Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Matyáš Brabec
Název práce	Using combined sequence and structural features to predict protein-ligand binding sites
Rok odevzdání	2024
Studijní program	Informatika
Studijní obor	Softwarové a datové inženýrství

Autor posudku	David Hoksza
Role	Vedoucí
Pracoviště	KSI

Text posudku:

Metody identifikace protein-ligand interakcí byly donedávna založeny primárně na analýze 3D struktury. S příchodem proteinových jazykových modelů se ovšem ukazuje, že predikce vazebných míst je možná čistě z proteinové sekvence a to v kvalitě přibližující se kvalitám metod založených na 3D struktuře. Cílem práce pak bylo kvantifikovat rozdíl mezi těmito dvěma přístupy a navrhnout možnost jejich integrace.

Předložená diplomová práce je dělená do tří hlavních sekcí: úvod a motivace, popis navržených přístupů a analýza implementovaných metod.

Práce je interdisciplinárního charakteru; úvodní část práce je tedy věnována vysvětlení biologického pozadí problému a vhledu do domény predikce interakcí. Druhá část úvodu je pak věnována popisu metod strojového učení.

Druhá část práce je pak věnována popisu použitých přístupů a dat, které se používají k benchmarkingu vyvinutých metod. Navržené přístupy zahrnují integraci proteinových jazykových modelů a různých možností zapojení struktury. To zahrnuje jak zapojení protruze, tj. strukturní charakteristiky, která se v dřívějších pracích ukázala jako hlavní prediktor vazebného místa, tak různé způsoby agregace embeddingů z jazykových modelů na základě 3D vzdálenosti příslušných aminokyselin.

Vyhodnocení implemetovaných přístupů je provedeno na standardním benchmarku různých typů ligandů vzhledem k různým metrikám, včetně vyhodnocení statistické významnosti přínosu implemetnovaných metod vzhledem k baseline metodě, která odpovídá přímočarému použití embeddingů jazykových modelů spolu s MLP. Dále práce obsahuje porovnání se state-of-the-art metodou P2Rank.

Experimentální část práce obsahuje velké množství testů, které vyžadovalo netriviální množství práce pro nastavení procesů umožňující takovéto large-scale GPU výpočty.

Relativně elementární zapojení strukturní informace je vyváženo právě komplexitou zbytku práce. I tak práce ukazuje, že zapojení strukturní informace dokáže přinést lehké vylepšení oproti sekvenční baseline. Nicméně také se ukazuje , že nárust je minimální a výrazné vylepšení výkonu by vyžadovalo více netriviální způsob zapojení, případně zapojení netriviálních strukturních charakteristik.

Student pracoval pravidelně po celou dobu řešení projektu, komunikoval s vedoucím a ke konci aktivně navrhoval vlastní přístupy k řešení problému.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 30. května 2024 **Podpis**