

<https://doi.org/10.1038/s41746-024-01024-9>

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



Tianhao Li^{1,5}, Sandesh Shetty^{2,5}, Advait Kamath³, Ajay Jaiswal¹, Xiaoqian Jiang⁴, Ying Ding¹ & Yejin Kim⁴ ✉

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-3² and GPT-4³, have been a game changer in foundation AI model⁴. 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 corpus⁵. Although LLMs have shown proficiency in few-shot learning in various fields², 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 disorders⁸. 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 mechanisms⁹. 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

¹School of Information, University of Texas at Austin, Austin, TX, USA. ²Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, USA. ³Department of Chemical Engineering, University of Texas at Austin, Austin, TX, USA. ⁴McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA. ⁵These authors contributed equally: Tianhao Li, Sandesh Shetty.

✉ e-mail: yejin.kim@uth.tmc.edu

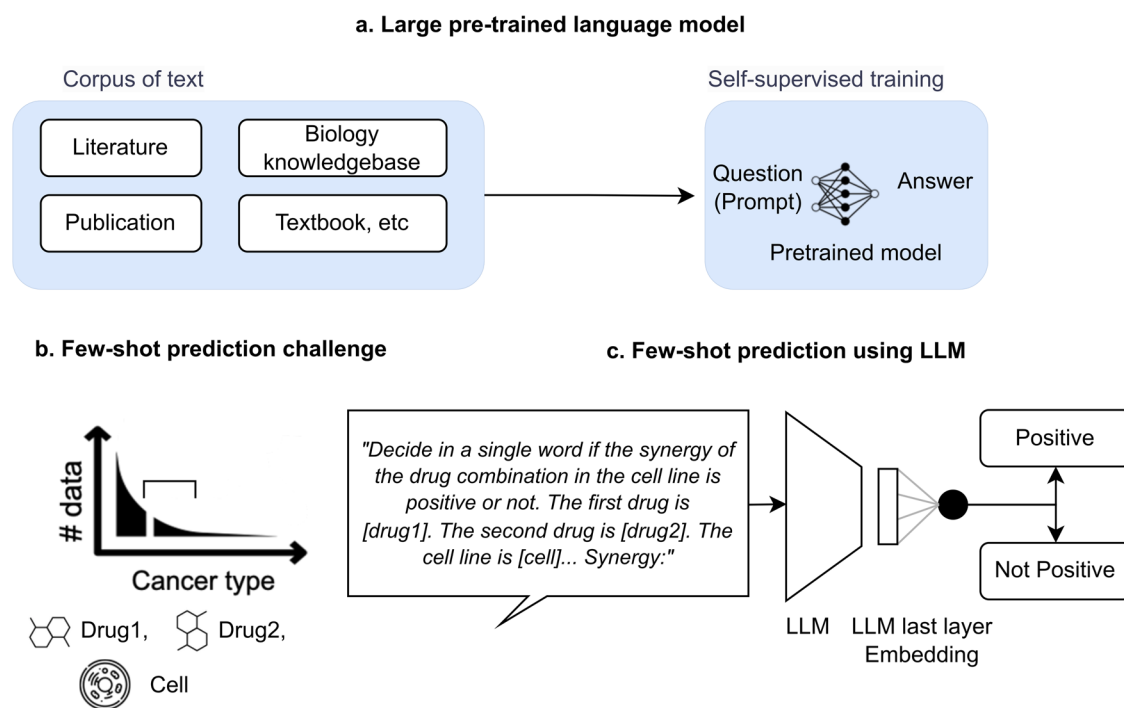


Fig. 1 | Few-shot prediction in biomedicine. **a** Different from a task-specific approach, a large pre-trained language model can perform new tasks for which it has not been explicitly trained. **b** Drug pair synergy prediction in rare tissues is an

important example of the numerous few-shot prediction tasks in biomedicine. **c** A large pre-trained language model can be an innovative approach for few-shot prediction in biomedicine, thanks to the prior knowledge encoded in its weight.

score regarding a specific cell line to cell lines in other tissues^{10–14}, ignoring the biological and cellular differences in these tissues. Another line of studies has sought to overcome the discrepancy between tissues by utilizing diverse and high-dimensional features, including genomic (e.g., gene expression of cell lines) or chemical profiles (e.g., drug structure)^{13,15–18}. However, most of the available data comes from common cancer types in certain tissues, such as breast and lung cancer; very limited experiment data are available on certain tissues, such as bone and soft tissue (Fig. 1). Obtaining cell lines from these tissues can be physically difficult and expensive, which limits the number of training data available for drug pair synergy prediction. This can make it challenging to train machine learning models that rely on large datasets. ref.¹⁸ used transfer learning to extend the prediction model trained in common tissues to some of the rare tissues with relatively rich data and cellular features. However, it cannot be utilized for rare tissues with extremely limited data and cellular information.

In this work, we aim to overcome the above challenge by LLMs. We hypothesize that cancer types with limited structured data and features still have rich information in scientific literature. Manually extracting predictive information on such biological entities from literature is a complex task. Our innovative approach is to leverage prior knowledge in scientific literature encoded in LLMs. We build a few-shot drug pair synergy prediction model that leverages representation from LLM as a prior knowledge and predict the outcome. Our experimental results demonstrate that our LLM-based few-shot prediction model achieved significant accuracy even in zero-shot setting (i.e., no training data) and outperformed strong tabular prediction models in most cases. This remarkable few-shot prediction performance in one of the most challenging biological prediction tasks has a critical and timely implication to a broad community of biomedicine because it shows a strong promise in the “generalist” biomedical artificial intelligence¹.

Results

Results overview

We developed CancerGPT, a few-shot drug pair synergy prediction model for rare tissues. Leveraging LLMs-based tabular data prediction model¹⁹, we converted the prediction task into a natural language task, derived

embedding from prior knowledge encoded in LLM’s pre-trained weight matrices, and built classification model for synergy prediction (Fig. 2). We presented our strategy to adapt the LLM to our task with only a few shots of training data in each rare tissue using k -shot fine-tuning strategy (Fig. 6).

We evaluated the accuracy of our proposed CancerGPT model, other LLM-based models (GPT-2, GPT-3, SciFive²⁰), and general data-driven prediction models (XGBoost, Collaborative Filtering, TabTransformer) (Methods, Fig. 3, 4, Supplementary Figs. 1, and 2, Supplementary Table 2-5) by the area under the precision-recall curve (AUPRC) and the area under the receiver operating curve (AUROC) under the different settings. We considered different few-shot learning scenarios, where the model is provided with a limited number k of training data to learn from ($k = 0$ to 128). By varying the number of shots, we can examine the model’s ability to adapt and generalize with minimal training data. Next, we compared the performance of CancerGPT and other LLM-based models across different tissues and model settings. We then investigated whether the LLM’s reasoning for its prediction is valid by checking its argument with scientific literature.

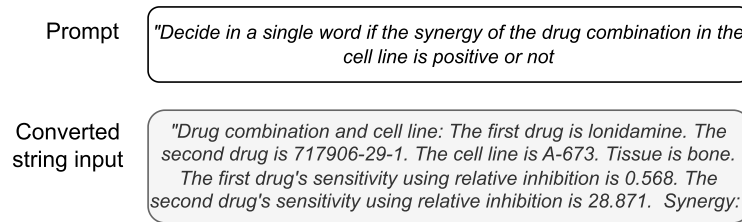
Accuracy

CancerGPT showed the highest accuracy in tissues such as the liver, soft tissue, and urinary tract. CancerGPT’s accuracy increases as the number of training shots increases (Figs. 3, 4), indicating that the information gained from a few shots of data complements the prior knowledge encoded in LLM and the information in external data. In contrast, the two data-driven models, Collaborative filtering and TabTransformer, showed higher accuracy in some tissues, including endometrium, stomach, and bone, indicating that the patterns learned from external data (common tissues) can be extrapolated to these tissues. The discrepancy in accuracy between CancerGPT (which utilizes prior knowledge and external data) and the data-driven model can be attributed to the similarity of the external data’s distribution to the data of interest. Certain rare tissue cancer cell lines show unique characteristics in comparison to common tissues. These tissues have specific cellular characteristics that are unique to their tissue of origin, which may not be accurately predicted through training with common tissue. For example, hepatic cell lines, which originate from liver tissue, have unique

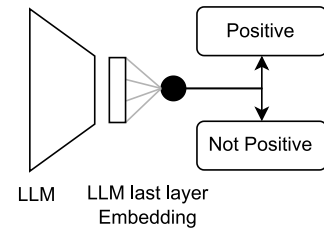
a. Drug pair synergy data

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	?

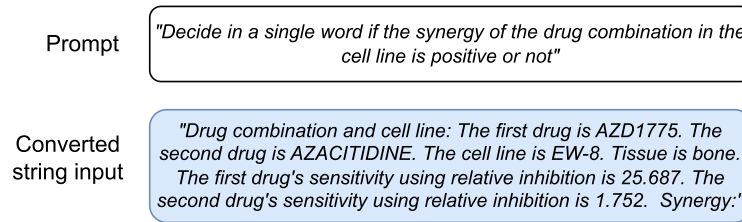
b. Convert tabular input and prediction task to natural text



c. k-shot finetuning



d. Predict drug pair synergy



e. Evaluate accuracy

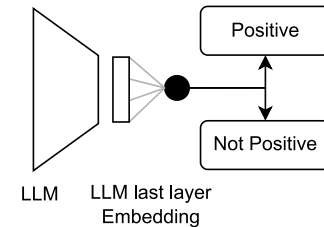


Fig. 2 | Study workflow. We converted the tabular input (a) to natural text and created a task-specific prompt (b). The prompt was designed for the LLM to give a binary answer (i.e., "Positive", "Not positive"), and the LLM's embedding from the last token was used for the final prediction (c, d). We evaluated and compared the prediction models with a different number of shots and tissues (e). We investigated the LLM's reasoning based on factual evidence. LLM Large Language Model.

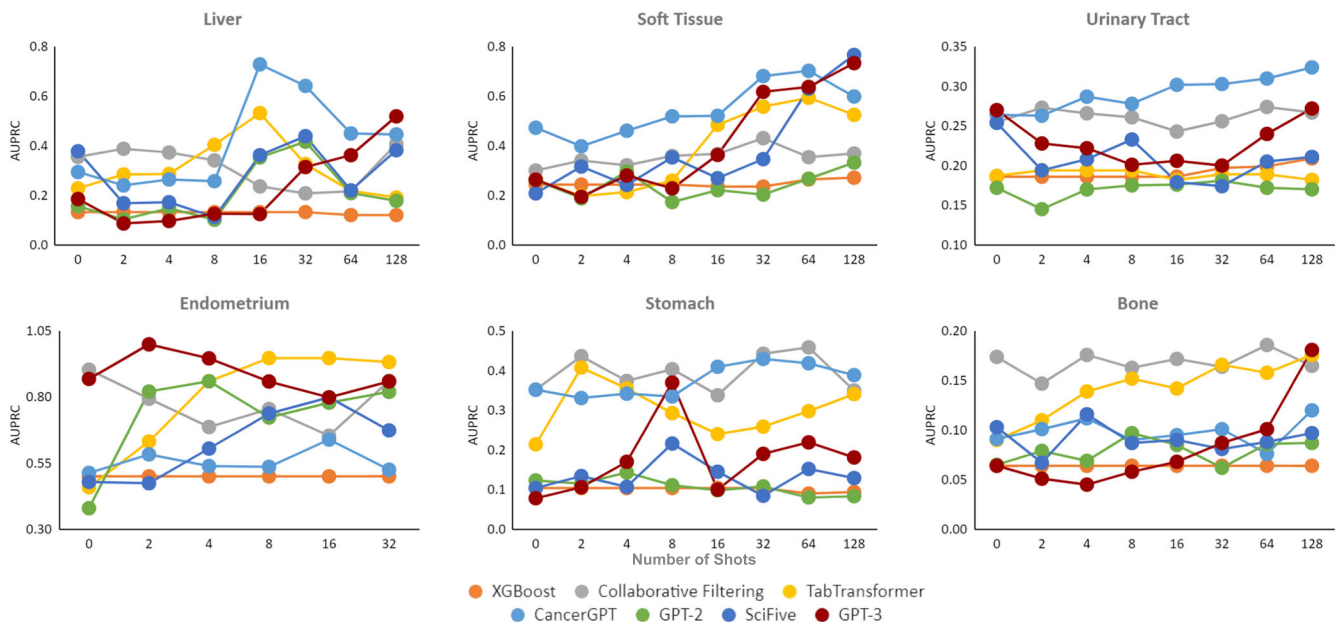


Fig. 3 | AUPRC of k-shot learning on seven tissue sets. We used 20% data as a test set in each rare tissue, while ensuring the binary labels were equally represented. AUPRC Area Under the Precision Recall Curve. Legend Orange: XGBoost, Gray: Collaborative Filtering, Yellow: TabTransformer, Light Blue: CancerGPT, Green: GPT-2, Dark Blue: SciFive, Red: GPT-3.

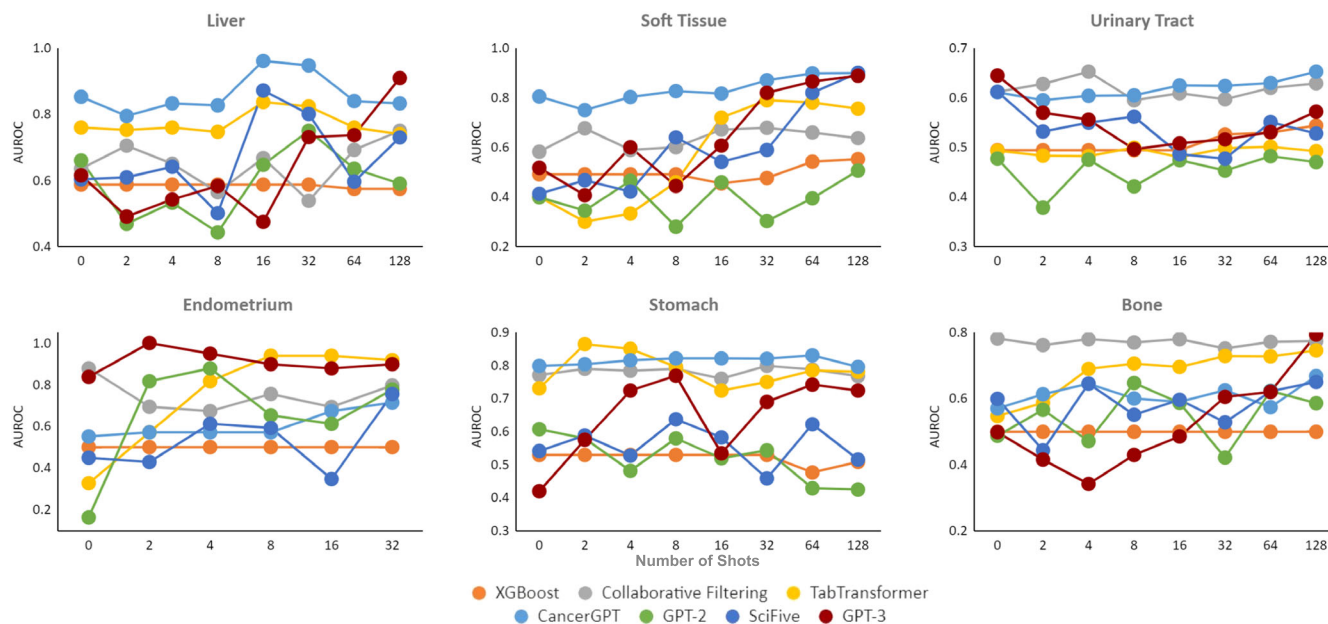


Fig. 4 | AUROC of k -shot learning on seven tissue sets. We used 20% data as a test set in each rare tissue, while ensuring the binary labels were equally represented. AUROC Area Under the Receiver Operating Curve. Legend Orange: XGBoost, Gray:

Collaborative Filtering, Yellow: TabTransformer, Light Blue: CancerGPT, Green: GPT-2, Dark Blue: SciFive, Red: GPT-3.

drug response characteristics due to high expression of drug-metabolizing enzymes such as cytochrome P450s²¹. When the data from common tissue is distributed differently to each rare tissue (i.e., external data is out-of-distribution), CancerGPT outperforms the general data-driven prediction model trained with out-of-distribution data, thanks to general prior knowledge already encoded in the LLM parameters. On the other hand, TabTransformer and Collaborative filtering, trained with external data, achieved the best accuracy in tissues such as endometrium, stomach, and bone. We further investigated the genomic characteristics of these tissues to ascertain any commonalities with common tissues. Supplementary Figure 3 showed the lower dimensional projection of gene expression of cancer cell lines grouped by tissue types²². In this figure, gene expression of tumorous cells from the endometrium (uterus), stomach (gastric), and bone were similar to that from other majority of tissues. Conversely, the gene expression of tumorous cells from liver, soft tissue (including Ewing sarcoma), and urinary tract formed unique clusters that were distinct from other majority of tissues. When the data from common tissues are distributed similarly to each rare tissue (in-distribution), the data-driven prediction model outperforms the few-shot LLM-based models. In all, the data-driven model showed superior accuracy when in-distribution external data was available, compared to all the LLM-based models, including CancerGPT. However, in the absence of external data, or when the external data is out-of-distribution, our customized LLM-based model achieved the best accuracy.

When comparing the accuracy of LLM-based prediction models, CancerGPT emerged as the most accurate one. This superior performance is attributable to its fine-tuning with external data, tailored specifically to the task at hand. GPT-3 also proved to be a competitive model, showing a potential to enhance its accuracy with an increasing number of shots. This characteristic renders it an optimal choice for datasets that permit the abundant additional training samples. Particularly, GPT-3 scored the highest accuracy in pancreas tissue, in which only zero-shot tuning is possible due to the limited sample size (Supplementary Tables 2 and 3). Despite SciFive's specialized design for scientific literature and its substantial parameter size of 220M, it failed to consistently outperform GPT-2 across all datasets. The SciFive's lower accuracy despite the larger parameters and adaptation to scientific literature might be due to its difference in base LLM (T5 and GPT). The higher accuracy of CancerGPT over GPT-2 underscores

the value of task-specific adjustments. These modifications can augment accuracy while preserving a model's versatility. However, the advantages of such fine-tuning may diminish with larger LLM models like GPT-3, particularly in scenarios demanding greater generalizability. Interestingly, CancerGPT, despite its smaller parameter size of 124M, outperformed GPT-3's accuracy, which has a larger parameter size of 175B. This suggests that further fine-tuning of GPT-3 may potentially yield even greater accuracy, provided that in-distribution data is readily available.

We also compared the accuracy when the models are fine-tuned with different strategies. Full training, which updates both LLM parameters and classification head during k -shot tuning, generally showed higher accuracy over last layer training (Supplementary Figs. 1 and 2), which only updates the classification head while freezing the LLM parameters. However, the marginal increase in accuracy, despite the extensive tuning of the millions or even billions of parameters was not significantly high. This suggests that the last layer representation of LLM already encapsulates a substantial amount of prior knowledge, effectively serving as a foundation for downstream tasks. The observation that last layer training, which leaves the LLM's backbone untouched, delivers accuracy comparable to that of a fully trained model, provides an important insight. An LLM pre-trained with an extensive collection of scientific literature can flexibly enhance biomedical predictions via transfer learning, which is similar to how a pre-trained model with ImageNet²³ can augment image analysis through transfer learning.

Fact check LLM's reasoning

We evaluated whether the LLM can provide the biological reasoning behind its prediction. In this experiment, we used zero-shot GPT-3 because other fine-tuned LLM-based models (GPT-2, SciFive, CancerGPT) compromised its language generative performance during the fine-tuning and were not able to provide coherent responses. To do this, we randomly selected one true positive prediction and examined whether its biological rationale was based on factual evidence or mere hallucination. Our example was the drug pair AZD4877 and AZD1208 at cell line T24 for urinary tract tissue. We prompted the LLMs with "Could you provide details on why the drug1 and drug2 are synergistic in the cell line for a given cancer type?" Details on prompt generation are discussed in Supplementary Note 1. We evaluated the generated answer by comparing it with existing scientific literature. We found that the LLM provided mostly accurate arguments, except for two

Table 1 | Example of generated answer when the LLM was asked to provide its reasoning for its prediction

Excerpt of the generated answer	Fact check and reference
<i>"The combination of AZD-4877 and AZD1208 has been studied in T24 cells...to be synergistic in reducing bladder cancer cell growth and metastasis"</i>	False. No study conducted on this drug pair
<i>"The combination was also found to target multiple pathways involved in the growth and spread of bladder cancer cells."</i>	True. AZD1208 is a PIM1 inhibitor. PIM1 is overexpressed in bladder cancer initiation and progression ⁽⁴⁵⁾ . AZD4877 is a drug designed to target bladder cancer ⁽⁴³⁾ .
<i>"...Specifically, AZD-4877 was found to inhibit the activation of proteins involved in the promotion of tumor growth..."</i>	True. AZD4877 is a drug designed to target bladder cancer ⁽⁴³⁾ .
<i>"...AZD1208 was found to inhibit proteins associated with the inhibition of tumor growth."</i>	True. AZD1208 inhibits the cell growth by suppressing p70S6K, 4EBP1 phosphorylation, and messenger RNA translation (in acute myeloid leukemia) ⁽⁴⁴⁾ .
<i>"This combination was also effective at reducing the production of inflammatory mediators such as cytokines, which are known to contribute to tumor progression."</i>	False. AZD1208 is a pan-PIM kinase inhibitor, and PIM kinases are downstream effectors of cytokine ⁽⁴⁵⁾ . However, AZD4877 has no evidence in reducing inflammatory mediators.
<i>"...these two drugs have been shown to reduce levels of apoptosis inhibitors, which can also play a role in tumor progression."</i>	True. AZD1208 induce cell apoptosis ⁽⁴⁶⁾ . AZD4877 is an inhibitor of Eg5, which promotes cell apoptosis ⁽⁴⁷⁾ .

Italic denotes the generated text from LLM. LLM Large Language Model.

cases (Table 1) in which no scientific literature exists. By combining these individual scientific facts the LLM inferred the unseen synergistic effect. Generally, drugs targeting non-overlapping proteins in similar pathways are more likely to be synergistic^{24,25}. In this case, both AZD4877 and AZD1208 target similar pathways that inhibit tumor cell divisions without overlapping protein targets. The Loewe synergy score of this pair at T24 was 46.82, meaning strong positive synergistic effect.

Example of prediction results

As an example, we listed predicted synergistic drug pairs for stomach and soft tissue using CancerGPT (Supplementary Table 6, 7) and for bone and liver tissue using GPT-3 (Supplementary Table 8, 9). We randomly selected two true positive, false positive, true negative, and false negative prediction examples. We discovered that Loewe synergy scores of the true negative or false negative prediction examples were close to the threshold used to categorize the label (i.e., Loewe score > 5). This suggests that accuracy may vary significantly by different thresholds for determining positive synergy. Setting more extreme thresholds (e.g., >10, >30), like previous models^{13,17,18}, may increase the prediction accuracy.

Discussion

Our study investigates the potential of LLMs as a widely applicable few-shot prediction model in the field of biomedicine. Specifically, we propose a new few-shot model for predicting drug pair synergy, which can be used in rare tissues with few or no training samples available. We transformed tabular data prediction into natural language tasks and fine-tuned LLMs (GPT-2, GPT-3, SciFive) with very few samples in each tissue. Our extensive experiments with seven rare cancer tissues, seven different models, and two different finetuning strategies resulted in the following lessons learned:

- The data-driven models showed superior accuracy when in-distribution external data was available. However, in the absence of external data, or when the external data is out-of-distribution, our customized LLM-based model, CancerGPT, achieved the best accuracy.
- The higher accuracy of CancerGPT over GPT-2 underscores the value of task-specific adjustments. These modifications can augment accuracy while preserving a model's versatility. However, the advantages of such fine-tuning may diminish with larger LLM models like GPT-3.
- CancerGPT, despite its smaller parameter size of 124M, outperformed GPT-3's accuracy, which has a larger parameter size of 175B, suggesting that further fine-tuning of GPT-3 may potentially yield even greater accuracy, provided that in-distribution data is readily available.
- Last layer training, which leaves the LLM's backbone untouched, delivered accuracy comparable to that of a fully trained model. An LLM pre-trained with an extensive collection of scientific literature can

flexibly enhance biomedical predictions using transfer learning based on last-layer training.

The prediction of drug pair synergy in rare tissues serves as an excellent benchmark task for evaluating LLMs in few-shot learning within the field of biomedicine. This prediction requires incorporating multiple pieces of information, such as the drugs and the cell line, as well as the sensitivity of drugs to the cell lines, 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, drug pair synergy scores are obtained through high-throughput testing facilities involving robot arms²⁶. Therefore, individual records of the experiments are rarely recorded in academic literature, decreasing the likelihood of their use as training data for LLMs and avoiding data leakage. Additionally, few studies have been conducted on rare tissues regarding their synergy prediction models, and their synergy prediction outcomes are not explicitly stated in text format. A similar task for evaluating LLMs in biomedicine is predicting the sensitivity of a single drug in a cell line; however, since the sensitivity of individual drugs is extensively researched and well-documented in publications, the LLM model may merely recollect from the text rather than infer unseen tasks.

It should be noted that it was not possible to compare our LLM-based models directly with previous predictions of drug pair synergy. The majority of previous models necessitates high-dimensional features of drugs and cells (e.g., genomic or chemical profiles), along with a substantial amount of training data, even the one specifically designed for rare tissue¹⁸. This kind of data is not easily accessible in rare tissues, making it challenging to carry out a significant comparison. Our model is designed to address a common but often overlooked situation where we have limited features and data. Thus, we compared the LLM-based models with other tabular models that share the same set of inputs.

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 has significant implications for drug development in these cancer types. By accurately predicting which drug pair will have a synergistic effect on these tissues, where cell line features are expensive to obtain, biologists can directly focus on the most probable drug pairs and perform experiments in a cost-effective manner.

Our study also delivers generalizable insights about LLMs in the broader context of biomedicine. Our research essentially bridges two distinct methodologies: data-driven machine learning models, which rely on

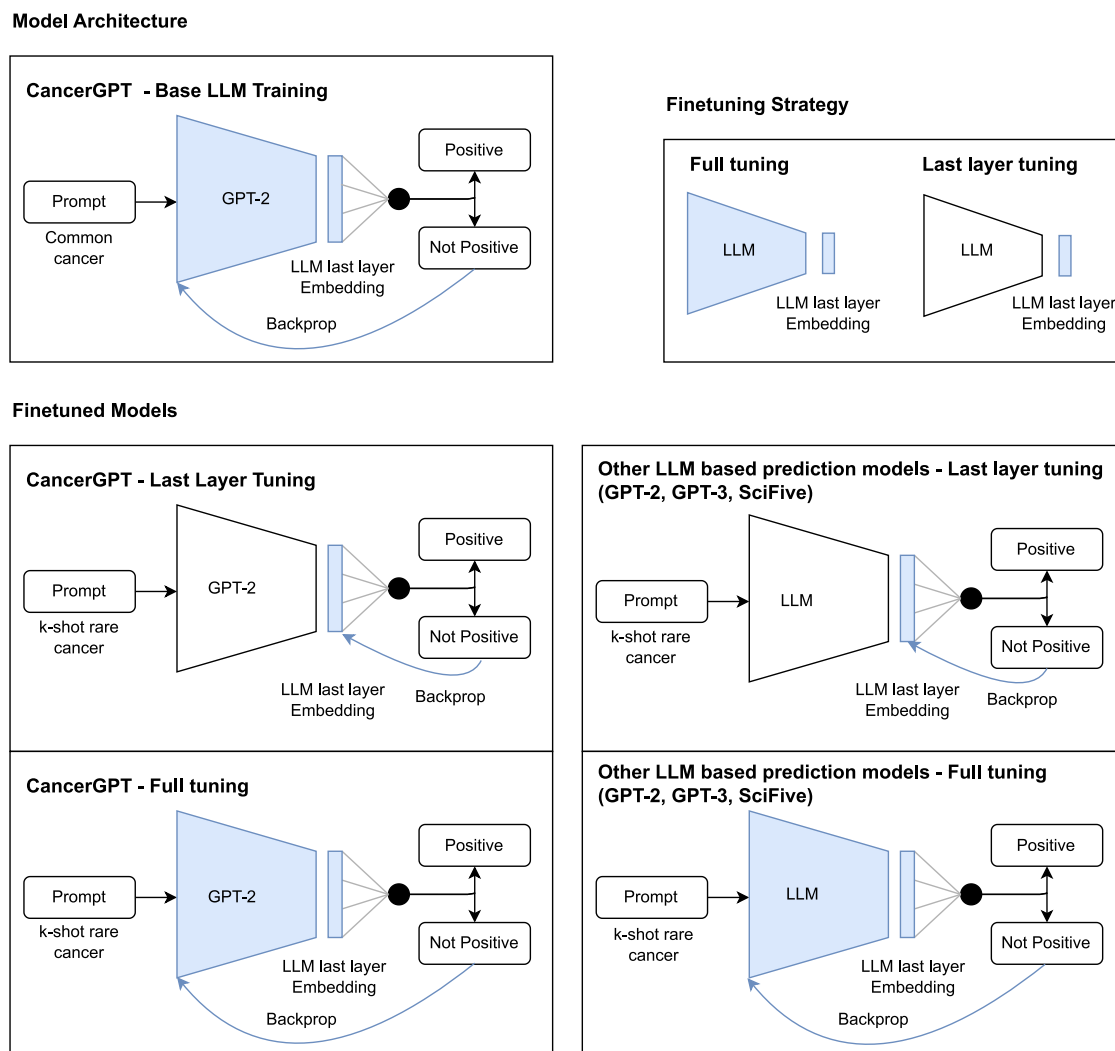


Fig. 5 | Model architecture. We built an LLM-based prediction model by adding a classification head to the LLM and fine-tuned it for the classification task. To obtain representation from LLM, we used GPT-2, GPT-3, and SciFive. We further tailored GPT-2 by fine-tuning it with a large amount of external data, in order to adjust GPT-2 in the context of drug pair synergy prediction (CancerGPT). We finetuned all

models with k shots of data in each of the rare tissues (k -shot fine-tuning strategy). We compared accuracy of two different fine-tuning strategies: fine-tuning the entire model's parameters versus freezing the LLM's parameters and only fine-tuning the classifier's parameters. LLM Large Language Model.

inductive reasoning; and knowledge-driven inference models, which use deductive reasoning. In many cases, especially when dealing with rare diseases like certain types of cancer, the data required to build data-driven models is scarce. This scarcity renders these models virtually ineffective. Conversely, knowledge-driven models pose their own challenges. Without extensive domain expertise, navigating the vast and intricate premises to predict biomedical outcomes is not only challenging, but also unscalable. Our approach combines the strengths of both models. We “automate” deductive reasoning using a LLM, which then forms the foundation for a few-shot prediction model. This combination of inductive (few-shot fine-tuning) and deductive (knowledge encapsulated in LLM) reasoning is a novel concept, made possible through our LLM-based prediction model.

Furthermore, this LLM-based few-shot prediction approach can be applied to a wide range of diseases beyond cancer, which are limited by the scarcity of available data. For instance, this approach can be used in infectious diseases, in which the prompt identification of new treatments and diagnostic tools is crucial. LLMs are able to help researchers quickly identify potential drug targets and biomarkers for these diseases, resulting in faster and more effective treatment development.

The present study, while aiming to showcase the potential of LLMs as a few-shot prediction model in the field of biomedicine, is not without its

limitations. To fully establish the generalizability of LLMs as a “generalist” artificial intelligence, a wider range of biological prediction tasks must be undertaken for validation. Additionally, it is crucial to investigate how the information gleaned from LLMs complements the existing genomic or chemical features that are traditionally the primary source of predictive information. In future research, we plan to investigate this aspect and develop an ensemble method that effectively utilizes both existing structured features and new prior knowledge encoded in LLMs.

Furthermore, while we observed that GPT-3's reasoning was similar to our own when fact-checking its argument with scientific literature in one example, it is important to note that the accuracy of its arguments cannot always be verified and may be susceptible to hallucination. It is also reported that LLMs can also contain biases that humans have²⁷. Therefore, further research is necessary to ensure that the LLM's reasoning is grounded in factual evidence. Despite these limitations, our study provides valuable insights into the potential of LLM to be a few-shot prediction model in biomedicine and lays the groundwork for future research in this area.

In conclusion, our study has effectively demonstrated the potential of LLMs in few-shot learning tasks within the complex field of biomedicine. Our study revealed that while data-driven models excel with sufficient in-distribution data, CancerGPT is superior in scenarios lacking such data. The

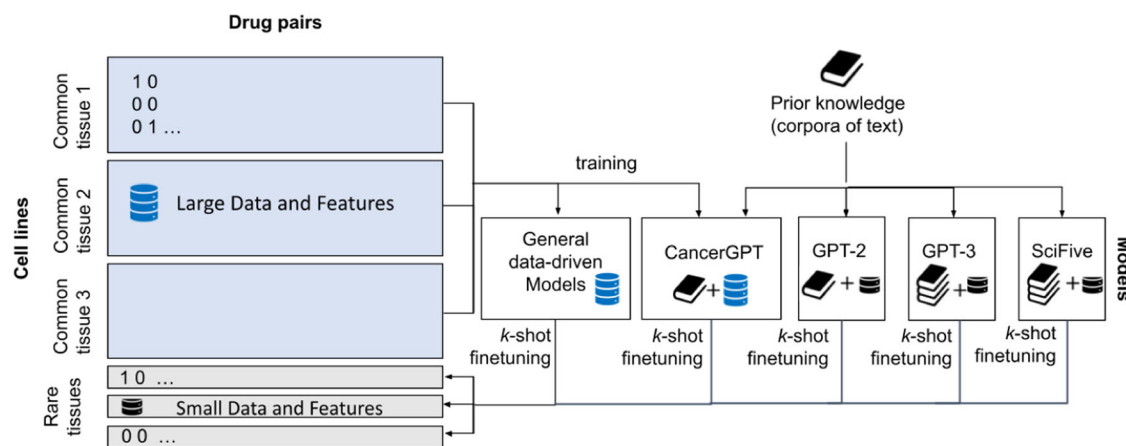


Fig. 6 | Training strategy of baseline and proposed LLM-based models. General data-driven models and CancerGPT were first trained with samples from common tissues (cancer type) then *k*-shot fine-tuned with each tissue of interest. GPT-2, GPT-3 and SciFive are pre-trained models, and we fine-tuned them with *k* shots of data in each tissue.

effectiveness of task-specific fine-tuning was highlighted by the superior performance of CancerGPT over GPT-2, demonstrating that such adjustments can significantly enhance accuracy without compromising the model’s flexibility. Our findings not only are pivotal for the prediction of drug pair synergy in rare tissues but also signify broader applicability for LLMs in biological inference tasks.

Methods

Problem formulation

Our objective is to predict whether a drug pair in a certain cell line has a synergistic effect, particularly focusing on rare tissues with limited training samples. Given an input

$$x = \{d_1, d_2, c, t, ri_1, ri_2\} \tag{1}$$

of drug pair (d_1, d_2), cell line c , tissue t , and the sensitivity of the two drugs using relative inhibition (ri_1, ri_2), the prediction model is

$$y \approx f(x) \tag{2}$$

where y is the binary synergy class (1 if synergy > 5 ; 0 otherwise). Prior research^{14,17} has employed three different scenarios for predicting drug pair synergy (random split, stratified by cell lines, stratified by drug combinations). Our task is to predict synergy when the data are stratified by tissue, which is a subset of cell lines.

Distributions learned in one tissue may not generalizable well to other tissues with different cellular environments²². This biological difference poses a challenge in predicting drug pair’s synergy in tissues with a limited number of samples. The limited sample size of data makes it even more difficult to incorporate typical cell line features, such as gene expression level, which has large dimensionality (e.g., ~ 20,000 genes). Due to this data challenge, the drug pair synergy prediction model is then reduced to build a prediction model with limited samples (few-shot or zero-shot learning) with only limited tabular input feature types. Specific input features were described in Methods.

Prompt selection

To use a LLM for tabular data, the tabular input and prediction task must be transformed into natural text. For each instance of tabular data (Fig. 2), we converted the structured features into text. A prior study¹⁹ investigated different strategies to “serialize” the structured instance to natural language input, such as List Template (a list of column names and feature values), Text Template (a textual enumeration of all features and values), and Table-To-Text (table to text generation via LLM). Among them, we used the Text Template strategy which proved to be most effective. 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^{28,29}.

We created a prompt that specifies our tasks and guides the LLM to generate a label of 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 prompts (Supplementary Note 1), the final prompt we used in this work was “Decide in a single word if the synergy of the drug combination in the cell line is positive or not. {Tabular Input}. Synergy:”.

LLM-based prediction model

We built an LLM-based prediction model by adding a classification head to the LLM and fine-tuned it for the classification task (Fig. 5). This was done to enhance the model’s performance in the classification task. The hidden representation of the final token in the LLM output was utilized, as it encapsulates the information of all preceding tokens. This strategy not only allowed for a fair comparison with other baseline models, such as Tab-Transformer, but also created an opportunity to fine-tune the model specifically for binary classification.

We also considered another modeling approach, which was to allow the LLM to directly generate text for the binary response (“positive”, “not positive”). However, regardless of the LLM’s ability to encapsulate relevant knowledge within their embedding, some LLMs (e.g., GPT-2, SciFive) could not ensure the response strictly adhered to one of the binary labels. This made the inclusion of the classification head an essential choice. The primary objective of our study was to assess the LLM’s biomedical knowledge and its adaptability to a specific task of interest, rather than comparing language generation capabilities.

To obtain representation from LLM, we used GPT-2, GPT-3, and SciFive (Fig. 6). GPT-2 is a Transformer-based large language model, which was pre-trained on a very large corpus of English data without human supervision. It achieved state-of-the-art results on several language modeling datasets in a zero-shot setting when it was released, and it is the predecessor of GPT-3 and GPT-4. GPT-2⁵ has several versions with different sizes of parameters, GPT-2, GPT-Medium, GPT-Large, and GPT-XL. We used GPT-2 with the smallest number of parameters (regular GPT-2, 124 million) in this work to make the model trainable on our server.

GPT-3² is a Transformer-based autoregressive language model with 175 billion parameters, and it achieved state-of-the-art performance on many zero-shot and few-shot tasks upon its release. GPT-3.5, including ChatGPT³⁰, a famous fine-tuned model from GPT-3.5, is an improved version of GPT-3. However, the GPT-3 model and its parameters are not publicly available. Although the weight of the GPT-3 model is undisclosed, OpenAI offers an API³¹ to fine-tune the model and evaluate its performance. We utilized this API to build drug pair synergy prediction models through *k*-shot fine-tuning. There are four models provided by OpenAI for fine-tuning, Davinci, Curie, Babbage, and Ada, of which Ada is the fastest model and has comparable performance with larger models for classification tasks. For that reason, we use GPT-3 Ada as our classification model. After uploading the train data, the API adjusted the learning rate, which is 0.05, 0.1, or 0.2 multiplied by the original learning rate based on the size of the data, and fine-tuned the model for four epochs. A model of the last epoch was provided for further evaluation.

In contrast to general-purpose LLM (i.e., GPT-2 and GPT-3), SciFive²⁰ is an LLM specialized in the biomedical field. Based on T5 model, which is a text-to-text language model trained on the unfiltered CommonCrawl dataset, SciFive re-trains the T5 model on various combinations of the CommonCrawl dataset, a corpus of PubMed abstracts, and a corpus of PubMed Central (PMC) full-text articles. We used the SciFive Pubmed Base model with 220M parameters.

CancerGPT

We further tailored GPT-2 by fine-tuning it with a large amount of external data (common cancer data), in order to adjust GPT-2 in the context of drug pair synergy prediction. We named this model CancerGPT (Fig. 5). The data from common tissues (Data Availability) are external data that do not necessarily align with the context of the rare tissue being studied. However, this data could potentially serve as a useful data source to “warm up” the LLM, thereby enhancing its ability to predict drug pair synergy. The intuition behind this “warming up” strategy is analogous to collaborative filtering³² that learn relational information and extrapolate it to the new combination of input. Certain drug pairs exhibit synergy similarly regardless of the cellular context, and therefore, the relational information between drug pairs in common tissues can be used to predict synergy in new cell lines in different tissues¹⁷. Subsequently, we utilized CancerGPT as one of the pre-trained LLMs and fine-tuned to *k* shots of data in each rare tissue (as discussed in the following section).

General data-driven models

The LLM-based prediction are knowledge-driven models as they utilize prior knowledge encoded in pre-trained weight. We compared them with general data-driven prediction model. We specifically used XGBoost³³, TabTransformer³⁴, and collaborative filtering¹⁸. XGBoost has been widely used in large-scale drug synergy data^{35,36}. All the variables (drugs, cell lines, and sensitivities) were used as input to predict the drug pair synergy. TabTransformer is a self-attention-based supervised learning model for tabular data. TabTransformer applies a sequence of multi-head attention-based Transformer layers on parametric embedding to transform them into contextual embedding, in which highly correlated features will be close to each other in the embedding space. Considering the highly correlated nature of drugs in our data, TabTransformer can be a very strong baseline in this work. We first converted the drugs and cell lines in the tabular data into indicators using one-hot coding. We first trained an embedding layer on the drugs and cell lines and passed them through stacked multi-headed attention layers, which we then combined with the continuous variables (sensitivities). This combination then passes through feed-forward layers, which have one layer of classification head. Collaborative filtering was used to extrapolate the relational information to the new combination of drugs and cell line. Drug combinations will react to a cell line similarly if these two drugs have responded similarly to other cell line¹⁸. Note that, tissue information was not used in training because the models will be tested in one

specific rare tissue that is not used in training. No further contextual information can be inferred through the unseen tissue indicator.

k-shot fine-tuning strategy

The LLM-based models had different training and fine-tuning strategies (Fig. 6). Samples of common tissues were split into 80% train data and 20% validation data for CancerGPT. The models were trained using train data and evaluated by validation data to determine the models with specific hyperparameters to be used for further fine-tuning on rare tissues. For the GPT-2 and GPT-3 based prediction models, we directly used pre-trained parameters from GPT-2² using Huggingface’s Transformers library³⁷ and GPT-3 Ada from OpenAI² respectively.

All these models were then fine-tuned with *k* shots of data in each of the rare tissues. For bone, urinary tract, stomach, soft tissue, and liver, we performed experiments with *k* from [0, 2, 4, 8, 16, 32, 64, 128]. For endometrium and pancreas, because of the limited number of data, we implemented experiments with *k* from [0, 2, 4, 8, 16, 32] from the endometrium, and only zero shot (*k* = 0) for the pancreas.

With the limited number of shots, a careful balance of binary labels in the train and test set was critical. We partitioned the data into 80% for training and 20% for testing in each rare tissue, while ensuring the binary labels were equally represented in both sets. We randomly selected *k* shots from the training for fine-tuning, while maintaining consistency with previously selected shots and adding new ones. Specifically, we maintained the previously selected *k* shots in the training set and incremented additional *k* shots to create $2 \times k$ shots. The binary label distribution in each *k* shot set followed that of the original data, with at least one positive and one negative sample included in each set. For evaluation stability, the test data was consistent across different shots for each tissue.

To investigate whether the LLMs inherently possesses the capability to predict the label only with minimal tuning, we compared accuracy of two different fine-tuning strategies: fine-tuning the entire model (LLM’s parameter and classifier) versus freezing the LLM’s parameter and fine-tuning only on the classifier. Note that for zero shot tuning, we measured accuracy by allowing the LLM to directly generate a binary answer without appending a classification head because we are unable to optimize the parameters in the classification head.

Hyperparameter setting

The LLM’s last layer hidden representation size was 768 (same as GPT-2), and we used left padding to ensure that the last token was from the prompt sentence. The cross-entropy loss was used to optimize the model during the fine-tuning process. All the LLM models used the tabular input that was converted to natural text and shared the same prompt.

The predicted output was a binary label indicating the presence of a synergistic effect, with a Loewe score greater than 5 indicating a positive result. We used AUROC and AUPRC to evaluate the accuracy of classification. Regression tasks were not possible in our LLM-based models because our model can only generate text-based answers (“positive” or “not positive”), with poor precision in accurately quantifying the synergy value.

XGBoost was used with a boosting learning rate of 0.3. The number of gradient boost trees was set to 1000 with a maximum tree depth of 20 for base learners. TabTransformer was used with a learning rate of 0.0001 and a weight decay of 0.01. The model was trained for 50 epochs on common tissues. During the training, the model with the best validation performance was selected for further fine-tuned on rare tissues. For each *k* shot in each tissue, the model was fine-tuned using the same learning rate and weight decay for 1 epoch and tested with AUPRC and AUROC.

CancerGPT was first fine-tuned with pre-trained regular GPT-2 for 4 epochs on common tissues. The learning rate was set to be $5e-5$ and weight decay was set to be 0.01. Then the model was fine-tuned for *k* shots in rare tissues. The same hyperparameters are used in training. The model was finally tested with AUPRC and AUROC. Note that, due to an imbalance in positive and non-positive labels, we reported both AUPRC and AUROC.

GPT-2 and GPT-3 are directly fine-tuned on rare tissues with pre-trained parameters from regular GPT-2 and GPT-3 Ada. For each k shot in each tissue, GPT-2 is fine-tuned for 4 epochs using a learning rate of $5e-5$ and a weight decay of 0.01. The hyperparameters of GPT-3 are adjusted by OpenAI API based on the data size. The model was also fine-tuned for 4 epochs. GPT-2, GPT-3 and SciFive fine-tuned models were finally tested with AUPRC and AUROC. Details in the hyperparameter setting are discussed in Supplementary Note 2.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

We utilized a publicly accessible extensive database of drug synergy from DrugComb Portal⁹, which is an open-access data portal where the results of drug combination screening studies for a large variety of cancer cell lines are accumulated, standardized, and harmonized. The database contains both drug sensitivity rows and drug pair synergy rows. After filtering the available drug pair synergy rows, the data contains 4226 unique drugs, 288 cell lines, with a total of 718,002 drug pair synergy rows. We employed the Loewe synergy score, which ranges from -100 (antagonistic effect) to 75 (strong synergistic effect), for drug combination synergy³⁸. The Loewe synergy score quantifies the excess over the expected response if the two drugs are the same compound^{39,40}. In this paper, we focused on cell lines from rare tissues. We defined the rare tissues as the ones with less than 4000 samples, which include the pancreas ($n = 39$), endometrium ($n = 68$), liver ($n = 213$), soft tissue ($n = 352$), stomach ($n = 1190$), urinary tract ($n = 2458$), and bone ($n = 3985$). We tested our models with each of the rare tissues. The remaining tissues, namely hematopoietic and lymphoid, lung, skin, ovary, kidney, colon, brain, breast, and prostate were regarded as common tissues.

Code availability

All models used in this paper were developed using open-sourced libraries such as PyTorch⁴¹, HuggingFace³⁷ and publically available APIs from OpenAI³¹. Method-specific code will be provided upon request.

Received: 19 May 2023; Accepted: 2 February 2024;

Published online: 19 February 2024

References

- Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
- Brown, T.B. et al. Language Models are Few-Shot Learners. Preprint at <https://arxiv.org/abs/2005.14165> (2020).
- OpenAI: GPT-4 Technical Report. Preprint at <https://arxiv.org/abs/2303.08774> (2023).
- Mitchell, M. & Krakauer, D. C. The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci.* **120**, 2215907120 (2023).
- Radford, A. et al. Language Models are Unsupervised Multitask Learners. Preprint at <https://d4mucfpksyww.cloudfront.net/better-language-models/language-models.pdf> (2018).
- Veit, A. et al. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 839–847 (2017).
- Wertheimer, H. Few-Shot learning with localization in realistic settings. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), vol. 0, pp. 6551–6560 (2019).
- Sun, W., Sanderson, P. E. & Zheng, W. Drug combination therapy increases successful drug repositioning. *Drug Discov. Today* **21**, 1189–1195 (2016).
- Zagidullin, B. et al. DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res.* **47**, 43–51 (2019).
- Chen, H., & Li, J. DrugCom: Synergistic Discovery of Drug Combinations Using Tensor Decomposition. In *2018 IEEE International Conference on Data Mining (ICDM)*, 899–904 (2018).
- Sun, Z., Huang, S., Jiang, P. & Hu, P. DTF: Deep tensor factorization for predicting anticancer drug synergy. *Bioinformatics* **36**, 4483–4489 (2020).
- Li, H., Li, T., Quang, D. & Guan, Y. Network propagation predicts drug synergy in cancers. *Cancer Res.* **78**, 5446–5457 (2018).
- Kuru, H. I., Tastan, O. & Cicek, A. E. MatchMaker: A deep learning framework for drug synergy prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 2334–2344 (2022).
- Liu, X. et al. Multi-way relation-enhanced hypergraph representation learning for anti-cancer drug synergy prediction. *Bioinformatics* **38**, 4782–4789 (2022).
- Preuer, K. et al. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* **34**, 1538–1546 (2018).
- Liu, Q. & Xie, L. TranSynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations. *PLoS Comput. Biol.* **17**, 1008653 (2021).
- Hosseini, S.-R. & Zhou, X. CCSynergy: an integrative deep-learning framework enabling context-aware prediction of anti-cancer drug synergy. *Brief. Bioinform.* **24**, bbac588 (2023).
- Kim, Y. et al. Anticancer drug synergy prediction in understudied tissues using transfer learning. *J. Am. Med. Inform. Assoc.* **28**, 42–51 (2021).
- Hegselmann, S. et al. TabLLM: Few-shot classification of tabular data with large language models. *AISTATS abs/2210.10723*, (2022).
- Phan, L. N. et al. SciFive: a text-to-text transformer model for biomedical literature. Preprint at <https://arxiv.org/abs/2106.03598> (2021).
- Guo, L. et al. Similarities and differences in the expression of drug-metabolizing enzymes between human hepatic cell lines and primary human hepatocytes. *Drug Metab. Dispos.* **39**, 528–538 (2011).
- Warren, A. et al. Global computational alignment of tumor and cell line transcriptional profiles. *Nature Communications* **12**, 22 (2021).
- Deng, J. et al. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009).
- Cheng, F., Kovács, I. A. & Barabási, A.-L. Network-based prediction of drug combinations. *Nat. Commun.* **10**, 1197 (2019).
- Tang, Y.-C. & Gottlieb, A. SynPathy: Predicting drug synergy through Drug-Associated pathways using deep learning. *Mol. Cancer Res.* **20**, 762–769 (2022).
- He, L. et al. Methods for high-throughput drug combination screening and synergy scoring. In: *Cancer Systems Biology: Methods and Protocols*, pp. 351–398 (2018).
- Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. & Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **4**, 258–268 (2022).
- Li, Y., Li, J., Suhara, Y., Doan, A. & Tan, W.-C. Deep entity matching with pre-trained language models. *Proc. VLDB Endowment* **14**, 50–60 (2020).
- Narayan, A., Chami, I., Orr, L. & Ré, C. Can Foundation Models Wrangle Your Data? *Proc VLDB Endow* **16**, 738–746 (2022).
- OpenAI. Introducing ChatGPT. <https://openai.com/blog/chatgpt> (2022).
- OpenAI. Fine-Tuning - OpenAI API. <https://platform.openai.com/docs/guides/fine-tuning> (2021).
- Suphavitai, C., Bertrand, D. & Nagarajan, N. Predicting cancer drug response using a recommender system. *Bioinformatics* **34**, 3907–3914 (2018).
- Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016).

34. Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. Preprint at <https://arxiv.org/abs/2012.06678> (2020).
35. Sidorov, P., Naulaerts, S., Ariey-Bonnet, J., Pasquier, E. & Ballester, P. J. Predicting synergism of cancer drug combinations using NCI-ALMANAC data. *Front Chem.* **7**, 509 (2019).
36. Celebi, R., Bear Don't Walk IV, O., Movva, R., Alpsy, S. & Dumontier, M. In-silico prediction of synergistic Anti-Cancer drug combinations using multi-omics data. *Sci. Rep.* **9**, 8949 (2019).
37. Wolf, T. et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (2020).
38. Greco, W. R., Bravo, G. & Parsons, J. C. The search for synergy: a critical review from a response surface perspective. *Pharmacol. Rev.* **47**, 331–385 (1995).
39. Ianevski, A., He, L., Aittokallio, T. & Tang, J. SynergyFinder: a web application for analyzing drug combination dose-response matrix data. *Bioinformatics* **33**, 2413–2415 (2017).
40. Yadav, B., Wennerberg, K., Aittokallio, T. & Tang, J. Searching for Drug Synergy in Complex Dose-Response Landscapes Using an Interaction Potency Model. *Comput. Struct. Biotechnol. J.* **13**, 504–513 (2015).
41. Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Preprint at <https://arxiv.org/abs/1912.01703> (2019).
42. Guo, S. et al. Overexpression of pim-1 in bladder cancer. *J. Exp. Clin. Cancer Res.* **29**, 161 (2010).
43. Jones, R. et al. Phase II study to assess the efficacy, safety and tolerability of the mitotic spindle kinesin inhibitor AZD4877 in patients with recurrent advanced urothelial cancer. *Invest New Drugs* **31**, 1001–1007 (2013).
44. Cortes, J. et al. Phase I studies of AZD1208, a proviral integration moloney virus kinase inhibitor in solid and haematological cancers. *Br. J. Cancer* **118**, 1425–1433 (2018).
45. National Cancer Institute. NCI Drug Dictionary. <https://www.cancer.gov/publications/dictionaries/cancer-drug/def/pan-pim-kinase-inhibitor-azd1208> (2011).
46. Cervantes-Gomez, F. et al. PIM kinase inhibitor, AZD1208, inhibits protein translation and induces autophagy in primary chronic lymphocytic leukemia cells. *Oncotarget* **10**, 2793–2809 (2019).
47. Borthakur, G. et al. Clinical, pharmacokinetic (PK), and pharmacodynamic findings from a phase I trial of an eg5 inhibitor (AZD4877) in patients with refractory acute myeloid leukemia (AML). *J. Clin. Orthodont.* **27**, 3580–3580 (2009).

Acknowledgements

XJ is CPRIT Scholar in Cancer Research (RR180012), and he was supported in part by Christopher Sarofim Family Professorship, UT Stars award, UTHealth startup, the National Institute of Health (NIH) under award number

R01AG066749, R01AG066749-03S1, R01LM013712 and U01TR002062, and the National Science Foundation (NSF) #2124789. YD is supported by NSF AI Center Grant (NSF 2019844), NSF-CSIRO (NSF 2303038), NIH Bridge2AI (OTA-21-008), and Bill & Lewis Suit Professorship at the University of Texas at Austin. YK is supported by the NIH under award number R01AG082721, R01AG066749-03S1, and CPRIT RR180012.

Author contributions

The first two authors contributed equally. X.J. and Y.D. provided the motivation of the study. T.L. and S.S. contributed in the development of models, hyperparameter tuning, data processing and manuscript writing. A.K. collected chemical features from the various drugs. Y.K. led in writing of the manuscript and study design, and is the corresponding author. A.J., X.J., Y.D. and Y.K. provided feedback during the study and for the manuscript. Y.K. and Y.D. provided supervision and led design study during the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains

supplementary material available at <https://doi.org/10.1038/s41746-024-01024-9>.

Correspondence and requests for materials should be addressed to Yejin Kim.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024