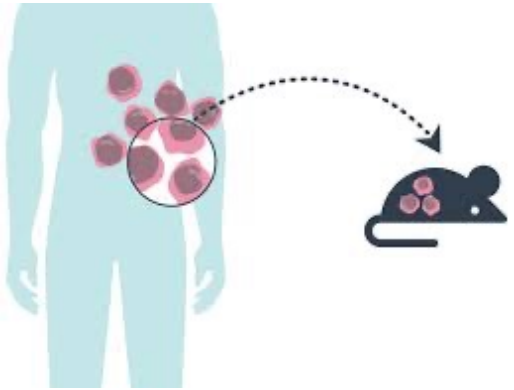
CancerGPT: Few-shot Drug Pair Synergy Prediction using Large Pre-trained Language Models

- Tianhao Li, Sandesh Shetty, Advaith Kamath, Ajay Jaiswal, Xiaoqian Jiang, Ying Ding & Yejin Kim
- University of Texas at Austin
- University of Massachusetts Amherst

Aim



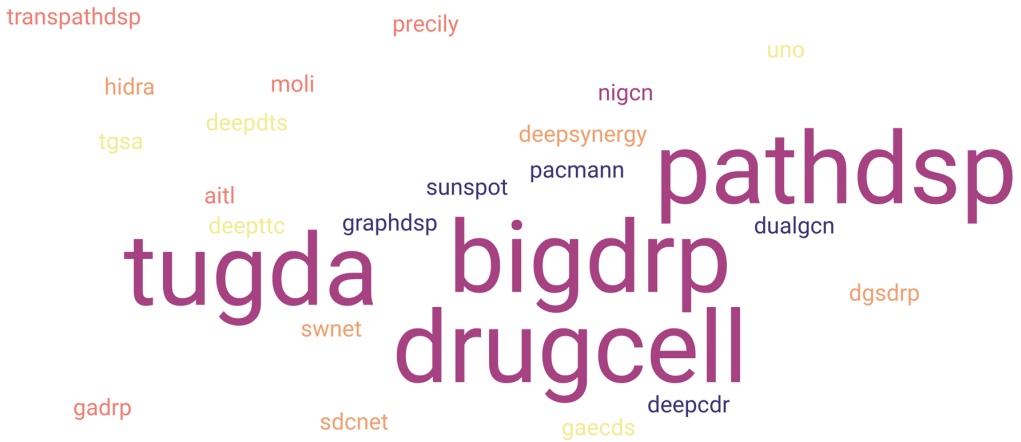
- Predict drug response between cell line or xenografts and drug
- Why?
  - Identify therapy (cancer cell lines)
- Datasets
  - Cancer Cell Line Encyclopedia (CCLE)
  - Genomics of Drug Study in Cancer (GDSC)
  - Cancer Therapeutics Response portal (CTRP)

Slide taken from Mayo Clinic

Prediction methods

- Many neural network method (Graph Autoencoder and Convolutional Neural Network)

- metrics
  - area under the precision-recall curve (AUROC)
  - area under the receiver operating curve (AUPRC)



Large language models

- Gemini/Gemma (Google)
- Llama (meta)
- Mistral

Pretrained large language model

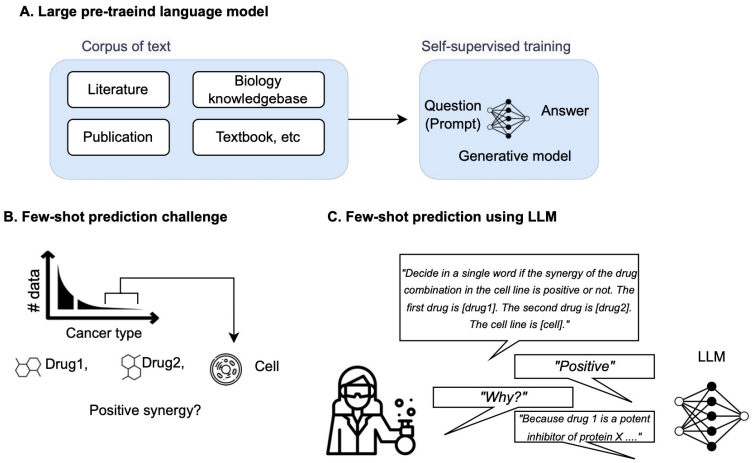


Figure 1: Few-shot prediction in biology. A. Different from task-specific approach, large pre-trained language model can perform new tasks which are not been explicitly trained for. B. Drug pair synergy prediction in rare tissues is an important examples of numerous few-shot prediction tasks in biology. C. Large pre-trained language model can be an innovative approach for few-shot prediction in biology thanks to its prior knowledge encoded in its weight.

Why drug pair synergy prediction to evaluate LLMs

Why drug pair synergy prediction to evaluate LLMs The prediction of drug pair synergy in uncommon tissues serves as an excellent benchmark task for evaluating LLMs in few-shot learning within the field of biology. This prediction requires incorporating multiple pieces of information, such as drug and cell line, as well as the sensitivity of drugs to the cell lines, in order to infer the synergistic effects. While detailed information on these entities can be found in scientific papers, the interaction effect, or synergistic effect, is primarily available through biological experiments. To effectively assess LLMs' inference capabilities, one must employ a prediction task where the ground truth is not explicitly available in text format but can be determined through alternative sources for model evaluation. Typically,

Motivation

Contribution The contribution of our study can be summarized as follows. In the area of drug pair synergy prediction in rare tissues, our study is the first to predict drug pair synergy on tissues with very limited data and features, which other previous prediction models have neglected. This breakthrough in drug pair synergy prediction could have significant implications for drug development in these cancer types. By accurately predicting which drug pair will have a synergistic effect on these tissues in which cell lines are expensive to obtain, biologists can directly zoom into the most probable drug pairs and perform in vitro experiments in a cost effective manner.

Our study also delivers generalizable insights about LLMs in the broader context of biology. To the best of our knowledge, our study was the first to investigate the use of LLMs as an few-shot inference tool based on prior knowledge in the field of biology, where much of the latest information is presented in unstructured free text (such as scientific literature). This innovative approach could have significant implications for advancing computational

Pipeline fo synergy prediction

1. Drug pair synergy data

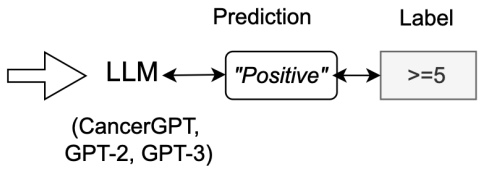
Drug1	Drug2	Cell line	Tissue	Drug1 sensitivity	Drug 2 sensitivity	Synergy
ABT-888	MK-8776	ES2	Bone	-1.625	48.756	<5
lonidamine	717906-29-1	A-673	Bone	0.568	28.871	>=5
AZD1775	AZACITIDINE	EW-8	Bone	25.687	1.752	?

2. Convert tabular input and prediction task to natural text

Prompt: *"Decide in a single word if the synergy of the drug combination in the cell line is positive or not*

Converted string input: *"Drug combination and cell line: The first drug is lonidamine. The second drug is 717906-29-1. The cell line is A-673. Tissue is bone. The first drug's sensitivity using relative inhibition is 0.568. The second drug's sensitivity using relative inhibition is 28.871. Synergy:"*

3. *k*-shot finetuning

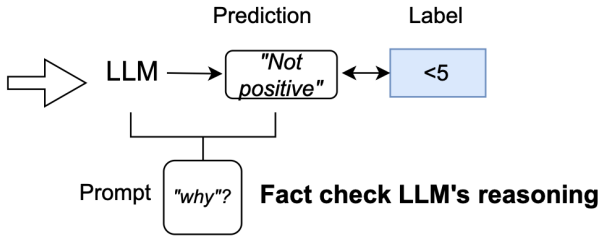


4. Predict drug pair synergy

Prompt: *"Decide in a single word if the synergy of the drug combination in the cell line is positive or not"*

Converted string input: *"Drug combination and cell line: The first drug is AZD1775. The second drug is AZACITIDINE. The cell line is EW-8. Tissue is bone. The first drug's sensitivity using relative inhibition is 25.687. The second drug's sensitivity using relative inhibition is 1.752. Synergy:"*

5. Evaluate accuracy



Synergy prediction models based on Large pre-trained language models

5.2. Synergy prediction models based on Large pre-trained language models

Converting tabular input to natural text To use an LLM for tabular data, the tabular input and prediction task must be transformed into a natural text. For each instance of tabular data (Fig. 2), we converted the structured features into text. For example, given the feature string (e.g., "drug1", "drug 2", "cell line", "tissue", "sensitivity1", "sensitivity2") and its value (e.g., "lonidamine", "717906-29-1", "A-673", "bone", "0.568", "28.871"), we converted the instance as *"The first drug is AZD1775. The second drug is AZACITIDINE. The cell line is SF-295. Tissue is bone. The first drug's sensitivity using relative inhibition is 0.568. The second drug's sensitivity using relative inhibition is 28.871."* Other alternative ways to convert the tabular instance into the natural text are discussed in previous papers

Synergy prediction models based on Large pre-trained language models

Converting prediction task into natural text We created a prompt that specifies our tasks and guides the LLM to generate a label of our interest. We experimented with multiple prompts. One example of the prompts we created was *"Determine cancer drug combination synergy for the following drugs. Allowed synergies: Positive, Not positive. Tabular Input . Synergy:"*. As our task is a binary classification, we created the prompt to only generate binary answers (*"Positive"*, *"Not positive"*). Comparing these multiple

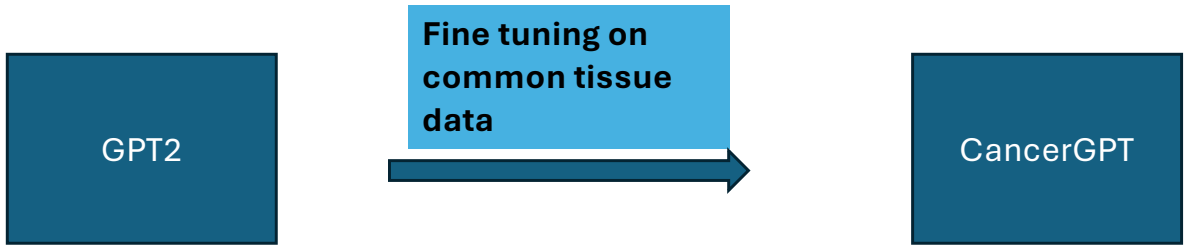
Synergy prediction models based on Large pre-trained language models

CancerGPT:

- Large language models: (using API for fine-tuning)
- To adjust the model for a binary classification task:

  + add a linear layer as a sequence classification head on top of GPT-2, which uses the last token of the output of GPT-2 to classify the input.

- The cross-entropy loss was used during the fine-tuning process



Synergy prediction models based on Large pre-trained language models

k-shot fine-tuning strategy

k from [0, 2, 4, 8, 16, 32, 64, 128]

For bone, urinary tract, stomach, soft tissues, and liver rare tissues

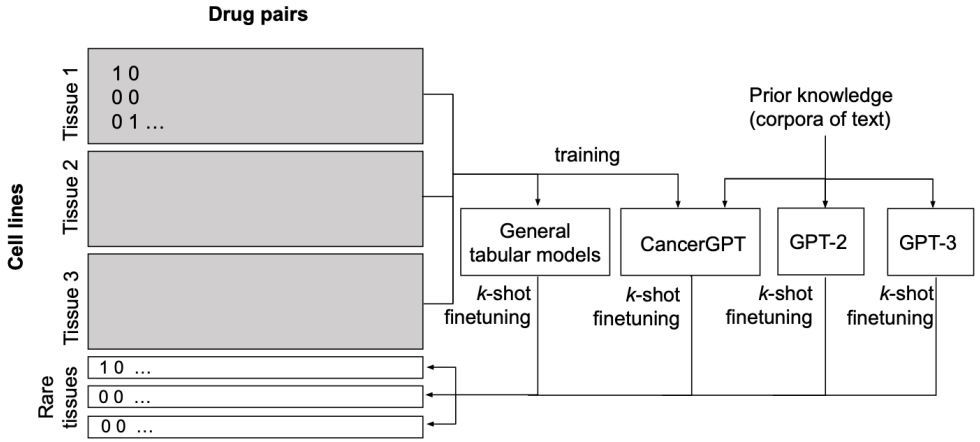


Figure 3: Training strategy of baseline and proposed LLM-based models. General tabular models and CancerGPT were first trained with samples from common tissues then k-shot fine-tuned with each tissue of interest. GPT-2 and GPT-3 are pre-trained models, and we fine-tuned them with k shots of data in each tissue.

Findings: Number of training data and accuracy

Number of training data and accuracy Overall, the LLM-based models (CancerGPT, GPT-2, GPT-3) achieved comparable or better accuracy in most of the cases compared to baselines. In the zero-shot scenario, the LLM-based models generally had higher accuracy than the baseline models in all experiments except stomach and bone. As the number of shots increased, we observed mixed patterns across various tissues and models. TabTransformer consistently exhibited an increase in accuracy with more shots. CancerGPT showed higher accuracy with more shots in the endometrium and soft tissue, and GPT-3 showed higher accuracy with more shots in the liver, soft tissues, and bone, indicating that the information gained from a few shots of data complements the prior knowledge encoded in CancerGPT and GPT-3.

However, the LLM-based models sometimes did not show significant improvements in accuracy in certain tissues, such as the stomach and urinary tract, suggesting that the additional training data do not always improve the LLM-based models' performance. With the maximum number of shots (k=128), the LLM-based model, specifically GPT-3, was on par with TabTransformer, achieving the highest accuracy with the pancreas, liver, soft tissue, and bone, while TabTransformer achieved the best accuracy with endometrium, stomach, and urinary tract.

Findings: Number of training data and accuracy

- Overall, the LLM-based models (CancerGPT, GPT-2, GPT-3) achieved comparable or better accuracy in most of the cases compared to baselines: XGBoost and TabTransformer

- In the zero-shot scenario, the LLM-based models generally had higher accuracy than the baseline models in all experiments except stomach and bone.

- TabTrans- former consistently exhibited an increase in accuracy with more shots., CancerGPT higher accuracy with more shots in the endometrium and soft tissue, GPT-3 in the liver, soft tissues, and bone

- LLM-based models sometimes did not show significant improvements in accuracy in certain tissues, such as the stomach and urinary tract

Findings: Comparing LLM-based models

Comparing LLM-based models When comparing LLM-based models, CancerGPT and GPT-3 demonstrated superior accuracy compared to GPT-2 in most tissues. GPT-3 exhibited higher accuracy than CancerGPT in tissues with limited data or unique characteristics, while CancerGPT performed better than GPT-3 in tissues with less distinctive characteristics, such as the stomach and urinary tract. The higher accuracy of CancerGPT compared to GPT-2 highlights that well-balanced adjustment to specific tasks can increase the accuracy while maintaining generalizability. However, the benefits of such adjustments

Findings: Fact check LLM's reasoning

We prompted the LLMs with "Could you provide details why are the drug1 and drug2 synergistic in the cell line for a given cancer type?".

Excerpt of the generated answer	Fact check and reference
"The combination of AZD-4877 and AZD1208 has been studied in T24 cells...to be synergistic in reducing bladder cancer cell growth and metastasis"	False. No study conducted on this drug pair
"The combination was also found to target multiple pathways involved in the growth and spread of bladder cancer cells."	True. AZD1208 is a PIM1 inhibitor. PIM1 is overexpressed in bladder cancer initiation and progression (Guo et al. (2010)). AZD4877 is a drug designed to target bladder cancer (Jones et al. (2013)).
"...Specifically, AZD-4877 was found to inhibit the activation of proteins involved in the promotion of tumor growth..."	True. AZD4877 is a drug designed to target bladder cancer (Jones et al. (2013)).
"...AZD1208 was found to inhibit proteins associated with the inhibition of tumor growth."	True. AZD1208 inhibits the cell growth by suppressing p70S6K, 4EBP1 phosphorylation, and messenger RNA translation (in acute myeloid leukemia) (Cortes et al. (2018)).
"This combination was also effective at reducing the production of inflammatory mediators such as cytokines, which are known to contribute to tumor progression."	False. AZD1208 is a pan-PIM kinase inhibitor, and PIM kinases are downstream effectors of cytokine (National Cancer Institute (2011)). However, AZD4877 has no evidence in reducing inflammatory mediators.
"...these two drugs have been shown to reduce levels of apoptosis inhibitors, which can also play a role in tumor progression."	True. AZD1208 induce cell apoptosis (Cervantes-Gomez et al. (2019)). AZD4877 is a inhibitor of Eg5, which promotes cell apoptosis (Borthakur et al. (2009)).

Table 3: Example of generated answer when the LLM was asked to provide its reasoning for its prediction

Limitations:

- The present study, while aiming to showcase the potential of LLMs as a few-shot prediction model in the field of biology. To fully establish the generalizability of LLMs as a “generalist” artificial intelligence, a wider range of biological prediction tasks must be undertaken to validate it.

- Crucial to investigate how the information gleaned from LLMs complements the existing genomic or chemical features that have traditionally been the primary source of predictive information.

- GPT-3’s reasoning, the accuracy of its arguments cannot always be verified and may be susceptible to hallucination.

Thank you!