

Aplikacja metod uczenia głębokiego opartych o uwagę do znajdowania nowych aktywnych związków chemicznych

W ciągu ostatnich 10 lat, metody uczenia głębokiego zyskały ogromną popularność i zostały z powodzeniem zastosowane w wielu dziedzinach nauki, w których mamy dostęp do dużej ilości danych. Obecnie trwają badania nad zastosowaniem modeli głębokiego uczenia do komputerowo wspomaganego projektowania leków (CADD). Jednym z podstawowych problemów CADD jest przewidywanie właściwości związków chemicznych przy pomocy komputera (*in silico*) i generowanie nowych cząsteczek o pożądanych właściwościach. Projektowanie nowego leku jest długim i skomplikowanym procesem, który może trwać nawet 10 lat i kosztować do 10 miliardów dolarów. Dzięki algorytmom uczenia maszynowego, cała procedura może być szybsza i tańsza.

Celem naszego projektu jest dostarczenie modelu głębokiego uczenia opartego o uwagę do przewidywania aktywności związków chemicznych. Będzie on uczony na grafowej reprezentacji cząsteczki, sięc dostanie na wejściu informację o sąsiedztwie atomów, odległościach między nimi oraz cechach wiązań zachodzących między atomami. Będziemy również chcieli zbadać, jaka jest najlepsza reprezentacja wejściowa dla atomów przekazywanych do takich modeli.

W oparciu o nasz model chcemy również przeprowadzić optymalizację cząstek chemicznych, tak aby model potrafił zwracać cząsteczki podobne do zadanej, jednak o lepszych cechach. Trening chcemy przeprowadzić zarówno w trybie nadzorowanym, jak i korzystając z metod uczenia ze wzmocnieniem. Powstały model może być pomocny w fazie optymalizacji (lead optimization) w procesie projektowania leku, w której to fazie znaleźliśmy już pewną liczbę dobrych i interesujących cząsteczek (nazywanych *lead*), jednak chcemy dalej optymalizować niektóre z ich właściwości.

W naszych eksperymentach będziemy wykorzystywać dane z publicznie dostępnych baz danych, takich jak ChEMBL lub ZINC, a także dane zebrane w wyniku współpracy naszej grupy z grupą prof. A. Bjarskiego - obecnie mamy dostęp do zbioru danych zawierającego ponad 300 000 wyników eksperymentów dokowania. Dane będą wymagać przetwarzania wstępnego, aby były odpowiednie dla naszych metod. Skupimy się głównie na przewidywaniu i optymalizacji właściwości fizykochemicznych cząsteczek (takich jak rozpuszczalność), stabilności metabolicznej i aktywności w stosunku do zadanych białek.

Nasze metody mogą potencjalnie zmniejszyć koszty i czas produkcji nowych leków. Można będzie je wykorzystywać podczas procesu wirtualnego screeningu, podczas którego możemy odfiltrować niepożądane cząsteczki, co znacznie zmniejszy wydatki podczas testów laboratoryjnych. Cząsteczki wygenerowane przez nasze modele będą mogły również być zsyntetyzowane i przetestowane w laboratorium, a niezależnie od tego, będą mogły naprowadzić chemików medycznych do nowych, nieoczywistych podstawień, które skutkować mogą pojawieniem się nowych leków.

Metody powstałe w wyniku naszego grantu, mogą spowodować, że nowe leki będą tańsze i bardziej dostępne.