Projekt

Celý projekt se skládá ze tří dílčích projektíků :-)

1) Predikce aktivity
2) Analýza
3) Vizualizace

Pro všechny tři části bude potřeba si sehnat aktivitní data, doporučoval bych na Chemblu si najít vhodný target, který má naměřeno alespoň 300 aktivit (spíš více a nejlépe ve stejné jednotce ;-). Nedáváme striktní zadání, ale spíš bychom chtěli, abyste si s tím nějak vyhráli. Odevzdejte to do složky `projekt` pod vaším jménem, nejlépe jako jupyterový notebook(y) i se vstupními daty.

Predikce aktivity

Pomocí scikit-learn natrénujte model, který bude predikovat aktivitu pro vámi zvolený target. Vaši sadu (asi náhodně) rozdělte na trénovací a testovací zhruba v poměru 70:30, na trénovací sadě naučte nějaký model (doporučil bych RandomForest s Morganovým fingerprintem) a sdělte nám přesnost na testovací sadě. Klidně si s tím můžete pohrát, udělat modelů víc a vybrat ten nejlepší. Nebo se rovnou rozšoupnout a zkusit klasifikovat agonisty a antagonisty (pravděpodobně to nedopadne, protože jsou si agonisté i antagonisté hodně podobný, zkuste serotonin 2A, který má odlišné [a/antago]nisty).

Analýza

Podívejte se jaké strukturní rysy obsahují vybrané ligandy. Můžete použít scaffoldy, MACCS keys, farmakofory nebo jakékoli jiné deskriptory. Můžete shlukovat, ukázat histogramy logP apod.

Vizualizace

Vizualizujte prostor Vámi vybraného targetu (hlavně jeho aktivních sloučenin) ať už pomocí PCA, MDS, Kohonenovy mapy. Barvou nebo velikostí naznačte aktivitu. Pokud jste v analýze udělali shlukování, můžete zobrazit clustery.

Různá vylepšení, co byste mohli (měli) použít

Standardizace struktur, alespoň odsolení a odstranění anorganiky. Aktivity by měli být v jedné jednotce, tzn. nekombinujte mM, nM, uM u koncentrací. Zvažte zda použít málo aktivní struktury (přes 10000 nM), nebo je naopak můžete použít pro klasifikaci jako neaktivní.