FÉLÉVES DOKTORI BESZÁMOLÓ

Bagoly Attila

2017/2018 tanév, 1. félév

Bevezetés

Az élő szervezet sejtjeiben egy komplex szignálhálózat működik, amely irányítja sejtjeink működését. Ezen hálózat legfontosabb elemei a fehérjék és szignálmolekulák (kis molekulák). A fehérjék különböző receptorokkal rendelkeznek, amelyekhez kis molekulák be tudnak kötni, így kiváltva valamilyen biológiai reakciót a fehérjéből (a reakció akár egy másik szignálmolekula felszabadulása is lehet, amely egy másik receptorba köt be). Adott receptorhoz bekötő molekulákat két csoportba szokás sorolni: agonista és antagonista. Az agonista molekulák aktiválják a receptort, azaz kiváltanak valamilyen biológiai hatást a fehérjéből, míg az antagonista molekulák bekötnek, és egyszerűen blokkolják a receptort (egy molekula lehet egy adott receptorra agonista, míg egy másikra antagonista is). Az embereket érintő betegségek jelentős része abból fakad, hogy a sejtekben működő szignálhálózatban hiba lép fel, például egy fehérje elkezd abnormális biológiai reakciót (pl. rossz szignálmolekulát szabadít fel, nem aktiválódik stb.) adni az aktivációjára. Ezért a mai gyógyszerfejlesztés egyik iránya, hogy meg kell keresni azt a receptort, amelynek, ha megakadályozzuk az aktivációját, akkor a betegség megszűnik. Ezt úgy érik el, hogy az adott receptorhoz terveznek antagonista molekulákat. Agonista molekulák alkalmazása jelentősen kisebb, mint az antagonista molekulák-é, mivel nehéz ilyen molekulákat találni (számolások korlátozottak, kipróbáláson alapuló keresés meg drága). Az antagonista gyógyszerjelölt molekulák, egy másik receptorra lehetnek agonisták is, így akár toxikus hatást is kiválthatnak. Ezért a gyógyszerfejlesztés folyamatában ezeket a lehetőségeket ki kell zárni, amely nagyrészt kísérletekkel történik, hozzájárulva a fejlesztési folyamat lassúságához és drágaságához. Az utóbbi években a figyelem központjába került deep learning lehetőséget biztosít a probléma egy újszerű megközelítésére, amely az eddigi eredmények alapján nagy előrelépést vetít előre a területen. Munkám során ezen technológiák alkalmazásával próbálok előrelépést elérni ezen a területen.

Az aktuális félévben elvégzett kutatások ismertetése

A félévben elkezdett munka során publikus adatbázisokban (pl. PubChem) elérhető adatok alapján dolgoztam. Az elérhető adatbázisok közül, nagyrészt a PubChem Bioassay adatbázisára fektettem a hangsúlyt, amelyben elérhető sok különböző receptorhoz tartozó agonista és antagonista molekulák listája. Kiindulópontként a molekulák háromdimenziós töltéseloszlásából kezdtem előre jelezni egy adott receptorra a molekulák aktivitását, háromdimenziós konvolúciós neuronhálózattal. Ez a feladatot két részfeladatra osztottam: töltéseloszlások generálása és deep learning rész. A PubChem Bioassay adatbázisából a 626-os ID-val rendelkező adathalmazt

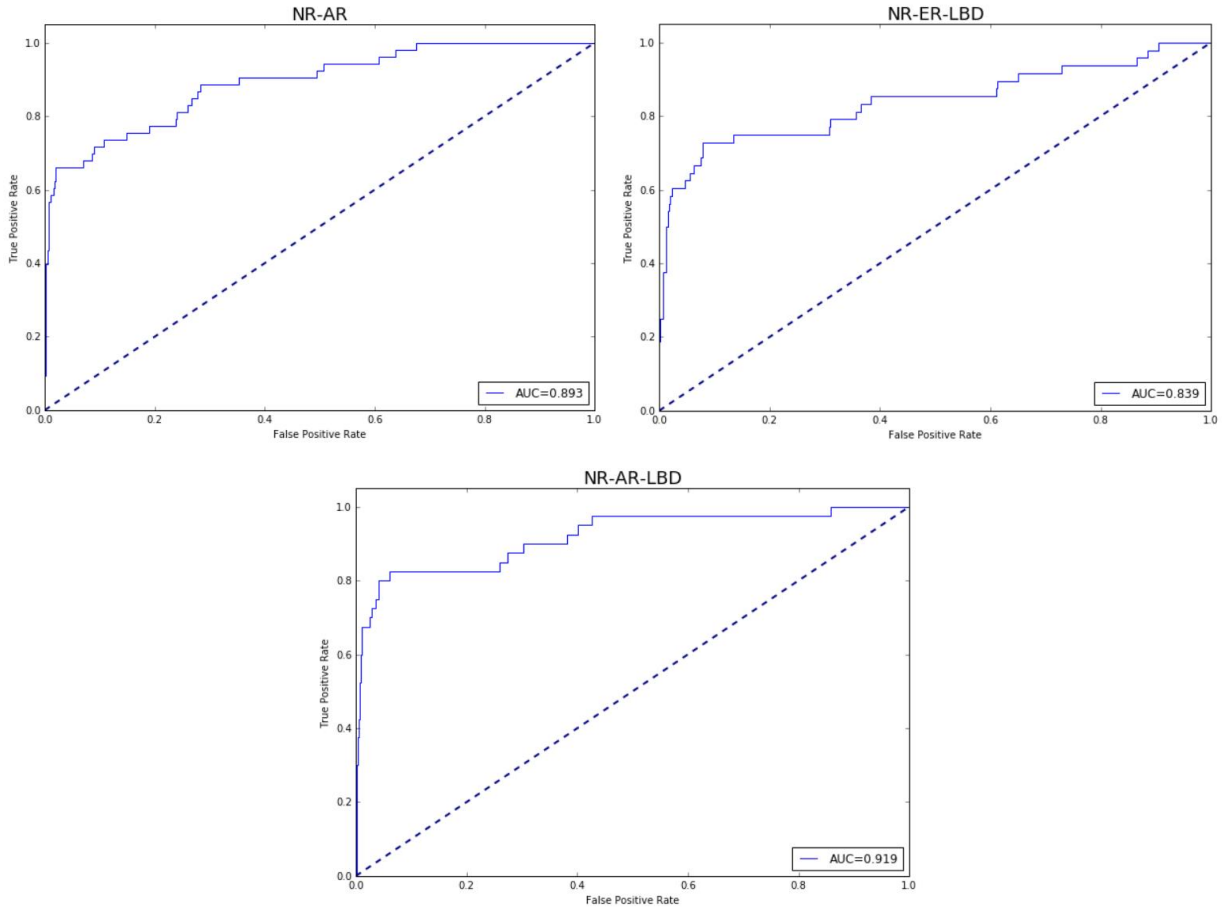
használtam, amely az M1 muszkarinos receptor agonista (aktív) és antagonista (inaktív) molekuláit tartalmazza, összesen 63682 molekula, amelyből 1938 molekula aktív.

Az első részben ezen közel 64 ezer molekula háromdimenziós töltéseloszlását generáltam le. Ennek elvégzésére az extended Hückel módszer segítségével meghatároztam a molekulapályákat, amelyek az atomi pályák kombinációjából állnak elő. A számolás elvégzésére használtam a YAeHMOP programot, amelynek kimenetéből kivettem a különböző molekulapályákhoz tartozó atomi pályák együtthatóit. A molekula pályák felösszegzéséből pedig megkaptam a töltéseloszlást. Mivel a töltéseloszlások kiértékelése háromdimenziós térfogatban, aránylag nagy felbontással történik rengeteg molekulára lényeges szempont az implementált számolás sebessége. Mivel a feladat másik részében a Google rendkívül népszerű nyílt forráskódú, lineáris algebra és deep learning könyvtárát használom, a TensorFlow-t, ezért az egész kvantumos eloszlások számolását egy C++-ban írt kiegészítőként implementáltam, amely egy új műveletként jelenik meg a TensorFlow-ban. A sebesség értékében GPU-ra párhuzamosítva CUDA-ban végeztem az implementációt, így egy naív python implementációhoz képest jelentős sebesség javulást értem el: egy molekula esetén a 10-20 perces számolást 1-2 másodpercesre tudtam csökkenteni.

A második lépésben háromdimenziós neuronhálókat tanítottam az aktivitás perdiktálására a legenerált töltéseloszlásokból. Mivel lényeges a molekuláknak a forgásinvarianciája, továbbá az adatszett mérete jelentősen elmarad a manapság tanításra használt adathalmazoktól, ezért fontos a molekulák random elforgatása. Mivel a deep learning területén belül a háromdimenziós objektumfelismerés még újabb terület, ezért a random háromdimenziós elforgatás még nem volt implementálva TensorFlow-ban (és más könyvtárakban sem), ezért ezt is egy új operációként implementáltam CUDA-ban. A PubChem Bioassay 626-os ID-val rendelkező adatszettre háromdimenziós töltéseloszlásból tanítással egy publikációban már foglalkoztak (töltéseloszlást azonban nem kvantumosan számolták) a Müncheni Technológiai Egyetemen, ezért elsőként az általuk kapott eredmény reprodukálása volt a célom (0.7 AUC az adatszetten), amelyet sikeresen teljesítettem. Ahhoz, hogy előrelépést lássunk fontos összevetni eredményeinket a régebbi eredményekkel. Erre a célra a 2014-ben rendezett TOX21 versenyt választottuk, ahol 11 ezer molekula különböző részhalmazainak affinitása van megadva összesen 12 receptorra. Ezeket a receptorokat aktiváló molekulák toxikusok az emberi szervezetre, ezért a molekulák aktivitásának meghatározása rendkívül fontos, ezért is hirdette meg ezen versenyt az amerikai toxikológiai intézet. Mivel ez az adathalmaz nagyon kicsi, esélytelen, egy teljes háromdimenziós neuronháló betanítása, ezért nagyobb adathalmazon halmazon más problémára betanított hálókat finomhangoltam, azaz valamilyen feladatra betanított háló súlyait használtam kezdőbeállításként a TOX21 adathalmazon. Az előzőekben említett feladaton betanított háló nem volt különösebben sikeres a TOX21 adathalmazon.

Az aktivitás előrejelzésére különböző molekula leírókon alapúló gépi tanuló algoritmusokkal szokták végezni. Tehát adott molekulákra előbb meghatároznak bizonyos jellemző értékeket (pl. tömeg, nehézatomok száma, gyűrűk száma stb.), majd ezeket az értékeket használják leíróként, azaz egyszerűbb klasszifikáló algoritmusok bemeneteként (pl. véletlenerdő, SVM stb.). A TOX21 verseny résztvevői is hasonló módon jártak el, a nyertes megkereste a hasznos leírók egy nagy halmazát és a legjobb algoritmusok kombinációját. Célunk lenne ezt legyőzni, deep learning

segítségével, amely kiváltja a molekulaleírók generálását, keresését és jóval meghaladhatja az ilyen módszerekkel elérhető eredményeket. Annak érdekében, hogy javítsunk az eredményen, új és egyszerűbb feladaton tanítottam a hálót. A legenerált töltéseloszláson a következő feladatot definiáltam: 3-nál kevesebb, 3 vagy 3-nál több gyűrű van az adott molekulán (így ezeknek a kategóriáknak egyenletes lett az eloszlása a halmazon). Ezen a feladaton tanítottam a hálót, 1 hét tanulás után 96% pontossággal tudta megmondani egy még nem látott molekuláról, hogy 3-nál kevesebb, 3 vagy 3-nál több gyűrű található benne. Ezen alapfeladaton betanított háló súlyait használva kezdeti beállításként, a hálót a TOX21 adathalmazon tanítva bíztató eredményeket értem el. A 12 alverseny (a 12 receptor mindegyike egy külön versenyhez tartozott) közül összesen 3 versenyben tudtam jobb eredményt elérni, mint a nyertes. Ezeknek a ROC görbéje és az általam elért AUC szám a teszthalmazon a következő ábrákon látható:



Tanulmányi tevékenység az aktuális félévben

Ebben a félévben a egyetem kínálatából elvégeztem a „Adatbázisok kezelése a csillagászatban” című tárgyat. Ezen kívül a coursera-n elvégeztem a gépi tanulás neves kutatójának, a Stanford tanárának a Deep Learning című specializáció első 4 kurzusát (Neural Networks and Deep

Learning, Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization, Structuring Machine Learning Projects, Convolutional Neural Networks).

Konferenciák az aktuális félévben

A Fizika Doktori Iskola támogatásával részt veszek január végén a háromnapos Applied Machine Learning Days című konferencián, amelyet Svájcban szervez az EPFL egyetem.