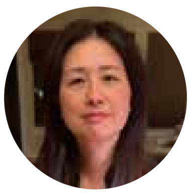
natureresearch

npj | Digital Medicine



Ying Ding

Tianhao Li

CancerGPT for few shot drug pair synergy prediction using large pretrained language models

Tianhao Li[1,5], Sandesh Shetty[2,5], Advaith Kamath[3], Ajay Jaiswal[1], Xiaoqian Jiang[4], Ying Ding[1] & Yejin Kim[4]

Large language models (LLMs) have been shown to have significant potential in few-shot learning across various fields, even with minimal training data. However, their ability to generalize to unseen tasks in more complex fields, such as biology and medicine has yet to be fully evaluated. LLMs can offer a promising alternative approach for biological inference, particularly in cases where structured data and sample size are limited, by extracting prior knowledge from text corpora. Here we report our proposed few-shot learning approach, which uses LLMs to predict the synergy of drug pairs in rare tissues that lack structured data and features. Our experiments, which involved seven rare tissues from different cancer types, demonstrate that the LLM-based prediction model achieves significant accuracy with very few or zero samples. Our proposed model, the CancerGPT (with ~ 124M parameters), is comparable to the larger fine-tuned GPT-3 model (with ~ 175B parameters). Our research contributes to tackling drug pair synergy prediction in rare tissues with limited data, and also advancing the use of LLMs for biological and medical inference tasks.

Foundation models have become the latest generation of artificial intelligence (AI)[1]. Instead of designing AI models that solve specific tasks one at a time, such foundation models or "generalist" models can be applied to many downstream tasks without specific training. For example, large pre-trained language models (LLMs), such as GPT-3[2] and GPT-4[3], have been a game changer in foundation AI model[4]. An LLM can apply its skills to unfamiliar tasks for which it has never been trained, known as few-shot learning or zero-shot learning. This is due in part to multitask learning, which enables LLM to unintentionally gain knowledge from implicit tasks in its training corpus[5]. Although LLMs have shown proficiency in few-shot learning in various fields[2], including natural language processing, robotics, and computer vision[2,6,7], their generalizability to unseen tasks in more complex fields, such as biology, has yet to be fully tested. In order to infer unseen biological reactions, knowledge of participating entities (e.g., genes, cells) and underlying biological mechanisms (e.g., pathways, genetic background, cellular environment) is required. While structured databases encode only a small portion of this knowledge, the vast majority is stored in free-text literature, which can be used to train LLMs. Thus, we envision that when there are limited structured data and limited sample sizes, LLMs can serve as an innovative approach for biological prediction tasks, by extracting prior knowledge from unstructured literature. One of such few-shot biological prediction tasks with a pressing need is a drug pair synergy prediction in understudied cancer types.

Drug combination therapy has become a widely accepted strategy for treating complex diseases such as cancer, infectious diseases, and neurological disorders[8]. In many cases, combination therapy can provide better treatment outcomes than single-drug therapy. Predicting drug pair synergy has become an important area of research in drug discovery and development. Drug pair synergy refers to the enhancement of the therapeutic effects of two (or more) drugs when used together compared to when each drug is used alone. The prediction of drug pair synergy can be challenging due to a large number of possible combinations and the complexity of the underlying biological mechanisms[9]. Several computational methods have been developed to predict drug pair synergy, particularly using machine learning. Machine learning models can be trained on large datasets of existing drug pair's experiment results to identify patterns and predict the likelihood of synergy for a new drug pair. Early studies in this area have relied on relational information or contextual information to extrapolate the synergy

[1]School of Information, University of Texas at Austin, Austin, TX, USA. [2]Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, USA. [3]Department of Chemical Engineering, University of Texas at Austin, Austin, TX, USA. [4]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA. [5]These authors contributed equally: Tianhao Li, Sandesh Shetty.
e-mail: yejin.kim@uth.tmc.edu

Li, Tianhao, Sandesh Shetty, Advaith Kamath, Ajay Jaiswal, Xiaoqian Jiang, Ying Ding, and Yejin Kim. (2024). "CancerGPT for Few Shot Drug Pair Synergy Prediction Using Large Pretrained Language Models." Npj Digital Medicine 7 (1): 40.