmy w stanie dowiedzieć się, co jest istotne dla aktywności danego związku chemicznego, np. przez wskazanie ważnych dla aktywności podstruktur.
- Dodawanie informacji o typie celu biologicznego, np. opis jego funkcji, poprawia przewidywania modelu.

Duże modele językowe w przewidywaniu aktywności

Kamienie milowe

- **Prezentacja śródsemestralna**
 - Zapoznanie się z tematami związanymi z inżynierią podpowidzi (prompt engineering)
 - Wybranie modeli językowych, na których prowadzone będą eksperymenty.
 - Wytypowanie celów biologicznych i znalezienie danych aktywności związków.
- **Prezentacja semestralna**
 - Konstrukcja zbiorów danych aktywnościowych.
 - Przygotowanie zapytań do modelu i ustanowienie metryk ewaluacyjnych.
 - Przygotowanie raportu wskazującego na skuteczność modeli językowych w przewidywaniu aktywności oraz możliwości wyjaśniania tych predykcji.