





Large Language Models for Translational Cancer Informatics

Yining Pan, MS¹; Yanfei Wang, PhD¹; Guangyu Wang, PhD^{2,3} ; Jing Su, PhD⁴ ; Umit Topaloglu, PhD⁵ ; and Qianqian Song, PhD¹ 

DOI <https://doi.org/10.1200/CCI-25-00108>

ABSTRACT

Accepted July 29, 2025
Published October 14, 2025

JCO Clin Cancer Inform
9:e2500108
© 2025 by American Society of
Clinical Oncology

PURPOSE Cancer remains a leading cause of death worldwide. The growing volume of high-throughput single-cell and spatial transcriptomic data sets—particularly those related to cancer—offers immense opportunities as well as analytical challenges for effective data analysis and interpretation. Large language models (LLMs), pretrained on vast data sets and capable of various biomedical tasks, offer a promising solution. This review explores the application of LLMs in cancer research from both cellular and pathologic perspectives, aiming to showcase their potential in advancing precision oncology.

MATERIALS AND METHODS We systematically review current LLMs in analyzing single-cell RNA sequencing, spatial transcriptomic, and histology image data, emphasizing their relevance to cancer biology and translational research.

RESULTS A total of 24 LLMs, published or in preprint between 2022 and 2025, were selected for review. In single-cell transcriptomics, LLMs have primarily been used for cell type annotation, batch integration, and drug-response prediction. In spatial transcriptomics, LLMs support multislide and multimodal spatial data integration, gene expression imputation, niche and region label prediction, spatial domain identification, cell-cell communication inference, and marker gene detection. In computational pathology, LLMs have been applied to cancer subtyping, detection of rare malignancies, genomic mutation prediction, image segmentation, as well as cross-modal retrieval. Despite these advances, many models remain underoptimized for cancer-specific applications, highlighting the need for domain-specific fine-tuning and scalable adaptation strategies.

CONCLUSION LLMs have the potential to significantly advance cancer research by providing scalable and effective tools for analyzing and interpreting single-cell, spatial transcriptomic, and pathology data. Future efforts should prioritize tailoring these models to cancer-specific contexts to enhance their utility in uncovering disease mechanisms, identifying biomarkers, and informing therapeutic strategies.

INTRODUCTION

Cancer remains one of the leading cause of death worldwide.¹ In recent years, the increasing accessibility of single-cell, spatial omics, and pathology images have facilitated cancer research, offering insights into cellular diversity, tumor metastasis, and drug resistance.^{2,3} The development of large-scale resource—as exemplified by Human BioMolecular Atlas Program (HuBMAP) for single-cell and spatial data,⁴ the Cancer Image Archive for hematoxylin and eosin whole-slide cancer pathology images,⁵ and HEST-1k for paired spatial transcriptomics and histology images⁶—necessitates the use of advanced computational tools for effective data integration, analysis, and interpretation. This explosion of complex, high-dimensional data creates an ideal landscape for the application of large language models

(LLMs), which are designed to learn from massive and heterogeneous data sets.^{7–13} LLMs, typically based on transformer architectures, follow a two-stage training process: pretraining on large-scale unlabeled corpora, followed by fine-tuning on domain-specific tasks.¹⁴ This approach enables them to generalize across a wide range of applications while retaining the flexibility to specialize, making them well suited to address the multifaceted challenges of biomedical research.^{15–17}

In single-cell transcriptomics, LLMs have enabled characterization of cancer cell subpopulations,¹⁰ modeling of drug response mechanisms,⁷ as well as reconstructing cell lineage trajectories.¹⁸ In spatial transcriptomics, LLMs facilitate cell type deconvolution,¹⁹ mapping of gene expression at high spatial resolution within tissue samples,¹⁹ and ligand-

receptor interactions inference.²⁰ Furthermore, in computational pathology, LLMs have emerged as powerful tools for analyzing digital pathology images, aiding in cancer subtyping,²¹ genomic mutation prediction,²² and assessment of tumor resectability.¹² These advancements highlight the remarkable capabilities of LLMs in analyzing vast and complex high-throughput omics data and histopathology imaging data to transform cancer research. Therefore, there is a need for a focused and integrative review of LLM applications specifically in cancer research, encompassing both molecular and pathologic perspectives. To address this growing interest, we examined current state-of-the-art LLMs that leverage single-cell and spatial transcriptomics, as well as pathology imaging data, encouraging further integration of LLMs into translational cancer research and clinical applications (Fig 1).

LLMs FOR SINGLE-CELL APPLICATIONS IN CANCER

The development of large-scale single-cell and single-nucleus molecular profiling data sets—such as those generated by the Human Cell Atlas (HCA) consortium,²³ HuBMAP,⁴ Tabula Sapiens,²⁴ and PanglaoDB²⁵—has provided a valuable foundation for training advanced models, including single-cell LLMs.²³ These data sets serve as ground-truth references and have enabled the development of LLMs with superior capabilities across a range of cancer biology applications, including cell type annotation, multitbatch and multiomics integration, and drug-response prediction (Tables 1 and 2).

A fundamental task in single-cell analysis is the annotation of cell types, which traditionally requires substantial time and expert knowledge.⁵⁶ Single-cell foundation models

(FMs), such as single-cell bidirectional encoder representations from transformers (scBERT),¹⁰ have emerged as powerful tools in automating this process. scBERT is a deep neural model trained on large volumes of unlabeled single-cell RNA sequencing (scRNA-seq) data and learns gene-gene interactions via transformer-based encoder representations. When benchmarked across data sets encompassing over 500,000 cells from diverse organs, tissues, and single-cell modalities, scBERT consistently outperformed conventional methods—including marker-gene-based tools, correlation-based methods (eg, Seurat v4⁵⁷), and machine learning (ML)-based models (eg, SciBet¹⁰). Notably, scBERT could accurately distinguish between closely related immune cell populations, such as CD8+ cytotoxic T cells, CD19+ B cells, CD34+ progenitors, and CD8+/CD45RA+ naïve cytotoxic cells—cell types often implicated in cancer progression.¹⁰ Notably, in the human brain—known for its intricate cellular architecture⁵⁸—a novel Transformer-based model, single-cell Hyena (scHyena), excels in cell type classification using full-length scRNA-seq data.²⁸ Trained on full-length scRNA-seq data, scHyena reached high accuracy in classifying key brain cell types (eg, astrocytes, microglia, oligodendrocyte progenitor cells, excitatory/inhibitory neurons, endothelial cells, and pericytes) across multiple data sets.²⁸ Extending beyond specific tissues, the Universal Cell Embedding (UCE) model provides a unified biological latent space, in which cells from any tissue or species can be embedded without additional training.²⁹ This allows for robust cross-data set and cross-species comparisons.²⁹ For instance, UCE can accurately transfer human lymph node annotations from human lymph node samples to an unseen data set from the mediastinal lymph node of a green monkey,²⁹ highlighting its potential for identifying conserved oncogenic mechanisms. Similarly, SCsimilarity, a

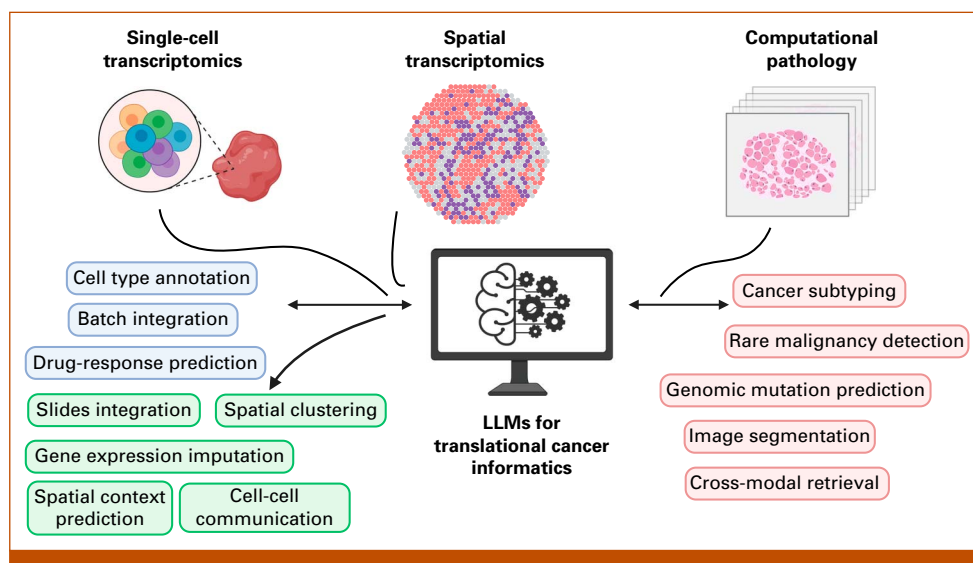


FIG 1. Application of LLMs in single-cell and spatial transcriptomics, and computational pathology. LLM, large language model.

TABLE 1. An Overview of Single-Cell, Spatial, and Pathology LLM Applications

Application Type	Ref	Year	Model	Goal and Aim	Pretaining Data Sets	Input Data	Parameter Size	Tokenization	Transformer Architecture	Model Derived	Cancer Research Application	Cancer Application Area
Single-cell LLMs	Yang et al ¹⁰	2022	scBERT	Understand gene-gene interactions, cell type annotation, batch-effect removal	16,000 gene profiles from Pan-gliadb comprising 209 human scRNA-seq data with 1,125,580 cells	Single-cell transcriptome	10M	Binned gene expression values	Encoder only	Pretrain and fine-tune	No	NA
	Zhao et al ¹⁵	2023	CellLM	Cell type annotation, drug sensitivity prediction	19,379 genes from 2 million scRNA-seq data	Single-cell transcriptome	50M	Binned gene expression values	BERT + contrastive learning	Pretrain and fine-tune	Yes	Incorporated cancer data in pretraining to distinguish tumor cells from healthy cells
	Wen et al ¹⁷	2023	CellPLM	Cell clustering, cell type annotation, perturbation prediction	9 million scRNA-seq cells	Single-cell transcriptome	80M	Continuous gene expression values	Transformer	Pretrain and fine-tune	No	NA
	Theodoris et al ⁹	2023	Geneformer	Predict chromatin and network dynamics	30 million human single-cell transcriptomes, Genecorpus-30M	Single-cell transcriptome	30M	Ranked gene expression values	Encoder only	Pretrain and fine-tune	No	NA
	Oh et al ¹⁸	2023	scHyena	Cell type annotation, scRNA-seq imputation for brain research	19,205 genes from two external brain scRNA-seq data sets	Single-cell transcriptome	Unknown	Continuous gene expression values	Transformer	Pretrain and fine-tune	No	NA
	Shen et al ¹⁶	2023	tgPT	Cell type annotation, cell lineage trajectory identification	22.3 million single-cell transcriptomes	Bulk and single-cell transcriptome	Unknown	Ranked genes	Decoder-only transformer	Pretrain and fine-tune	Yes	Applied to bulk cancer tissue sequencing data, the features captured by tgPT effectively supported cell type annotation, cancer subtyping, prognosis prediction, and immunotherapy outcome assessment
	Rosen et al ¹⁴	2023	UCE	Cell type annotation for multiple species	More than 300 data sets consisting of 126 over 36 million cells from human and other species downloaded from the CELLXGENE corpus	Single-cell transcriptome	650M	Continuous gene expression values	Transformer	Pretrain and fine-tune	No	NA
	Hao et al ¹⁷	2024	scFoundation	Drug response prediction, single-cell perturbation prediction, cell type annotation, gene expression enhancement, and gene module inference	50 million human single-cell transcriptomic data from various databases	Bulk and single-cell transcriptome	100M	Continuous gene expression values	Transformer	Pretrain and fine-tune	Yes	Incorporated single cells of diverse cancer types for pretraining. The model has been applied to cancer drug response prediction
	Cui et al ²⁰	2024	scGPT	Cell type annotation, multibatch integration, multomic integration, perturbation response prediction, gene network inference	33 million cells from CELLXGENE portal	Single-cell transcriptome	50M	Binned gene expression values	Encoder only	Pretrain and fine-tune	Yes	Fine-tuned to distinguish cancer types and perform immune cell annotation in a myeloid data set, as well as to identify cancer, tissue, and cell types across pan-cancer data sets
	Heimberg et al ⁸	2024	SCimilarity	Identify similar cell profiles across diverse diseases and tissues	28,231 genes from 23.4-million-cell atlas of 412 scRNA-seq data	Single-cell transcriptome	NA	Continuous gene expression values	Transformer	Pretrain	Yes	Applied on cancer data sets to identify similar cells across different cancers, such as fibrosis-associated macrophage in uveal melanoma, PDAC, and colon cancer
	Bian et al ¹¹	2024	scMulan	Cell type annotation, batch integration	10 million single-cell transcriptomic data and their corresponding metadata	Single-cell transcriptome	368M	Customized cell sentence	Decoder-only transformer	Pretrain and fine-tune	No	NA

(continued on following page)

TABLE 1. An Overview of Single-Cell, Spatial, and Pathology LLM Applications (continued)

Application Type	Ref	Year	Model	Goal and Aim	Pretraining Data Sets	Input Data	Parameter Size	Tokenization	Transformer Architecture	Model Derived	Cancer Research Application	Cancer Application Area
Spatial transcriptomics LLMs	Wen et al ²⁷	2023	CellPLM	Understand cell-cell interactions and impute gene expressions	2 million spatial transcriptomic cells	Spatial transcriptome	80M	Continuous gene expression values	Transformer	Pretrain and fine-tune	Yes	Applied on lung cancer and liver cancer data sets for spatial transcriptomic gene imputation task
	Liu et al ³²	2024	Geneverse	Identify marker genes for cell types	Spatial transcriptomics data from an external study on human breast tissue	Gene names	7B/8B	BPE	Decoder-only transformer; cross-attention transformer; transformer	Fine-tune	No	NA
	Schaar et al ³³	2024	Nicheformer	Niche and region label prediction, spatial density prediction	57 million dissociated and 53 million spatially resolved cells across 73 tissues from both human and mouse	Single-cell and spatial transcriptome	49.3M	Ranked gene expression values	Transformer	Pretrain and fine-tune	Yes	Incorporated various cancer samples in pretraining. It can predict neighborhood density in Xenium data from lung and colon cancers
	Blampey et al ³⁴	2024	Novae	Integrate data from different platforms and gene panels, spatial domain identification	78 slides comprising 30 million cells across 18 tissues	Spatial transcriptome	Unknown	Graph-based representation	Self-supervised graph attention network	Pretrain and fine-tune	Yes	It performed spatial domain assignment on breast samples and identified domain expansion associated with clonal proliferation of cancer cells
	Wang et al ¹⁹	2025	scGPT-spatial	Multimodal and multislide integration, cell type deconvolution, missing gene imputation	30 million spatial transcriptomic profiles	Spatial transcriptome	Unknown	Binned gene expression values	Transformer Encoder and MoE-based decoder	Pretrain and fine-tune	No	NA
	Ji et al ²⁰	2024	spaCCC	Infer cell-cell communications	OmniPath v1.0.4, STRING v12.0, and CellChat database were used for previous knowledge of ligand-receptor interactions	Spatial transcriptome	50M	Binned gene expression values	Encoder only	Pretrain and fine-tune	Yes	Inferred cell-cell interactions in human RCC and human breast cancer Visium data sets
	Lin et al ³⁵	2024	ST-Align	Spatial cluster identification and gene prediction	1.3 million image-gene pairs from 573 human tissue slides	Spatial transcriptome and histopathology images	50M	Binned gene expression values	Adaptive encoder and attention-based fusion network	Pretrain and fine-tune	No	NA
Pathology LLMs	Zhang et al ³⁶	2024	BiomedGPT	Visual question answering and image captioning	200k vision and language data sets, 46k object-detection data sets, and 600k masked image modeling data sets	Pathology and radiology image and text	33M/93M/182M	Pix2Seq, BPE, and VQGAN	Transformer	Pretrain and fine-tune	Yes	Performed cancer tissue identification using PathMNIST data set
	Wang et al ³⁷	2024	CHIEF	Cancer classification and survival prediction	60,530 WSIs from 14 cohorts	Histopathologic image and text	Unknown	BPE and tile encoding	Vision and language transformer	Pretrain and fine-tune	Yes	It is a foundation model for cancer diagnosis and prognosis prediction. It performs systematic cancer evaluation such as cancer cell detection, tumor origin identification, and prognostic prediction
	Lu et al ¹³	2024	CONCH	Cancer subtyping, cross-modal retrieval, WSI segmentation	21,442 WSIs spanning over 350 cancer subtypes	Histopathologic image and text	200M	BPE and tile encoding	Vision and language transformer	Pretrain and fine-tune	Yes	It was evaluated on slide-level cancer subtyping and ROI-level tissue classification tasks
	Xu et al ²²	2024	Prov-GigaPath	Cancer subtyping, mutation prediction, vision-language alignment	1.3 billion 256 × 256 pathology image tiles in 171,189 whole slides	Histopathologic image	23M	Patch and tile encoding	LongNet	Pretrain and fine-tune	Yes	The model was pretrained on cancer pathology images from Providence. It was evaluated on nine cancer subtyping tasks using Providence and TCGA data
	Ding et al ²¹	2024	TITAN	Rare cancer detection, cross-modal retrieval, pathology report generation	335,645 WSIs	Histopathologic image	100M	BPE, patch and tile encoding	Vision and language transformer	Pretrain and fine-tune	Yes	It was evaluated on cancer subtyping, rare cancer retrieval, and cancer prognosis
	Chen et al ³⁸	2024	UNI	Cancer subtyping, ROI retrieval, ROI cell type segmentation	Over 100,000 diagnostic H&E-stained WSIs across 20 major tissue types	Histopathologic image	300M	Patch encoding	Vision transformer	Pretrain and fine-tune	Yes	The model was demonstrated to be capable of classifying 108 cancer types

Abbreviations: BERT, bidirectional encoder representations from transformers; BPE, byte pair encoding; CellLM, single-cell language model; CellPLM, single-cell pretrained language model; CHIEF, Clinical Histopathology Imaging Evaluation Foundation; CONCH, contrastive learning from captions for histopathology; H&E, hematoxylin and eosin; LLM, large language model; MoE, Mixture-of-Experts; NA, not applicable; PDAC, pancreatic ductal adenocarcinoma; RCC, renal cell carcinoma; Ref, reference; ROI, region of interest; scBERT, single-cell bidirectional encoder representations from transformers; scGPT, single-cell generative pretrained transformer; scHyena, single-cell Hyena; scMulan, single-cell multitask generative pretrained language model; scRNA-seq, single-cell RNA sequencing; spaCCC, spatial transcriptomics cell-cell communications; TCGA, The Cancer Genome Atlas; UCE, Universal Cell Embedding; VQGAN, vector quantized generative adversarial network; WSI, whole-slide image.

TABLE 2. LLM Applications and Performance Summary: Tasks, Data Sets, and Evaluation Metrics³⁹⁻⁵⁵

Model	Tasks	Data Sets	Metrics
scBERT	Cell type annotation	Nine scRNA-seq data sets covering 17 major organs/tissues, more than 50 cell types, over 500,000 cells, and mainstream single-cell omics technologies (Drop-seq, 10×, SMART-seq, and Sanger-Nuclei)	F1-score, 0.691; accuracy, 0.759
	Novel cell type discovery	Nine scRNA-seq data sets covering 17 major organs/tissues, more than 50 cell types, over 500,000 cells, and mainstream single-cell omics technologies (Drop-seq, 10×, SMART-seq, and Sanger-Nuclei)	Accuracy, 0.329
CellLM	Cell type annotation	Human PBMC data set Zheng68k and pancreas data set Baron	Mean F1, 0.889
	Single-cell drug sensitivity prediction	Human lung cancer cells (GSE149383) and human oral squamous cancer cells (GSE117872)	Mean F1, 0.934
	Single-omics cell line drug sensitivity prediction	CCLE and GDSC	Pearson correlation coefficient, 0.934
CellPLM	scRNA-seq denoising	PBMC 5k and Jurkat from 10× genomics	Fine-tuned CellPLM: RMSE, 0.725 ± 0.001; MAE, 0.551 ± 0.001
	Cell type annotation	hPancreas ³⁹ and MS ⁴⁰	Fine-tuned CellPLM: F1, 0.766 ± 0.007 in MS; F1, 0.749 ± 0.010 in hPancreas
Geneformer	Gene dosage sensitivity predictions	Collins et al ⁴¹	AUC, 0.91
	Chromatin dynamics predictions	PanglaoDB	AUC 0.93 and 0.88; bivalent versus unmethylated or H3K4me3-only, respectively
	Network dynamics predictions	30,000 normal ECs from the Heart Atlas	AUC, 0.81
scHyena	Cell type annotation	Lau, ⁴² Leng, ⁴³ Smajić, ⁴⁴ and Zhu ⁴⁵	F-1 score: macro, 0.994-0.998; micro, 0.967-0.998; weighted, 0.994-0.998
	scRNA-seq imputation	Smajić ⁴⁴	MSE, 0.133; Pearson correlation coefficients, 0.827
tGPT	Cell clustering	HCA, HCL, Macaque Retina, Tabula Muris, and TCGA	NMI: HCA, 0.75; HCL, 0.77; Macaque Retina, 0.77; GTEx, 0.90; Tabula Muris, 0.87; and TCGA, 0.77
	Cell lineage trajectories inference	GTEx	NMI: GTEx, 0.90
UCE	Data integration across different species	Integrated Mega-scale Atlas (36M cells)	Avg.Bio, 0.88
	Cell type matching	Tabula Sapiens v2	Avg.Bio, 0.65
scFoundation	Bulk tissue drug response prediction	CCLE and Genomics of Cancer Drug Sensitivity data sets	Pearson correlation coefficient: low-grade gliomas, 0.93; WZ-1-84, 0.94
	Single-cell drug response prediction	Genomics of Cancer Drug Sensitivity database, Ho et al, ⁴⁶ Kinker et al ⁴⁷	AUC: NVP-TAE684, 0.84; sorafenib, 0.84; PLX4720, 0.66; etoposide, 0.68
	Perturbation response prediction	Adamson data set with 87 one-gene perturbations, the Dixit data set with 24 one-gene perturbations, and the Norman data set with 131 two-gene perturbations and 105 one-gene perturbations	Pearson correlation coefficient, 0.18
	Cell type annotation	Zheng68k and Segerstolpe	Macro F1: Zheng68k, 0.736; Segerstolpe, 0.914
scGPT	Cell type annotation	A human pancreas data set and a MS data set	Accuracy: pancreas data set > 0.8; MS around 0.85
	Prediction of unseen gene perturbations	Three Perturb-seq data sets of leukemia cell lines: the Adamson data set consisting of 87 one-gene perturbations, ⁹⁸ the curated Replogle data set consisting of 1,823 one-gene perturbations, ⁹⁹ and the Norman data set consisting of 131 two-gene perturbations and 105 one-gene perturbations ¹⁰⁰	Pearson correlation delta score: Adamson, 0.789; Norman, 0.742; Replogle (all perturbations), 0.464; Replogle (TF only), 0.340
	In silico reverse perturbation prediction	A subset of the Norman data set	scGPT identified on average 91.4% relevant perturbations (6.4 of 7) within the top one predictions and 65.7% correct perturbations (4.6 of 7 test cases) within the top eight predictions
	Multi-batch integration	PBMC 10k data set	Avg.Bio (fine-tuned), 0.821
	Multi-omics integration	10× multiome PBMC data set	Avg.Bio (fine-tuned), 0.758
SCimilarity	Cell type matching through similarity	CELLxGENE	Accuracy, 0.865

(continued on following page)

TABLE 2. LLM Applications and Performance Summary: Tasks, Data Sets, and Evaluation Metrics³⁹⁻⁵⁵ (continued)

Model	Tasks	Data Sets	Metrics
scMulan	Cell type annotation	AHCA_BoneMarrow, Simonson et al, ⁴⁸ Suo et al ⁴⁹	F1, AHCA_BoneMarrow, 0.937; Simonson et al, ⁴⁸ 0.934; Suo et al, ⁴⁹ 0.894
	Data integration	A lung data set with 32,472 cells and an immune cell data set with 274,346 cells from healthy participants and patients with COVID-19 disease	Avg.Bio: lung, 0.5-0.6 (exact value not reported); COVID, 0.4-0.45 (exact value not reported)
CellPLM	Spatial transcriptomic gene imputation	Lung2 and Liver2 from lin (source reference not provided in the preprint)	Fine-tuned CellPLM: corr, 0.318 ± 0.015; cosine, 0.481 ± 0.011 in Lung2; corr, 0.328 ± 0.011; cosine, 0.481 ± 0.010 in Liver2
Geneverse	Marker genes identification	Human breast tissue from Lin et al ⁵⁰	LLaVa-7B fine-tuned by LoRA: factual score, 0.9, structural score, 1
Nicheformer	Spatial density prediction	Brain (mouse—MERFISH), liver (CosMx—human), lung (CosMx—human, Xenium—human), and colon (Xenium—human)	Pearson correlation coefficient. Exact values not reported
	Spatial label prediction (cell type, niche, region)	Brain (mouse—MERFISH), liver (CosMx—human), lung (CosMx—human, Xenium—human), and colon (Xenium—human)	Pearson correlation coefficient. Exact values not reported
	Spatial composition prediction	Brain (mouse—MERFISH), liver (CosMx—human), lung (CosMx—human, Xenium—human), and colon (Xenium—human)	Pearson correlation coefficient. Exact values not reported
Novae	Spatial domains prediction	Various data sets from vendor websites (Xenium, MERSCOPE, CosMX)	ARI, FIDE score. Exact values not reported
	Batch effects correction	Various data sets from vendor websites (Xenium, MERSCOPE, CosMX)	ARI, FIDE score. Exact values not reported
scGPT-spatial	Spatial data integration in multislide and multi-modal settings	Visium and Xenium slides from the developing fetal lung data set from Quach et al ⁵¹	Avg.Bio, 0.86
	Reference-based cell type deconvolution	Visium Human Breast data set from Kumar et al ⁵²	Macro F1, 0.58
	Gene imputation	Developing Human Thalamus data set from Kim et al ⁵³	Pearson correlation score over 0.6
spaCCC	Ligand-receptor interaction inference	A RCC data set is obtained from the STOmicsDB database (data set ID: STDS0000223) and a human breast cancer data set is obtained from the 10× Genomics website (Visium Demonstration, Human Breast Cancer, Block A Section 1)	Accuracy, precision, sensitivity, specificity, and MCC reported in Table S1 of the paper, but the table is not available to the public for now
	Cell type annotation	A RCC data set is obtained from the STOmicsDB database (data set ID: STDS0000223) and a human breast cancer data set is obtained from the 10× Genomics website (Visium Demonstration, Human Breast Cancer, Block A Section 1)	Accuracy, renal data set, 96.2% (Table S1 of the paper shows complete accuracy, precision, recall, and macro F1-score metrics, but the table is not available to the public for now)
ST-Align	Spatial cluster identification	Six independent human brain slices from Maynard et al ⁵⁴	ARI, 0.3396
	Gene prediction	Six independent human brain slices from Maynard et al ⁵⁴	MSE, 0.1682
BiomedGPT	Question answering	VQA-RAD, SLAKE, and PathVQA	Accuracy: VQA-RAD, 0.609; SLAKE, 0.854; PathVQA, 0.28
	Radiology reports generation	IU x-ray, Peir Gross, MIMIC-CXR	BiomedGPT-B. Rouge-L: IU x-ray, 28.5; Peir Gross, 36; MIMIC-CXR, 28.7. METEOR, IU x-ray, 12.9; Peir Gross, 15.4; MIMIC-CXR, 15.0. CIDEr: IU x-ray, 40.1; Peir Gross, 122.7; MIMIC-CXR, 23.4
CHIEF	Cancer cell detection	Fifteen independent data sets with a total of 13,661 WSIs	AUROC, 0.9397
	Tumor origin identification	CPTAC	Accuracy range, 0.9319-0.9778
	Prevalent genetic mutations prediction	13,432 WSIs across 30 cancer types and 53 genes with the top five highest mutation rates in each cancer type	AUROC > 0.8
	Survival outcomes prediction	9,404 WSIs in 17 data sets (from both publicly available and institutional sample sources) and focused on seven cancer types (COADREAD, LUSC, BRCA, GBM, UCEC, LUAD, and RCC)	Average concordance index (C-index) of 0.74

(continued on following page)

TABLE 2. LLM Applications and Performance Summary: Tasks, Data Sets, and Evaluation Metrics³⁹⁻⁵⁵ (continued)

Model	Tasks	Data Sets	Metrics
CONCH	Histology image classification	TCGA BRCA (invasive breast carcinoma subtyping), TCGA NSCLC (subtyping), TCGA RCC (subtyping), and DHMC LUAD (histologic pattern classification), and three ROI-level tasks: CRC100k (CRC tissue classification), WSSS4LUAD (LUAD tissue classification), and SICAP (Gleason pattern classification)	Accuracy in zero-shot setting, NSCLC subtyping, 0.907; RCC subtyping, 0.902
	Segmentation	SICAP for prostate tumor versus normal tissue segmentation and on DigestPath for malignant versus benign tissue segmentation in CRC specimens	SICAP: mean Dice score, 0.601; mean recall, 0.751; mean precision, 0.672; DigestPath: mean Dice score, 0.615; mean recall, 0.709; mean precision, 0.663
	Text-to-image and image-to-text retrieval	Source A and source B (both are held-out sources from model pretraining that cover a diverse range of general pathology concepts) and TCGA LUAD	Mean recall, 0.44
Prov-GigaPath	Mutation prediction	Providence and TCGA	Macro-AUROC, 0.626
	Cancer subtyping	Providence and TCGA	AUROC, range from 0.874 in EGC to 0.978 in OVT
TITAN	Rare cancer retrieval tasks	TCGA, CPTAC, DHMC, OT108, EBRAINS, MGB	Similarity, PGNG, 0.794; PHC, 0.651
	Morphologic classification tasks	TCGA, CPTAC, DHMC, OT108, EBRAINS, MGB	Accuracy (for multiclass)/AUROC (for binary), 0.781
	Molecular classification tasks	TCGA, CPTAC, DHMC, OT108, EBRAINS, MGB	Accuracy (for multiclass)/AUROC (for binary), 0.796
	Survival prediction tasks	TCGA, CPTAC, DHMC, OT108, EBRAINS, MGB	Concordance index, 0.716
UNI	ROI cell type segmentation	SegPath	Dice scores, epithelial, smooth muscle, and RBC types are 0.827, 0.690, and 0.803, respectively
	Slide classification using few-shot class prototypes	PANDA ⁵⁵	Quadratic weighted Cohen's κ , 0.946
	Cancer subtyping	Brain tumor from EBRAINS	Balanced ACC, 0.883; weighted F1, 0.926; AUROC, 0.996

Abbreviations: ACC, accuracy; ARI, adjusted rand index; AUROC, area under the receiver operating characteristic curve; BRCA, breast invasive carcinoma; CCLE, cancer cell line encyclopedia; CellLM, single-cell language model; CellPLM, single-cell pretrained language model; CHIEF, Clinical Histopathology Imaging Evaluation Foundation; COADREAD, colon adenocarcinoma and rectum adenocarcinoma; CONCH, contrastive learning from captions for histopathology; CPTAC, clinical proteomic tumor analysis consortium; CRC, colorectal cancer; DHMC, Dartmouth Hitchcock Medical Center; EBRAINS, European brain research infrastructures; EC, endothelial cell; EGC, early gastric cancer; FIDE, faithfulness, informativeness, diversity, and error; GBM, glioblastoma; GDSC, genomics of drug sensitivity in cancer; GTEx, genotype-tissue expression; HCA, Human Cell Atlas; HCL, human cell landscape; LLM, large language model; LoRA, low-rank adaptation; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MAE, mean absolute error; MCC, Matthews correlation coefficient; MGB, Mass General Brigham; MIMIC-CXR, MIMIC chest x-ray; MS, Multiple Sclerosis; MSE, mean squared error; NMI, normalized mutual information; NSCLC, non-small cell lung cancer; OVT, oncolytic virotherapy; PBMC, peripheral blood mononuclear cell; PGNG, paraganglioma; PHC, pheochromocytoma; RCC, renal cell carcinoma; RMSE, root mean square error; ROI, region of interest; scBERT, single-cell bidirectional encoder representations from transformers; scGPT, single-cell generative pretrained transformer; scHyena, single-cell Hyena; scMulan, single-cell multitask generative pretrained language model; scRNA-seq, single-cell RNA sequencing; SICAP, prostate cancer grade assessment; SLAKE, Semantically-Labeled Knowledge-Enhanced data set; spaCCC, spatial transcriptomics cell-cell communications; TCGA, The Cancer Genome Atlas; TF, transcription factor; UCE, Universal Cell Embedding; UCEC, uterine corpus endometrial carcinoma; WSI, whole-slide image.

metric-learning framework, can also reveal shared cellular states across different organs and diseases.⁸ It uncovered fibrosis-associated macrophage populations in multiple cancers, including uveal melanoma, pancreatic ductal adenocarcinoma (PDAC), and colon cancer, potentially highlighting shared mechanisms of cancer progression.⁸ This can help reveal common pathways and mechanisms driving cancer progression, regardless of the cancer type.

Integrating scRNA-seq data sets across multiple batches and modalities in single-cell multiomics presents unique challenges—specifically, the need to remove technical batch effects while preserving critical biological signals. Single-cell generative pretrained transformer (scGPT) has demonstrated effectiveness in addressing this issue.³⁰

Compared with established methods such as Seurat,⁵⁷ scVI,⁵⁹ and Harmony,⁶⁰ scGPT produces more distinct and biologically meaningful cell clusters when analyzing paired gene expression with chromatin accessibility or protein abundance data set.³⁰ Notably, scGPT was able to identify major cell types and subtypes that were missed by other methods.³⁰ For instance, it successfully identified a cluster of CD8+ naïve cells—key targets of immune checkpoint blockade and adoptive T-cell therapy for mediate antitumor responses⁶¹—using joint single-cell RNA and assay for transposase-accessible chromatin with high-throughput sequencing data.³⁰ Although models such as scGPT,³⁰ scBERT,¹⁰ Geneformer,⁹ and scFoundation⁷ perform well in multibatch and multiomics integration tasks, they often require extensive fine-tuning for multibatch or multiomics

integration. However, fine-tuning typically demands large amounts of data and substantial computation resources. Single-cell multitask generative pre-trained language model (scMulan) addresses this limitation by encoding both gene expression and metadata terms into a structured cell sentence,³¹ enabling the model to learn both microscopic (gene-level) and macroscopic (metadata-level) features during pretraining.³¹ Through the approach, scMulan supports zero-shot batch integration, reducing the need for data set-specific tuning while maintaining robust performance.³¹

Drug resistance continues to be a major challenge in cancer therapy. Accurate prediction of drug sensitivities, especially at the single-cell level, is crucial for guiding anticancer drug design and elucidating mechanisms of resistance.⁶² Several LLMs have been evaluated for this purpose. Among them, the single-cell language model (CellLM)²⁶ is the first single-cell LLM pretrained on a cancer database, CancerSCEM.⁶³ This domain-specific pretraining enables CellLM to capture cancer-specific transcriptomic features, resulting in more accurate single-cell representations.²⁶ As a result, CellLM outperformed scBERT in predicting drug sensitivity in single-cell and single-omics data sets, including those derived from human lung cancer and oral squamous cell carcinoma.²⁶ Another notable LLM, scFoundation,⁷ has also achieved state-of-the-art performance in predicting cancer drug responses. It accurately predicted single-cell responses to four anticancer agents: sorafenib, NVP-TAE684, PLX4720, and etoposide.⁷ One reason for scFoundation's strong performance may stem from its tokenization method. Unlike scBERT, scGPT, or Geneformer, which rely on binned or ranked normalized gene expression values, scFoundation uses continuous normalized expression values as input. This approach could help preserve the full dynamic range of gene expression, allowing the model to capture subtle expression differences and detect weak but informative biological variations. These findings indicate that different LLMs may excel under different settings, depending on cancer type, drug class, and data modality. For example, CellLM performs well with lung cancer and oral squamous cancer cell line data, while scFoundation excels well with drug sorafenib, drug WZ-1-84, and low-grade glioma cancer types. However, these conclusions are based on limited benchmark data sets, underscoring the need for comprehensive, systematic evaluations across diverse cancer types and drug treatments.

LLMs FOR SPATIAL TRANSCRIPTOMICS APPLICATIONS IN CANCER

To provide spatial context of single-cell data, spatial transcriptomics has emerged as a powerful approach that preserves tissue architecture while enabling gene expression profiling.^{64–68} LLMs have recently demonstrated utility in a variety of spatial transcriptomics tasks, including multislide and multimodal spatial data integration, gene expression imputation, niche and region label prediction, spatial

clustering identification, cell-cell communication inference, and marker gene identification (Tables 1 and 2).

The complexity of spatial transcriptomics data arises largely from the inclusion of spatial context and protocol-specific biases.¹⁹ To capture modality-specific features and facilitate generalizability across different sequencing protocols, scGPT-spatial¹⁹ FM uses a Mixture-of-Experts (MoE) decoder to perform protocol-aware integration of tissue samples.¹⁹ The MoE-based decoder is consisted of a learnable gating network and four feed-forward network of experts, enabling the model to capture gene expression profiles in a protocol-aware fashion.¹⁹ The MoE decoder architecture is paired with the gene expression prediction objective to generate query gene expression values from post-transformer embeddings.¹⁹ This is advantageous over conventional single feed-forward network decoders, which typically apply a uniform set of parameters across all inputs and do not account for protocol-specific variability. In zero-shot settings, it outperformed traditional approaches such as principal component analysis and Seurat⁵⁷ in integrating Visium and Xenium lung data sets for downstream cell type clustering. It also reduced batch effects in multislide integration, as shown in fetal thalamus¹⁹ data sets using five MERFISH slices. Another LLM, Novae, leverages a self-supervised graph attention network to encode local microenvironment into spatial representations. Novae is capable of integrating data from different slides, platforms, and gene panels.³⁴ For example, when applied to two slides from a breast data set (one MERSCOPE and one Xenium slide), Novae better aligns the slides in the latent space compared with existing methods including STAGATE⁶⁹ and SpaceFlow.³⁴ The spatial representations learned by Novae proved effective in downstream analyses: when applied to nondiseased lymph node and lymphoma tissues, Novae uncovered distinct spatial domains that reflected differences in gene expression, cell type proportions, and network organization—highlighting its potential for revealing biologically meaningful patterns between diseased and healthy tissues.³⁴

Two key prediction tasks in spatial transcriptomics involve different directions of information transfer: gene-wise imputation, where gene expression in spatial data is inferred using scRNA-seq as a reference⁷⁰; and cell-wise spatial context prediction, where the spatial origin of dissociated cells in scRNA-seq is inferred using spatial transcriptomics as a reference. A number of LLMs have demonstrated capabilities in these tasks. In the context of gene imputation, single-cell pretrained language model (CellPLM)²⁷ was pretrained on spatial transcriptomics data and incorporates spatial relationships among cells to improve imputation accuracy. It outperformed Tangram in MERSCOPE-derived lung and liver cancers data sets.²⁷ Similarly, scGPT-spatial¹⁹ uses spatially aware learning to reconstruct the gene expression profile of the central spot on the basis of its neighboring spots. It demonstrated superior results in MERFISH-to-scRNA-seq and Xenium-to-Visium reference

tasks. Notably, scGPT-spatial) more accurately predicted spatial expression patterns of key marker genes such as *MKI67* (a proliferation marker) and *IGFBP5* (an epithelial marker), aligning more closely with ground truth compared with alternative methods.¹⁹ In the cell-wise spatial context prediction task, Nicheformer—a transformer-based LLM—integrates human and mouse single-cell and spatial transcriptomics data to learn spatially informed cellular representations.³³ This enables it to transfer niche and regional annotations to dissociated scRNA-seq profiles.³³ When applied to mouse primary motor cortex scRNA-seq data set, Nicheformer accurately predicted spatial niche labels for non-neuronal cell types and assigned brain region labels to dissociated neurons, particularly excitatory neurons, demonstrating its utility in capturing nuanced spatial information.³³

Beyond the primary functions described above, other LLMs have also demonstrated utility in more specialized spatial transcriptomics tasks. For example, ST-Align³⁵ is a multimodal foundation model that aligns image-gene pairs by incorporating spatial context, effectively bridging pathologic imaging with transcriptomic features. In zero-shot settings, ST-Align outperformed both unimodal and other multimodal approaches in identifying spatial clusters within human brain tissue.³⁵ Notably, it more accurately delineated the boundary between white matter and cortical layer L6 compared with existing frameworks such as contrastive language-image pre-training⁷¹ and pathology language-image pretraining (PLIP).⁷² Another model, spatial transcriptomics cell-cell communications (SpaCCC),²⁰ combines a fine-tuned version of scGPT with functional gene interaction networks to infer ligand-receptor interactions. SpaCCC effectively corrects biases toward highly expressed ligands and receptors, enabling the identification of biologically meaningful yet previously overlooked interactions—for example, *L1CAM:ERBB3* and *SLIT3:ROBO1* between neutrophils and mast cells.²⁰ Additionally, LLaVa-7B, fine-tuned via low-rank adaptation (LoRA) on image data and incorporated into the Geneverse³² model collection, has exhibited strong capabilities in marker gene identification across cell types by analyzing gene expression patterns in spatial transcriptomics data sets.

APPLICATIONS OF LLMs IN CANCER PATHOLOGY

Building on the success of LLMs in spatial transcriptomics, there is growing interest in coupling spatial transcriptomic expression profiles with histopathologic imaging to enable a more comprehensive understanding of tissue organization and disease states.⁷³ This convergence has driven the application of LLMs in computational pathology, where models leverage high-resolution whole-slide images (WSIs) to perform tasks such as cancer subtyping, rare malignancies detection, genomic mutation prediction, image segmentation, and cross-modal retrieval (Tables 1 and 2).

Cancer subtyping and the detection of rare malignancies are crucial for precision oncology, but they remain challenging because of the heterogeneity of tumor histology and the limited availability of annotated data for rare cancer types.⁷⁴ To address these challenges, Prov-GigaPath—a model built on a vision transformer architecture adapted from LongNet—integrates both tile- and slide-level encoders to effectively capture local and global morphologic features.²² This approach enables accurate subtyping across six major cancer types: ovarian cancer, brain tumors, renal cell carcinoma (RCC), colorectal adenocarcinoma, hepatobiliary cancers, and diffuse intrinsic pontine glioma.²² Likewise, UNI, a general-purpose self-supervised model, has been evaluated on 108 cancer types and subtypes, including 90 rare cancers.³⁸ At region of interest (ROI) level, UNI can perform competitively on tasks such as breast metastasis detection and colon cancer microsatellite instability screening.³⁸ Specializing in rare cancers, Virchow⁷⁵ has demonstrated robust generalization across seven rare tumor types, including those of the liver, stomach, brain, ovary, cervix, testis, and bone.⁷⁵ Similarly, contrastive learning from captions for histopathology (CONCH) is a visual-language FM developed using diverse sources of histopathology images, biomedical text, as well as image-caption pairs.¹³ CONCH was trained using a combination of contrastive alignment objectives that align image and text modalities in a shared representation space and a captioning objective that teaches the model to predict descriptive text for a given image.¹³ The contrastive objective helps the model learn fine-grained associations by pulling matched image-text pairs closer together and pushing apart mismatched ones, which strengthens cross-modal understanding. CONCH achieved state-of-the-art performance in both zero-shot and few-shot histology classification tasks, including slide-level classification of breast cancer, RCC, and non-small cell lung cancer, as well as ROI-level classification in colorectal cancer (CRC).¹³ When combined with weakly supervised learning, where they set the contrastive loss to zero and retained only the captioning loss in the training objective, CONCH achieved the best performance for classification of 30 categories of rare brain tumors,¹³ outperforming other methods such as CTransPath.⁷⁶ Similarly, TITAN,²¹ a multimodal WSI FM, integrates visual self-supervised learning with vision-language alignment using both pathology reports and synthetic captions. Evaluated on a curated data set of 43 rare cancer types and a public data set of 29 rare cancers, TITAN effectively retrieves slides with similar morphologic features, such as those associated with paraganglioma.²¹

Because of the high cost and limited feasibility of routine genomic profiling in clinical practice, computational methods to infer tumor genotypes from histopathologic images have gained traction.⁷⁷ For example, Prov-GigaPath demonstrated superior performance in identifying mutations in five key lung adenocarcinoma (LUAD)-associated genes—*EGFR*, *FAT1*, *KRAS*, *TP53*, and *LRP1B*.²² It also outperformed HIPT,⁷⁸ CTransPath,⁷⁶ and REMEDIS¹¹ in

predicting 18 pan-cancer biomarkers.²² The enhanced performances can be attributed to the LongNet Vision Transformer architecture, which effectively captures global image patterns.²² Similarly, Clinical Histopathology Imaging Evaluation Foundation (CHIEF) model, pretrained on 60,530 WSI spanning 19 anatomic sites, achieved excellent accuracy in predicting molecular profiles across diverse cancer types.³⁷ For instance, CHIEF successfully identified critical gene mutations across various cancer, such as *TP53* mutations in low-grade glioma, adrenal carcinoma, and uterine corpus endometrial carcinoma; *GTF2I* mutations in thymic epithelial tumors; and *BAP1* mutation for uveal melanoma.³⁷ In addition, the model predicted genes associated with US Food and Drug Administration (FDA)–approved targeted therapies, such as *EZH2* in diffuse large B-cell lymphoma.³⁷ Additionally, CHIEF predicted isocitrate dehydrogenase status for glioma classification and microsatellite instability status for predicting drug response of immune checkpoint blockade in patients with CRC.³⁷

Cross-modal retrieval tasks, such as text-to-image and image-to-text matching, can assist in research cohort identification for clinical trials and the detection rare disease morphologies. Models such as CONCH have demonstrated the ability to perform these tasks in a zero-shot setting by learning an aligned latent space for visual and language embeddings.¹³ For example, CONCH outperforms models such as PLIP,⁷² BiomedCLIP,⁷⁹ and OpenAICLIP⁷¹ in retrieving diverse pathology cases, such as LUAD.¹³ In addition, BiomedGPT,⁸⁰ a lightweight vision-language FM, has extended cross-modal retrieval to other domains, including chest x-rays and clinical photographs. Beyond specialized biomedical models, general-purpose LLMs such as generative pretrained transformer (GPT)–4 and GPT–3.5 have also been evaluated for analyzing and generating text-based pathology reports.^{12,81} For example, Sushil et al⁸¹ assessed the capability of GPT–4⁸² and GPT–3.5⁸³ in breast cancer pathology classification, extracting information such as tumor grade, tumor histology, sites of disease, and margin status. Across 12 classification tasks, zero-shot GPT–4 demonstrated comparable or even significantly better performances compared with task-specific supervised models,⁸¹ such as long short-term memory.⁸⁴ The study highlighted that a major source of error in GPT–4's predictions came from grouping results into the “others” category.⁸¹ In another application, GPT–4⁸² and GPT–3.5⁸³ were evaluated for generating synoptic pathology reports for PDAC from computed tomography (CT) reports.¹² Applied to 180 consecutive PDAC CT staging reports, GPT–4 outperformed GPT–3.5, achieving near-perfect report generation and accurately determining tumor resectability through chain-of-thought prompting strategies.¹²

WSIs often span gigapixels and exhibit substantial heterogeneity, containing a variety of cell types, morphologies, and tissue structures—each occupying only a small portion of the slide.¹³ This makes slide-level segmentation both a

computationally intensive and biologically valuable task, as it allows for the identification of biologically distinct regions and reduces the number of tiles needed for downstream analyses.¹³ CONCH model addressed this challenge by dividing each slide into tiles and using zero-shot classification to categorize each tile as tumor or benign. It demonstrated the best performance on prostate tumor and colon cancer data sets, outperforming PLIP,⁷² BiomedCLIP,⁷⁹ and OpenAICLIP.¹³ For pan-cancer ROI cell type segmentation task, UNI model outperformed hierarchical vision backbones CTransPath⁷⁶ and convolutional neural networks on a data set comprising eight major cell types, including epithelial cells, leukocytes, lymphocytes, plasma cells, myeloid cells, and others.

FINE-TUNING STRATEGIES FOR LLMs FOR CANCER-SPECIFIC APPLICATIONS

Given the scarcity of high-quality, annotated cancer data sets, fine-tuning LLMs for cancer-specific tasks requires strategies that are both domain-adaptive and data-efficient. One approach is domain-adaptive pretraining, where models are further pretrained on oncology-specific corpora. For example, Huemann et al⁸⁵ adapted bidirectional encoder representations from transformers (BERT) models using nuclear medicine domain knowledge by incorporating text reports and images from fluorodeoxyglucose positron emission tomography-CT scans in patients with lymphoma. These domain-adapted models significantly outperformed standard BERT models in predicting the five-point Deauville score, which assesses treatment response in lymphoma, with a 4.4% improvement in accuracy. In settings with limited labeled data, parameter-efficient fine-tuning strategies—such as LoRA—can adapt large models by updating only a small subset of parameters. For instance, a benchmarking study on single-cell drug response prediction showed that, in tumor tissues, UCE model achieved an F1 score of 0.7230 in a zero-shot setting, which was improved to 0.774 after fine-tuned with LoRA using cancer-specific data sets.⁸⁶ Additionally, multitask learning allows simultaneous training across related cancer tasks enabling knowledge sharing and improved generalization. For example, M3FM is a medical multimodal-multitask FM for 3D low-dose CT lung cancer screening.⁸⁷ It supports 17 tasks such as lung nodule detection and characterization and lung cancer risk prediction by integrating 3D CT, voxel size, clinical texts, and user queries. To summarize, incorporating domain-adaptive pretraining, parameter-efficient tuning, and multitask learning will be critical to building robust and scalable LLMs tailored to real-world cancer research challenges.

OPEN DATA SETS AND BENCHMARKING RESOURCES FOR CANCER LLMs

Reproducibility and standardized benchmarking are essential for the responsible development and evaluation of LLMs in cancer research. In terms of reproducibility, we found that

all the models reviewed in this study have released their model weights, code for pretraining and fine-tuning, and provided references to the data sets used for pretraining or evaluation. These resources are typically made available through platforms such as GitHub, Zenodo, or Hugging Face. For standardized benchmarking, many open-access data sets are currently available. In single-cell analysis, HuB-MAP,⁴ HCA consortium,²³ Tabula Sapiens,²⁴ and PanglaoDB²⁵ provide multiomics data paired with clinical annotations. In spatial transcriptomics, resources such as SpatialHuman30M comprised 30 million spatial transcriptomics profiles, encompassing both image- and sequencing-based protocols.¹⁹ In computational pathology, HEST-1k,⁶ which links histology images with spatial gene expression profiles, could serve as valuable starting points. Furthermore, the Gene Expression Omnibus (GEO) offers a large collection of public data sets for benchmarking. In addition, Genomics Data Automatic Exploration Benchmark presents a standardized analysis pipeline for gene expression analysis to identify disease-associated genes. Its benchmark data set includes The Cancer Genome Atlas and GEO, supporting reproducibility and consistent evaluation of LLMs in cancer genomics research.⁸⁸ In terms of benchmarking platforms, scDrugMap is a recent example.⁸⁶ It benchmarks eight existing single-cell LLMs on tumor cell drug response prediction using 53 curated single-cell transcriptomics data sets across multiple drugs, tissue types, and cancer types.⁸⁶ However, such standardized benchmarking frameworks remain limited and needs further efforts.

DISCUSSION

The application of LLMs to single-cell, spatial transcriptomics, and histologic images marks significant progress in processing and analyzing high-dimensional biomedical data. Despite these advances, each domain presents distinct challenges that should be addressed to fully harness the potential of LLMs in cancer research.

LLMs in Single-Cell Analysis

Although many single-cell LLMs are pretrained on large and diverse data sets, they often fall short in capturing the cancer-specific cellular heterogeneity and the complexity of the tumor microenvironment. Notably, our review found that five of 11 single-cell LLMs—CellPLM, Geneformer, scHyena, UCE, and scMulan—have yet to be evaluated on cancer data sets, underscoring a gap in their applicability to oncology. Additionally, most existing single-cell LLMs primarily rely on gene expression data and often do not incorporate additional omics or contextual information that is critical in cancer, such as somatic mutations, copy-number variations, or epigenetic states. This could limit their ability to fully model tumor heterogeneity, clonal evolution, and microenvironmental interactions—factors that are central to cancer biology and therapeutic response.⁸⁹

LLMs in Spatial Transcriptomics Analysis

In the spatial domain, challenges such as data sparsity—particularly in technologies such as Xenium and MERFISH—small gene panels, limited availability of annotated data sets, and a lack of standardized and accurate cell type annotations hinder the robustness and generalizability of LLM-based models. Furthermore, to contextualize model generalizability, differences between spatial transcriptomics platforms, such as sample preparation protocols, RNA detection approaches, gene selection for panel design, as well as cell segmentation processes for image-based technologies should also be considered.⁹⁰ For example, MERFISH and Xenium are two subcellular resolution image-based spatial transcriptomics technologies. In RNA detection, MERFISH uses combinatorial barcoding, multiple hybridization rounds, and error correction to identify RNA transcripts.⁹¹ Xenium, however, uses padlock probes and a ligation-based detection method to quantify RNA molecules.⁹¹ The modality differences may influence how well models trained on one platform generalize to others, highlighting the need for platform-aware benchmarking and cross-platform evaluations when developing or deploying LLMs in spatial transcriptomics.

LLMs in Computational Pathology

LLMs trained on general pathology images may struggle to recognize the nuanced morphologic patterns unique to cancer. For example, although CONCH is pretrained on millions of images and shows superior performance on general pathology data sets, its effectiveness on cancer-specific tasks remains limited. Although it has demonstrated improved performance on lung cancer data sets compared with previous methods, further gains will likely depend on enhancements in both the quantity and quality of histopathology data sets. Overcoming these limitations is essential to unlock the full potential of LLMs in advancing cancer research and clinical applications. Future efforts focused on pretraining models with large-scale, cancer-specific data sets will be critical for developing the next generation of LLMs—models that are not only robust and generalizable, but also interpretable and optimized for translational oncology.

Limitations of Zero-Shot Evaluation in LLMs

Claims of strong zero-shot performance in models such as CONCH and GPT-4 should be interpreted with caution. A key concern is the lack of independence between training and test distributions. FMs are typically trained on massive web-scale data sets curated by developers, but the exact contents of these data sets are rarely disclosed.³⁶ By contrast, evaluation benchmarks are often constructed using publicly available data and openly released, making it likely that some of this content was included—directly or indirectly—in the model's training data or will be used

in future training.³⁶ This raises the risk of data leakage, where models appear to generalize but are in fact recalling previously seen information. Assessing train-test overlap is therefore critical for reliable evaluation. However, a recent study found that only 9 of 30 reviewed papers reported any statistics on train-test overlap, limiting the community's ability to assess the validity of zero-shot claims.³⁶ Therefore, model developers are encouraged to provide documentation or transparency reports describing the sources and characteristics of the pretraining data, even if the data itself cannot be released. Additionally, developing evaluation benchmarks that use private or newly collected data sets—unlikely to appear in web-scale training corpora—and applying temporally split evaluations, where training data predate the test data, can also reduce the risk of data leakage and improve the validity of zero-shot performance assessments in LLMs.

Need for Standardized Benchmarks in Cancer Informatics

As LLMs continue to be applied in cancer research, there is an urgent need for standardized benchmarking frameworks to ensure reproducibility and fair comparisons across data types and model architectures. In natural language processing (NLP), benchmarks such as GLUE⁹² and SuperGLUE⁹³ provide unified evaluation protocols across a diverse set of NLP tasks. Similarly, ImageNet has transformed computer vision by offering a large-scale, hierarchically labeled image data set.⁹⁴ Inspired by these successes, the cancer informatics community would greatly benefit from a domain-specific benchmarking framework. Such a framework should, first, reflect a range of real-world use cases such as molecular tasks (eg, drug-response prediction)⁸⁶ and clinical tasks (eg, pathology report generation).¹³ Second, it should support multigranularity evaluation metrics, going beyond accuracy to include train-test statistics, factual consistency, biological plausibility, interpretability, and clinical utility.⁹⁵ Third, high-quality, well-documented, and securely managed data sets—such as the HEST-1k cancer data set for spatial transcriptomics and histology images—should be curated to improve robustness and reduce the risk of train-test leakage, especially in zero-shot evaluations.⁶ A recent example is scDrugMap, which provides benchmarks and data sets for evaluating single-cell LLMs on tumor cell drug response prediction.⁸⁶ Overall, a consensus and robust and framework could be essential to guide the responsible

development, validation, and deployment of LLMs in cancer research.

Translational Challenges for LLMs in Cancer Research

In conclusion, despite recent advances, several translational hurdles must be addressed before LLMs can be effectively deployed in clinical cancer research and care. First, model interpretability remains a key challenge. Our study found that six LLMs examined in this review explicitly incorporate interpretability mechanisms: scBERT, CellLM, tGPT, SCimilarity, scMulan, and CHIEF. Specifically, scBERT¹⁰ and CellLM²⁶ use top-attention-gene lists to interpret cell annotation results. tGPT calculates gene importance scores that help distinguish cancer types and predict therapeutic outcomes, particularly for immunotherapy.¹⁸ SCimilarity applies integrated gradients to quantify the importance of each gene for specific cell types, recovering known biological markers such as surfactant genes for lung alveolar type 2 cells.⁸ scMulan identifies signature genes for cell type annotation that overlap with established marker genes.³¹ Finally, CHIEF visualizes attention scores over quantitative image feature vectors to interpret computational pathology predictions.³⁷ We advocate for broader adoption of interpretability tools in future LLM development for cancer research. Second, data set bias is a concern, as many LLMs are trained on data that may not represent the diversity of patient populations, cancer subtypes, or clinical settings, potentially leading to poor generalization or unintended disparities in real-world applications. Third, most LLMs have only been evaluated on retrospective data sets,^{13,21,38} and their performance in prospective clinical trials remains largely unexplored. Prospective validation is critical to establishing safety, effectiveness, and reproducibility. Fourth, regulatory hurdles remain a major barrier to clinical translation. LLMs developed for cancer research should comply with emerging guidelines on artificial intelligence-/ML-based software as a medical device, such as the Good Machine Learning Practice by FDA.⁹⁶ However, the regulatory pathway for FMs remains largely undefined. Finally, although LLMs show promise for clinical decision support, none is currently authorized by the FDA as a clinical decision support device.⁹⁷ Regulatory approval will require robust evidence of clinical utility, explainability, and consistent performance across diverse health care environments. Addressing these challenges is essential to ensure that LLMs are not only powerful but also reliable, equitable, and suitable for real-world cancer applications.

AFFILIATIONS

¹Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL

²Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX

³Department of Cardiothoracic Surgery, Weill Cornell Medicine, Cornell University, New York, NY

⁴Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN

⁵Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD

CORRESPONDING AUTHOR

Qianqian Song, PhD; e-mail: qsong1@ufl.edu.

SUPPORT

Q.S. is supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM151089). J.S. is supported by the National Library of Medicine of the National Institutes of Health (R01LM013771). J.S. is also supported by the National Institute on Alcohol Abuse and Alcoholism (R21AA031370 and U24AA026969), the National Institute of Health Office of the Director (OT2OD031919), the Indiana University Melvin and Bren Simon Comprehensive Cancer Center Support Grant from the National Cancer Institute (P30CA082709), and the Indiana University Precision Health Initiative. G.W. is supported by the National Institute of General Medical Sciences of the National Institutes of Health (1R35GM150460).

AUTHOR CONTRIBUTIONS

Conception and design: Yining Pan, Guangyu Wang, Jing Su, Umit Topaloglu, Qianqian Song

Collection and assembly of data: Yining Pan, Qianqian Song

Data analysis and interpretation: Yining Pan, Yanfei Wang, Jing Su, Umit Topaloglu, Qianqian Song

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians ([Open Payments](http://OpenPayments.org)).

Umit Topaloglu

This author is an Associate Editor for *JCO Clinical Cancer Informatics*. Journal policy recused the author from having any role in the peer review of this manuscript.

Stock and Other Ownership Interests: Care Directions

Consulting or Advisory Role: N-Power Medicine

No other potential conflicts of interest were reported.

REFERENCES

- Bray F, Laversanne M, Weiderpass E, et al: The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* 127:3029-3030, 2021
- Gagan J, Van Allen EM: Next-generation sequencing to guide cancer therapy. *Genome Med* 7:80, 2015
- Wang D, Liu B, Zhang Z: Accelerating the understanding of cancer biology through the lens of genomics. *Cell* 186:1755-1771, 2023
- Jain S, Pei L, Spraggins JM, et al: Advances and prospects for the Human BioMolecular Atlas Program (HuBMAP). *Nat Cell Biol* 25:1089-1100, 2023
- Clark K, Vendt B, Smith K, et al: The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J Digit Imaging* 26:1045-1057, 2013
- Jaume G, Doucet P, Song AH, et al: HEST-1k: A dataset for spatial transcriptomics and histology image analysis. *arXiv* [10.48550/arXiv.2406.16192](https://arxiv.org/abs/10.48550/arXiv.2406.16192)
- Hao M, Gong J, Zeng X, et al: Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 21:1481-1491, 2024
- Heimberg G, Kuo T, DePianto DJ, et al: A cell atlas foundation model for scalable search of similar human cells. *Nature* 638:1085-1094, 2025
- Theodoris CV, Xiao L, Chopra A, et al: Transfer learning enables predictions in network biology. *Nature* 618:616-624, 2023
- Yang F, Wang W, Wang F, et al: scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* 4:852-866, 2022
- Azizi S, Culp L, Freyberg J, et al: Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nat Biomed Eng* 7:756-779, 2023
- Bhayana R, Nanda B, Dehkharghanian T, et al: Large language models for automated synoptic reports and resectability categorization in pancreatic cancer. *Radiology* 311:e233117, 2024
- Lu MY, Chen B, Williamson DFK, et al: A visual-language foundation model for computational pathology. *Nat Med* 30:863-874, 2024
- Liu X, Wang Q, Zhou M, et al: DrugFormer: Graph enhanced language model to predict drug sensitivity. *Adv Sci* 11:2405861, 2024
- Bian H, Chen Y, Luo E, et al: General-purpose pre-trained large cellular models for single-cell transcriptomics. *Natl Sci Rev* 11:nwae340, 2024
- Lipkova J, Kather JN: The age of foundation models. *Nat Rev Clin Oncol* 21:769-770, 2024
- Guo F, Guan R, Li Y, et al: Foundation models in bioinformatics. *Natl Sci Rev* 12:nwaf028, 2025
- Shen H, Liu J, Hu J, et al: Generative pretraining from large-scale transcriptomes for single-cell deciphering. *iScience* 26:106536, 2023
- Wang C, Cui H, Zhang A, et al: scGPT-spatial: Continual pretraining of single-cell foundation model for spatial transcriptomics. *bioRxiv* [10.1101/2025.02.05.636714](https://doi.org/10.1101/2025.02.05.636714)
- Ji B, Wang X, Qiao D, et al: SpaCCC: Large language model-based cell-cell communication inference for spatially resolved transcriptomic data. *Big Data Min Anal* 7:1129-1147, 2024
- Ding T, Wagner SJ, Song AH, et al: Multimodal whole slide foundation model for pathology *arXiv* [10.48550/arXiv.2411.19666](https://arxiv.org/abs/10.48550/arXiv.2411.19666)
- Xu H, Usuyama N, Bagga J, et al: A whole-slide foundation model for digital pathology from real-world data. *Nature* 630:181-188, 2024
- Rood JE, Wynne S, Robson L, et al: The Human Cell Atlas from a cell census to a unified foundation model. *Nature* 637:1065-1071, 2025
- Tabula Sapiens Consortium, Jones RC, Karkanas J, et al: The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* 376:eab4896, 2022
- Franzén O, Gan LM, Björkregren JLM: PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database* 2019:baz046, 2019
- Zhao S, Zhang J, Nie Z: Large-scale cell representation learning via divide-and-conquer contrastive learning. *arXiv* [10.48550/arXiv.2306.04371](https://arxiv.org/abs/10.48550/arXiv.2306.04371)
- Wen H, Tang W, Dai X, et al: CellPLM: Pre-training of cell language model beyond single cells. *bioRxiv* [10.1101/2023.10.03.560734](https://doi.org/10.1101/2023.10.03.560734)
- Oh G, Choi B, Jung I, et al: scHyena: Foundation model for full-length single-cell RNA-seq analysis in brain. *arXiv* [10.48550/arXiv.2310.02713](https://arxiv.org/abs/10.48550/arXiv.2310.02713)
- Rosen Y, Roohani Y, Agarwal A, et al: Universal Cell Embeddings: A foundation model for cell biology. *bioRxiv* [10.1101/2023.11.28.568918](https://doi.org/10.1101/2023.11.28.568918)
- Cui H, Wang C, Maan H, et al: scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 21:1470-1480, 2024
- Bian H, Chen Y, Dong X, et al: scMulan: A multitask generative pre-trained language model for single-cell analysis. *Lecture Notes in Computer Science*, 2024, pp 479-482
- Liu T, Xiao Y, Luo X, et al: Geneverse: A collection of open-source multimodal large language models for genomic and proteomic research. *Findings of the Association for Computational Linguistics: EMNLP* 2024, 2024, pp 4819-4836
- Schaar AC, Tejada-Lapuente AP, Palla G, et al: Nichformer: A foundation model for single-cell and spatial omics. *bioRxiv* [10.1101/2024.04.15.589472](https://doi.org/10.1101/2024.04.15.589472)
- Blampey Q, Benkirane H, Bercovici N, et al: Novae: A graph-based foundation model for spatial transcriptomics data. *bioRxiv* [10.1101/2024.09.09.612009](https://doi.org/10.1101/2024.09.09.612009)
- Lin Y, Luo L, Chen Y, et al: ST-Align: A multimodal foundation model for image-gene alignment in spatial transcriptomics. *arXiv* [10.48550/arXiv.2411.16793](https://arxiv.org/abs/10.48550/arXiv.2411.16793)
- Zhang AK, Klyman K, Mai Y, et al: Language model developers should report train-test overlap. *arXiv* [10.48550/arXiv.2410.08385](https://arxiv.org/abs/10.48550/arXiv.2410.08385)
- Wang X, Zhao J, Marostica E, et al: A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* 634:970-978, 2024
- Chen RJ, Ding T, Lu MY, et al: Towards a general-purpose foundation model for computational pathology. *Nat Med* 30:850-862, 2024
- Chen PT, Wu T, Wang P, et al: Pancreatic cancer detection on CT scans with deep learning: A nationwide population-based study. *Radiology* 306:172-182, 2023
- Schirmer L, Velmeshev D, Holmqvist S, et al: Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* 573:75-82, 2019
- Collins RL, Glessner JT, Porcu E, et al: A cross-disorder dosage sensitivity map of the human genome. *Cell* 185:3041-3055.e25, 2022
- Lau SF, Cao H, Fu AKY, et al: Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. *Proc Natl Acad Sci U S A* 117:25800-25809, 2020
- Leng K, Li E, Eser R, et al: Molecular characterization of selectively vulnerable neurons in Alzheimer's disease. *Nat Neurosci* 24:276-287, 2021
- Smajić S, Prada-Medina CA, Landoulsi Z, et al: Single-cell sequencing of human midbrain reveals glial activation and a Parkinson-specific neuronal state. *Brain* 145:964-978, 2022
- Zhu B, Park JM, Coffey SR, et al: Single-cell transcriptomic and proteomic analysis of Parkinson's disease brains. *Sci Transl Med* 16:eabo1997, 2024

46. Ho YJ, Anaparthi N, Molik D, et al: Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. *Genome Res* 28:1353-1363, 2018
47. Kinker GS, Greenwald AC, Tal R, et al: Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat Genet* 52:1208-1218, 2020
48. Simonson B, Chaffin M, Hill MC, et al: Single-nucleus RNA sequencing in ischemic cardiomyopathy reveals common transcriptional profile underlying end-stage heart failure. *Cell Rep* 42:112086, 2023
49. Suo C, Dann E, Goh I, et al: Mapping the developing human immune system across organs. *Science* 376:eabo0510, 2022
50. Lin Y, Cao Y, Kim HJ, et al: scClassify: Sample size estimation and multiscale classification of cells using single and multiple reference. *Mol Syst Biol* 16:e9389, 2020
51. Quach H, Farrell S, Wu MJM, et al: Early human fetal lung atlas reveals the temporal dynamics of epithelial cell plasticity. *Nat Commun* 15:5898, 2024
52. Kumar T, Nee K, Wei R, et al: A spatially resolved single-cell genomic atlas of the adult human breast. *Nature* 620:181-191, 2023
53. Kim CN, Shin D, Wang A, et al: Spatiotemporal molecular dynamics of the developing human thalamus. *Science* 382:eadf9941, 2023
54. Maynard KR, Collado-Torres L, Weber LM, et al: Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* 24:425-436, 2021
55. Bulten W, Kartasalo K, Chen PHC, et al: Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nat Med* 28:154-163, 2022
56. Szalata A, Hrovatin K, Becker S, et al: Transformers in single-cell omics: A review and new perspectives. *Nat Methods* 21:1430-1443, 2024
57. Hao Y, Rao S, Andersen-Nissen E, et al: Integrated analysis of multimodal single-cell data. *Cell* 184:3573-3587.e29, 2021
58. Hodge RD, Bakken TE, Miller JA, et al: Conserved cell types with divergent features in human versus mouse cortex. *Nature* 573:61-68, 2019
59. Lopez R, Regier J, Cole MB, et al: Deep generative modeling for single-cell transcriptomics. *Nat Methods* 15:1053-1058, 2018
60. Korsunsky I, Millard N, Fan J, et al: Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 16:1289-1296, 2019
61. Philip M, Schietinger A: CD8+ T cell differentiation and dysfunction in cancer. *Nat Rev Immunol* 22:209-223, 2022
62. Van de Sande B, Lee JS, Mutasa-Gottgens E, et al: Applications of single-cell RNA sequencing in drug discovery and development. *Nat Rev Drug Discov* 22:496-520, 2023
63. Zeng J, Zhang Y, Shang Y, et al: CancerSCEM: A database of single-cell expression map across various human cancers. *Nucleic Acids Res* 50:D1147-D1155, 2022
64. Mo CK, Liu J, Chen Q, et al: Tumour evolution and microenvironment interactions in 2D and 3D space. *Nature* 634:1178-1186, 2024
65. Cui Zhou D, Jayasinghe RG, Chen S, et al: Spatially restricted drivers and transitional cell populations cooperate with the microenvironment in untreated and chemo-resistant pancreatic cancer. *Nat Genet* 54:1390-1405, 2022
66. Ma C, Balaban M, Liu J, et al: Inferring allele-specific copy number aberrations and tumor phylogeography from spatially resolved transcriptomics. *Nat Methods* 21:2239-2247, 2024
67. Hu J, Coleman K, Zhang D, et al: Deciphering tumor ecosystems at super resolution from spatial transcriptomics with TESLA. *Cell Syst* 14:404-417.e4, 2023
68. Song Q, Su J: DSTG: Deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief Bioinform* 22:bbaa414, 2021
69. Dong K, Zhang S: Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat Commun* 13:1739, 2022
70. Li B, Tang Z, Budhkar A, et al: SpaLM: Single-cell spatial transcriptomics imputation via style transfer. *bioRxiv* 10.1101/2025.01.24.634756
71. Radford A, Kim JW, Hallacy C, et al: Learning transferable visual models from natural language supervision. *arXiv* 10.48550/arXiv.2103.00020
72. Huang Z, Bianchi F, Yuksekogonul M, et al: A visual-language foundation model for pathology image analysis using medical Twitter. *Nat Med* 29:2307-2316, 2023
73. Pentimalli TM, Karaiskos N, Rajewsky N: Challenges and opportunities in the clinical translation of high-resolution spatial transcriptomics. *Annu Rev Pathol* 20:405-432, 2025
74. Zhao L, Lee VHF, Ng MK, et al: Molecular subtyping of cancer: Current status and moving toward clinical applications. *Brief Bioinform* 20:572-584, 2019
75. Vorontsov E, Bozkurt A, Casson A, et al: A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat Med* 30:2924-2935, 2024
76. Wang X, Yang S, Zhang J, et al: Transformer-based unsupervised contrastive learning for histopathological image classification. *Med Image Anal* 81:102559, 2022
77. Tjota MY, Segal JP, Wang P: Clinical utility and benefits of comprehensive genomic profiling in cancer. *J Appl Lab Med* 9:76-91, 2024
78. Chen RJ, Chen C, Li Y, et al: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp 16123-16134
79. Zhang S, Xu Y, Usuyama N, et al: BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv* 10.48550/arXiv.2303.00915
80. Zhang K, Zhou R, Adhikarla E, et al: A generalist vision-language foundation model for diverse biomedical tasks. *Nat Med* 30:3129-3141, 2024
81. Sushil M, Zack T, Mandair D, et al: A comparative study of large language model-based zero-shot inference and task-specific supervised classification of breast cancer pathology reports. *J Am Med Inform Assoc* 31:2315-2327, 2024
82. OpenAI: GPT-4 technical report. *arXiv* 10.48550/arXiv.2303.08774
83. OpenAI: ChatGPT (GPT-3.5, March 23 version) [large language model], 2023. <https://openai.com/chatgpt>
84. Hochreiter S, Schmidhuber J: Long short-term memory. *Neural Comput* 9:1735-1780, 1997
85. Huemann Z, Lee C, Hu J, et al: Domain-adapted large language models for classifying nuclear medicine reports. *Radiol Artif Intell* 5:e220281, 2023
86. Wang Q, Pan Y, Zhou M, et al: scDrugMap: Benchmarking large foundation models for drug response prediction. *arXiv* 10.48550/arXiv.2505.05612
87. Niu C, Lyu Q, Carothers CD, et al: Medical multimodal multitask foundation model for lung cancer screening. *Nat Commun* 16:1523, 2025
88. Liu H, Chen S, Zhang Y, et al: GenoTEX: An LLM agent benchmark for automated gene expression data analysis. *arXiv* 10.48550/arXiv.2406.15341
89. Steele CD, Abbasi A, Islam SMA, et al: Signatures of copy number alterations in human cancer. *Nature* 606:984-991, 2022
90. He S, Bhatt R, Brown C, et al: High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat Biotechnol* 40:1794-1806, 2022
91. Williams CG, Lee HJ, Asatsuma T, et al: An introduction to spatial transcriptomics for biomedical research. *Genome Med* 14:68, 2022
92. Wang A, Singh A, Michael J, et al: GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv* 10.48550/arXiv.2410.08385
93. Sarlin PE, DeTone D, Malisiewicz T, et al: SuperGlue: Learning feature matching with graph neural networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp 4937-4946
94. Russakovsky O, Deng J, Su H, et al: ImageNet large scale visual recognition challenge. *Int J Comput Vis* 115:211-252, 2015
95. Banerjee S, Agarwal A, Singh E: The vulnerability of language model benchmarks: Do they accurately reflect true LLM performance? *arXiv* 10.48550/arXiv.2412.03597
96. US Food and Drug Administration: Good machine learning practice for medical device development: guiding principles. FDA webpage, 2021
97. Weissman GE, Mankowitz T, Kanter GP: Unregulated large language models produce medical device-like output. *NPJ Digit Med* 8:148, 2025
98. Adamson B, Norman TM, Jost M, et al: A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* 167:1867-1882, 2016
99. Replogle JM, Saunders RA, Pogson AN, et al: Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* 185:2559-2575, 2022
100. Norman TM, Horlbeck MA, Replogle JM, et al: Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* 365:786-793, 2019