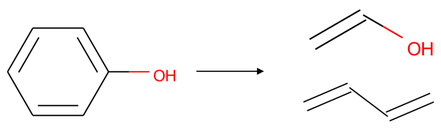
Modelování molekulární podobnosti pomocí fragmentů

Autor: Matyáš Lamprecht | Vedoucí: Mgr. Petr Škoda | Matematicko-fyzikální fakulta, Univerzita Karlova

Úvod

Nedílnou součástí vývoje léčiv je tzv. virtuální screening, jehož cílem je počítačová identifikace biologicky aktivních molekul. Jednou z variant virtuálního screeningu je ligandový virtuální screening, jenž je založen na využití známých biologicky aktivních molekul a podobnostního vyhledávání. Molekulu lze reprezentovat jako graf, molekulární podobnost lze pak modelovat na základě stejných fragmentů (podgrafů) mezi dvěma molekulami. Běžnou praxí je fragmenty hashovat do omezeného číselného intervalu a používat tato hashovaná čísla pro výpočet molekulární podobnosti. Při tomto hashování ovšem může dojít ke kolizím. Obecně jsou kolize považovány za nežádoucí, jelikož dochází ke ztrátě informace o molekule. Pro získání fragmentů molekuly se dnes používají molekulové otisky – AP, TT, ECFP, FCFP.



Vlevo je příklad grafu molekuly a vpravo je příklad jejích fragmentů.

Cíl práce

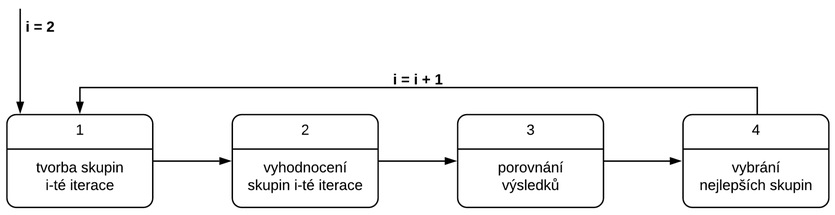
V této práci chceme otestovat, zda-li kolize fragmentů mohou vést k lepším výsledkům, než jsou výsledky běžně používaných metod a vybraných metod strojového učení. Za tímto účelem jsme navrhli několik podobnostních modelů postavených na fragmentech. Pro účely vyhodnocení jsme implementovali testovací prostředí, jenž umožňuje snadné testování a vyhodnocení různých modelů.

Postup řešení

Na kolize fragmentů se můžeme dívat jako na vytváření skupin fragmentů. Z fragmentů aktivních molekul odstraníme vícenásobné výskyty stejného fragmentu. Poté ze zbylých fragmentů skládáme dvojice (skupiny 2. iterace), ty vyhodnocujeme, porovnáváme s výsledkem běžně používané metody a ponecháme si pouze takové skupiny fragmentů, které dosáhly lepšího výsledku. Tyto skupiny poté skládáme mezi sebou – vzniknou 2 dvojice (pokud nemají stejný fragment) nebo 1 trojice (mají stejný fragment). Tyto skupiny nazýváme skupiny 3. iterace. Ty následně vyhodnocujeme a porovnáváme s výsledkem běžně používané metody. Takto iterujeme dále a vytváříme větší skupiny. Jelikož skupin, které dosáhly lepšího výsledku, bylo většinou mnoho a vyhodnocení skupin je časově náročné, tak jsme v každé iteraci vybírali pouze *n* nejlepších skupin.

Pro tento postup řešení jsme implementovali software, který má 2 části:

- Benchmarkovací program – slouží pro výpočet vyhodnocení jednoho modelu. Pro snadné přidávání nových modelů jsme použili návrhový vzor factory.
- Program skupiny – vytváří skupiny dané iterace pro daný model a vícejádrově je vyhodnocuje.



Průběh vytváření skupin.

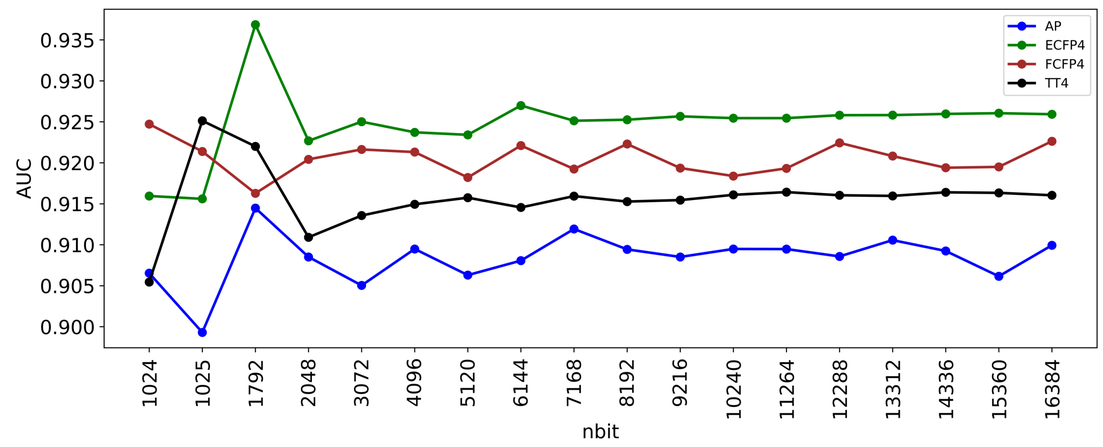
Současné metody a strojové učení

Typ	AUC	EF 1%
AP	0.913	63.7
TT4	0.914	58.8
ECFP4	0.927	63.7
FCFP4	0.92	53.9
rozhodovací strom	0.962	39.2
lineární regrese	0.819	4.9

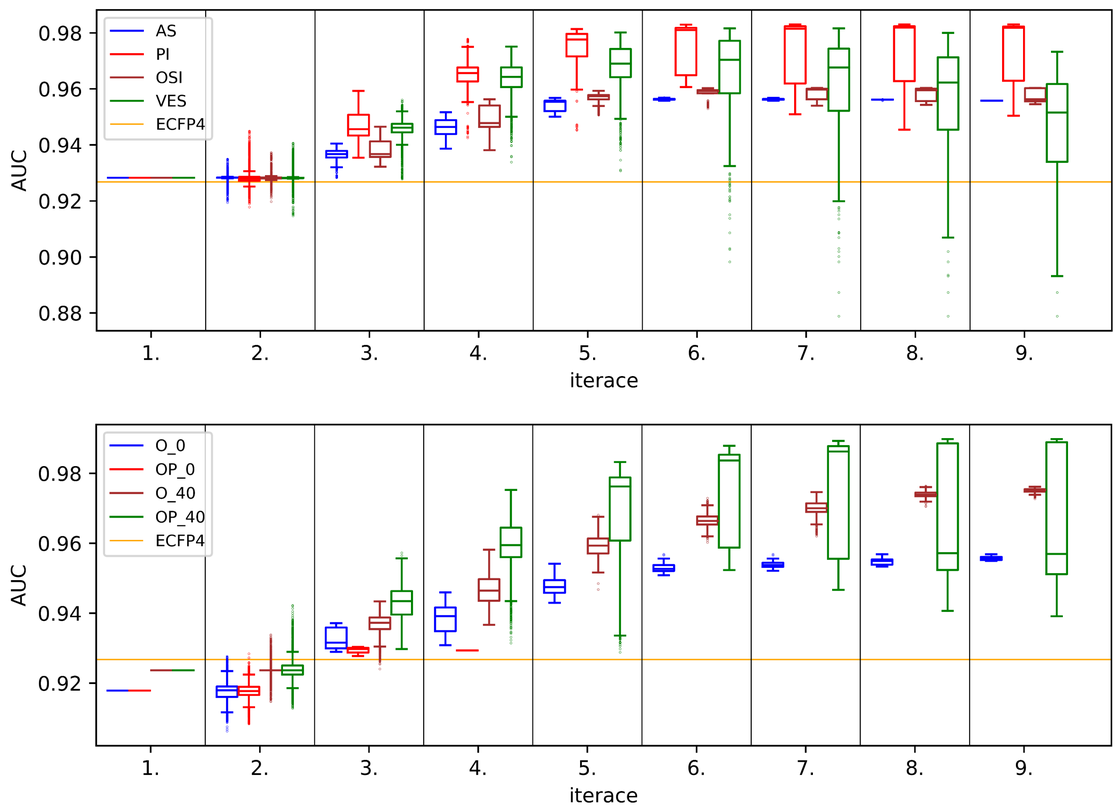
Porovnání běžně používaných metod a metod strojového učení.

Výsledky

Na následujícím obrázku je vidět, že pro určité hodnoty hashování můžeme dostat lepší výsledek, než je výsledek pro stejný molekulový otisk uvedený v tabulce.



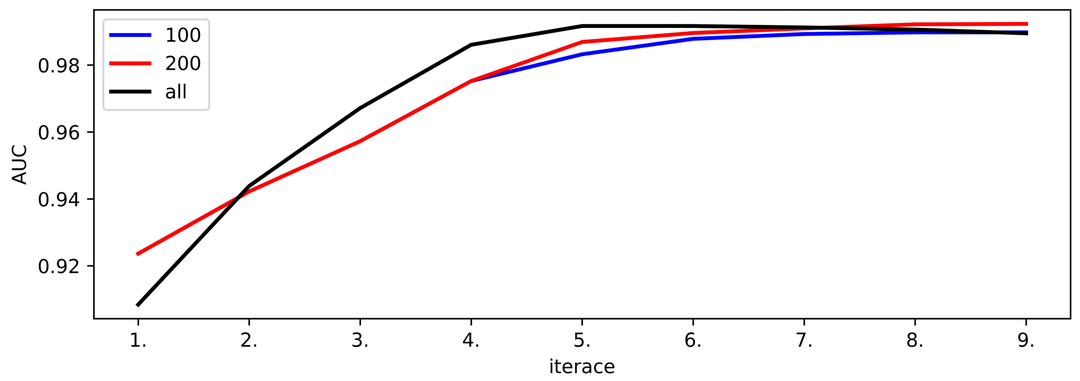
Vytvořili jsme 8 modelů, které pro výpočet podobností používají skupiny. Pro získání fragmentů, které budeme dávat do skupin, jsme použili fragmenty z ECFP4 a v každé iteraci jsme vybírali 100 nejlepších skupin. Výsledky jednotlivých modelů jsou uvedeny na následujících 2 obrázcích.



Porovnání 8 modelů.

Vidíme, že pro model OP_40 existují skupiny, které dosahují AUC ≐ 0.99. Pro skupinu s nejvyšším AUC navíc dostáváme EF 1% = 73.5.

Model OP_40, který měl ze všech modelů nejvyšší dosažené AUC, jsme otestovali pro výběr 200 nejlepších skupin v každé iteraci a poté jsem dali do trénovací sady i aktivní molekuly z testovací sady (all). Nejvyšší dosažená AUC pro jednotlivé iterace jsou vidět na následujícím obrázku.



Porovnání modelu OP_40.

Závěr

V námi implementovaných modelech mohou kolize vést k velice dobrým výsledkům, kterých jsme, v běžně používaných přístupech či přístupech strojového učení, nedosáhli. U většiny modelů platilo, že většina námi provedených kolizí byla prospěšná. Nicméně u určité kolize nedokážeme předem s jistotou říci, jestli díky ní budeme dostávat lepší či horší výsledky. Tedy pouze na základě fragmentů aktivních molekul nedokážeme určit skupiny, které budou dávat dobré výsledky. Ovšem pro modely existují fragmenty, které při tvorbě skupin mají větší pravděpodobnost pro vykazování dobrých výsledků.

Celkově můžeme říci, že správné kolize mohou vést ke zlepšení nacházení biologicky aktivních molekul.