nature portfolio

Peer Review File

A Multimodal Knowledge-enhanced Whole-slide Pathology Foundation Model

Corresponding Author: Professor Hao Chen

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Communications.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

My comments on the authors' replies to Reviewer 1's comments

Response to the Summary comment

I agree with Reviewer 1's comment about the novelty of the study being limited.

I think the authors' 1st point on multimodality is relatively weak, as all those mentioned modalities, ie pathology reports and gene expression profiles, have been used for help at slide/patient level tasks.

I do agree with the author's 2nd point that allowing the tile encoder to be updated with slide/patient level pathology reports and gene expression profiles is a novel contribution. However, I do find it difficult to understand the evaluation of contrastive learning for pathology reports and gene expression profiles. As far as I understand, pathology reports do not usually contain information about gene expression.

As for the 3rd point about the task, I do appreciate the added amount of work. I think authors are very thorough compared to existing pathology foundation model papers. I do have concerns over the usability of some clinical end points, like disease-free survival, which is often study and cohort-dependent. It could give a false sense of generality.

I think the authors have addressed comments 2-7,9,10.

As for comment 8, I find it confusing to include pathology reports for tasks like staging, grading and subtyping. The final labels in those tasks are likely to be stated in the pathology report already.

To my comment on novelty, the authors give the same reply as to Reviewer 1's question. Please see comments above.

The authors have addressed all my comments.

Minor question

What do the authors mean by "microscopic-level pathology slides" as an additional modality? I can see pathology reports and gene expression profiles being used to train the slide-level model. I can not link "microscopic-level pathology slides" to the training framework presented in the paper.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
We thank the authors for their detailed response and commend the improved clarity and expanded ablation studies in the revised manuscript. The new experiments systematically assess different combinations of pre-training modalities: pathology images, pathology reports, and RNA-Seq data. Notably, the results indicate that for mSTAR, pre-training on pathology alone performs comparably to combinations with either reports or RNA-Seq. A marginal performance gain of +0.01 in Avg C-Index is observed only when all three modalities are combined. The authors interpret this as evidence of scalable multimodal integration. But the tremendous pre-training effort that is required does not justify to use the mSTAR method compared to other models like UNI or CONCH.
Previous literature has already demonstrated that dual-modality pre-training (e.g., pathology + reports or pathology + RNA-Seq) can outperform unimodal pre-training. Thus, the added value of including a third modality with the specialized slide-level contrastive learning and fine-tuning applied in this study remains insufficiently substantiated.
A key limitation of this work is the entanglement of two experimental variables: the introduction of a third pre-training modality and a novel pre-training scheme involving fine-tuning of the image encoder (UNI). In contrast, models such as GigaPath, Tangle, and CHIEF rely on frozen feature extractors, allowing clearer attribution of performance gains to specific design choices. Without a controlled disentanglement of these factors, it becomes difficult to isolate the contribution of each component to the final performance.
Furthermore, the study lacks an investigation into scaling laws with respect to model size, which are essential for understanding the cost-benefit trade-offs of integrating additional modalities. Given the minimal performance improvements reported, a detailed analysis of training efficiency and computational overhead is necessary to assess the practical utility of the proposed approach.
Finally, if the goal is to evaluate modality scaling, we encourage the authors to consider broader avenues beyond RNA-Seq, such as including IHC or other specialized stainings (e.g., https://arxiv.org/abs/2408.02859), which have shown promising results in conjunction with scaled transformer architectures.
In summary, while the manuscript presents a technically sound and ambitious effort, the performance improvements remain marginal and lack sufficient justification given the added complexity. Clarification and additional analyses are required to establish the value of the proposed multimodal pre-training strategy.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
I want thank the authors for throughly addressing all my comments. I have no further concerns with the manuscript.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I thank the authors for their efforts in addressing my concerns. All concerns have been addressed.

(Remarks on code availability)

Dear reviewers,

We sincerely appreciate your acknowledgement of the novelty of approach and the opportunity to submit a revised version of our manuscript. We have carefully addressed all the remaining points raised during the second round of review, majorly including **1) re-illustration of novelty, 2) investigation of impacts of various modalities, and 3) exploration in the efficiency and effectiveness of modality scaling**. The revision of manuscript has been highlighted in red for reviewers' convenience.

We believe that these new revisions adequately address the reviewers' concerns. This study could provide valuable contributions to the community of computational pathology.

Sincerely,

Authors of the manuscript

Next, we will offer the response to every reviewer point by point, where our response to your comments is marked in blue.

Our response to Reviewer#2:

1. I think the authors' 1st point on multimodality is relatively weak, as all those mentioned modalities, ie pathology reports and gene expression profiles, have been used for help at slide/patient level tasks.

Response:

We appreciate the reviewer’s feedback regarding the novelty of our study. We respectfully highlight two key aspects that distinguish our work from prior research:

1. **Novelty in Multimodal Alignment and Scalability for Pathology Foundation Models**: While previous studies have indeed explored combinations of gene expression profiles [1] for slide/patient-level tasks, they were limited to two modalities and focused primarily on slide/patient-level aggregation. While TITAN [2] (29 Nov 2024 UTC) and PRISM2 [3] (16 Jun 2025 UTC) also used pathology reports (vision-language models) in their pathology foundation models, they were released later than our work (22 Jul 2024 UTC) and have not yet been formally published.
   Our work is the first and only one to seamlessly align and validate the integration of three distinct modalities: (e.g., pathology images, pathology reports and gene expression profiles) within a unified framework. Crucially, we demonstrate the scalability of this approach—showing that adding more modalities (beyond two modalities) consistently improves performance. This scalability, empirically validated in our experimental results (Section 2), has not been previously explored and offers a meaningful advance for multimodal learning in pathology foundation models, where multimodal data is abundant but underutilized.
2. **Impact on Patch Extractor, Not Just Slide Aggregator**: Related works [1-2] incorporating additional modalities (e.g., gene expression profiles), focused primarily on the role of aggregators. However, our study uniquely demonstrates that these modalities also enhance the feature extractor itself, by comparing performances of UNI [14] (baseline, before pretraining patch extractor) and mSTAR in Section 2. Crucially, we highlight that ***improving the feature extractor (e.g., ViT-based encoder) offers far greater scalability potential than optimizing aggregators***. In computational pathology, aggregators are typically shallow (e.g., 1–2 MLP layers) [4] or use lightweight linear transformers [5] to handle thousands of patches efficiently. ***This limited parameters inherently limits their capacity to absorb multimodal information during pretraining.*** In contrast, feature extractors (e.g., ViT) have orders of magnitude more parameters, enabling more powerful abilities in multimodal information integration. This scaling law has has been validated in natural vision and language domains [6-7]. Furthermore, since slide-level multimodal self-supervised signals fail to guide patch-level feature extraction, ***the pretraining objectives' misalignment of these two independent stages inevitably results in suboptimal performance.***

[1] Jaume G, Oldenburg L, Vaidya A, et al. Transcriptomics-guided slide representation learning in computational pathology[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 9632-9644.

[2] Ding T, Wagner S J, Song A H, et al. Multimodal whole slide foundation model for pathology[J]. arXiv preprint arXiv:2411.19666, 2024.

[3] Shaikovski G, Vorontsov E, Casson A, et al. PRISM2: Unlocking Multi-Modal General Pathology AI with Clinical Dialogue[J]. arXiv preprint arXiv:2506.13063, 2025.

[4] Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning[C]//International conference on machine learning. PMLR, 2018: 2127-2136.

[5] Shao Z, Bian H, Chen Y, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification[J]. Advances in neural information processing systems, 2021, 34: 2136-2147.

[6] Oquab M, Darcet T, Moutakanni T, et al. Dinov2: Learning robust visual features without supervision[J]. arXiv preprint arXiv:2304.07193, 2023.

[7] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.

2. I do agree with the author's 2nd point that allowing the tile encoder to be updated with slide/patient level pathology reports and gene expression profiles is a novel contribution. However, I do find it difficult to understand the evaluation of contrastive learning for pathology reports and gene expression profiles. As far as I understand, pathology reports do not usually contain information about gene expression.

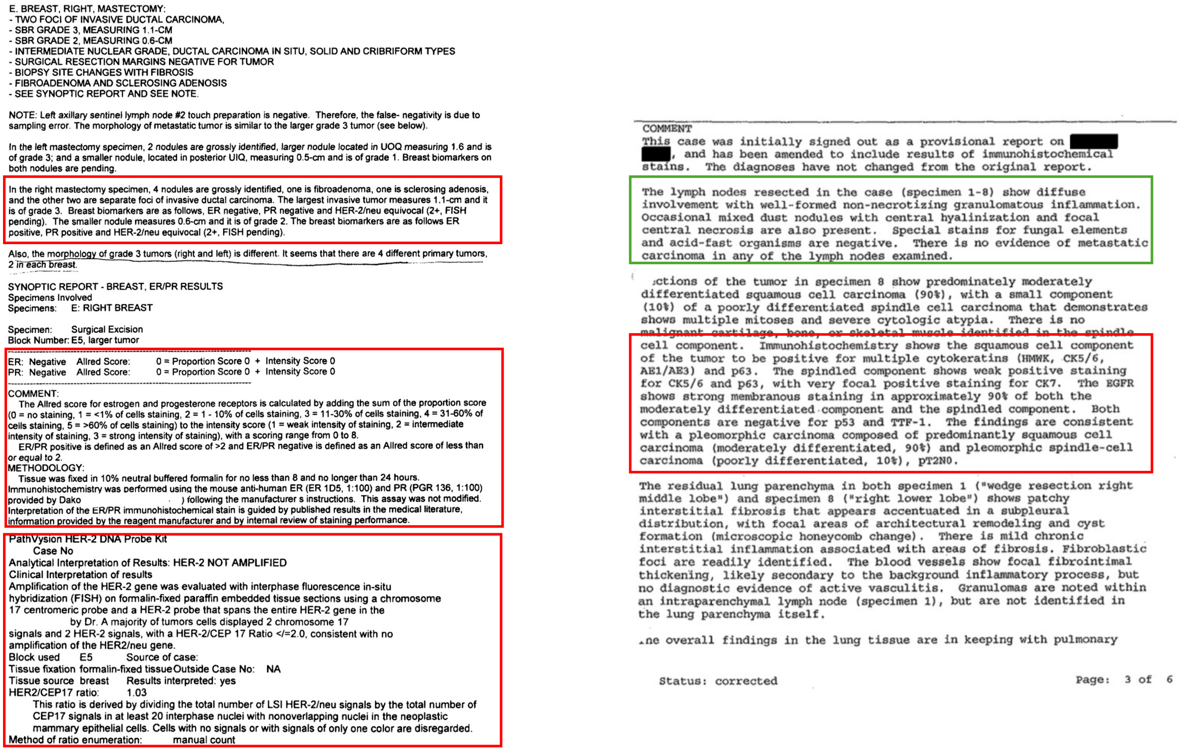
Response:

We thank the reviewer for raising this important point. We clarify the intrinsic link between pathology reports and gene expression through two key aspects supported by clinical practice and biological evidence:

1. **Pathology reports routinely include molecular profiling data**. Standard pathology reports (especially for cancer diagnostics) frequently incorporate molecular testing results such as: Immunohistochemistry (IHC) markers (e.g., ER/PR/HER2 status in breast cancer, PD-L1 in lung cancer), targeted gene sequencing (e.g., EGFR, KRAS mutations), in-situ hybridization (ISH) for gene amplification (e.g., HER2 FISH), etc. As shown in Figure 1, two examples of pathology reports used in pretraining explicitly documenting these molecular information highlighted in the red box, directly reflecting gene expression phenotypes.
2. **Morphological features in histology are biomarkers of gene activity**. The connection between histopathology and gene expression is biologically fundamental: (1) IHC visualizes protein expression products of specific genes (e.g., HER2 protein overexpression from ERBB2 gene amplification [8]). (2) Histomorphological patterns included in reports (e.g., granulomatous inflammation in lung cancer in the green box of Figure 1) correlate with upregulation of checkpoint molecules (e.g., PD-L1 [9-10]).

Thus, pathology reports establish phenotype-genotype correlations through both direct molecular profiling and morphology-derived biomarkers.

[8] Yoon H H, Shi Q, Sukov W R, et al. Association of HER2/ErbB2 expression and gene amplification with pathologic features and prognosis in esophageal adenocarcinomas[J]. Clinical Cancer Research, 2012, 18(2): 546-554.
[9] Braun N A, Celada L J, Herazo-Maya J D, et al. Blockade of the programmed death-1 pathway restores sarcoidosis CD4+ T-cell proliferative capacity[J]. American journal of respiratory and critical care medicine, 2014, 190(5): 560-571.
[10] Cornejo C M, Haun P, English III J, et al. Immune checkpoint inhibitors and the development of granulomatous reactions[J]. Journal of the American Academy of Dermatology, 2019, 81(5): 1165-1175.



E. BREAST, RIGHT, MASTECTOMY:
- TWO FOCI OF INVASIVE DUCTAL CARCINOMA,
- SBR GRADE 3, MEASURING 1.1-CM
- SBR GRADE 2, MEASURING 0.6-CM
- INTERMEDIATE NUCLEAR GRADE, DUCTAL CARCINOMA IN SITU, SOLID AND CRIBRIFORM TYPES
- SURGICAL RESECTION MARGINS NEGATIVE FOR TUMOR
- BIOPSY SITE CHANGES WITH FIBROSIS
- FIBROADENOMA AND SCLEROSING ADENOSIS
- SEE SYNOPTIC REPORT AND SEE NOTE.

NOTE: Left axillary sentinel lymph node #2 touch preparation is negative. Therefore, the false- negativity is due to sampling error. The morphology of metastatic tumor is similar to the larger grade 3 tumor (see below).

In the left mastectomy specimen, 2 nodules are grossly identified, larger nodule located in UOQ measuring 1.6 and is of grade 3; and a smaller nodule, located in posterior UIQ, measuring 0.5-cm and is of grade 1. Breast biomarkers on both nodules are pending.

In the right mastectomy specimen, 4 nodules are grossly identified, one is fibroadenoma, one is sclerosing adenosis, and the other two are separate foci of invasive ductal carcinoma. The largest invasive tumor measures 1.1-cm and it is of grade 3. Breast biomarkers are as follows, ER negative, PR negative and HER-2/neu equivocal (2+, FISH pending). The smaller nodule measures 0.6-cm and it is of grade 2. The breast biomarkers are as follows ER positive, PR positive and HER-2/neu equivocal (2+, FISH pending).

Also, the morphology of grade 3 tumors (right and left) is different. It seems that there are 4 different primary tumors, 2 in each breast.

SYNOPTIC REPORT - BREAST, ER/PR RESULTS
Specimens Involved
Specimens: E: RIGHT BREAST

Specimen: Surgical Excision
Block Number: E5, larger tumor

ER: Negative Allred Score: 0 = Proportion Score 0 + Intensity Score 0
PR: Negative Allred Score: 0 = Proportion Score 0 + Intensity Score 0

COMMENT:
The Allred score for estrogen and progesterone receptors is calculated by adding the sum of the proportion score (0 = no staining, 1 = <1% of cells staining, 2 = 1 - 10% of cells staining, 3 = 11-30% of cells staining, 4 = 31-60% of cells staining, 5 = >60% of cells staining) to the intensity score (1 = weak intensity of staining, 2 = intermediate intensity of staining, 3 = strong intensity of staining), with a scoring range from 0 to 8.
ER/PR positive is defined as an Allred score of >2 and ER/PR negative is defined as an Allred score of less than or equal to 2.
METHODOLOGY:
Tissue was fixed in 10% neutral buffered formalin for no less than 8 and no longer than 24 hours. Immunohistochemistry was performed using the mouse anti-human ER (ER 1D5, 1:100) and PR (PGR 136, 1:100) provided by Dako) following the manufacturer s instructions. This assay was not modified. Interpretation of the ER/PR immunohistochemical stain is guided by published results in the medical literature, information provided by the reagent manufacturer and by internal review of staining performance.

PathVysion HER-2 DNA Probe Kit
Case No
Analytical Interpretation of Results: HER-2 NOT AMPLIFIED
Clinical Interpretation of results
Amplification of the HER-2 gene was evaluated with interphase fluorescence in-situ hybridization (FISH) on formalin-fixed paraffin embedded tissue sections using a chromosome 17 centromeric probe and a HER-2 probe that spans the entire HER-2 gene in the by Dr. A majority of tumors cells displayed 2 chromosome 17 signals and 2 HER-2 signals, with a HER-2/CEP 17 Ratio </=2.0, consistent with no amplification of the HER2/neu gene.
Block used E5 Source of case:
Tissue fixation formalin-fixed tissue Outside Case No: NA
Tissue source breast Results interpreted: yes
HER2/CEP17 ratio: 1.03
This ratio is derived by dividing the total number of LSI HER-2/neu signals by the total number of CEP17 signals in at least 20 interphase nuclei with nonoverlapping nuclei in the neoplastic mammary epithelial cells. Cells with no signals or with signals of only one color are disregarded.
Method of ratio enumeration: manual count

COMMENT
This case was initially signed out as a provisional report on [redacted], and has been amended to include results of immunohistochemical stains. The diagnoses have not changed from the original report.

The lymph nodes resected in the case (specimen 1-8) show diffuse involvement with well-formed non-necrotizing granulomatous inflammation. Occasional mixed dust nodules with central hyalinization and focal central necrosis are also present. Special stains for fungal elements and acid-fast organisms are negative. There is no evidence of metastatic carcinoma in any of the lymph nodes examined.

.ctions of the tumor in specimen 8 show predominately moderately differentiated squamous cell carcinoma (90%), with a small component (10%) of a poorly differentiated spindle cell carcinoma that demonstrates shows multiple mitoses and severe cytologic atypia. There is no malignant cartilage, bone, or skeletal muscle identified in the spindle cell component. Immunohistochemistry shows the squamous cell component of the tumor to be positive for multiple cytokeratins (HMWK, CK5/6, AE1/AE3) and p63. The spindled component shows weak positive staining for CK5/6 and p63, with very focal positive staining for CK7. The EGFR shows strong membranous staining in approximately 90% of both the moderately differentiated component and the spindled component. Both components are negative for p53 and TTF-1. The findings are consistent with a pleomorphic carcinoma composed of predominantly squamous cell carcinoma (moderately differentiated, 90%) and pleomorphic spindle-cell carcinoma (poorly differentiated, 10%), pT2N0.

The residual lung parenchyma in both specimen 1 ("wedge resection right middle lobe") and specimen 8 ("right lower lobe") shows patchy interstitial fibrosis that appears accentuated in a subpleural distribution, with focal areas of architectural remodeling and cyst formation (microscopic honeycomb change). There is mild chronic interstitial inflammation associated with areas of fibrosis. Fibroblastic foci are readily identified. The blood vessels show focal fibrointimal thickening, likely secondary to the background inflammatory process, but no diagnostic evidence of active vasculitis. Granulomas are noted within an intraparenchymal lymph node (specimen 1), but are not identified in the lung parenchyma itself.

.ne overall findings in the lung tissue are in keeping with pulmonary

Status: corrected Page: 3 of 6

(a) TCGA-E2-A158 from TCGA-BRCA

(b) TCGA-18-4086 from TCGA-LUSC

Figure 1. **Examples of pathology reports**. (a) from breast cancer and (b) from lung cancer, where the text highlighted in the red box demonstrates the connection between pathology reports and gene expression profiles, and the one in the green box showcases the morphological descriptions in pathology reports.

3. As for the 3rd point about the task, I do appreciate the added amount of work. I think authors are very thorough compared to existing pathology foundation model papers. I do have concerns over the usability of some clinical end points, like disease-free survival, which is often study and cohort-dependent. It could give a false sense of generality.

Response:

We sincerely appreciate the reviewer's recognition of our comprehensive evaluation framework and their thoughtful concerns regarding the clinical endpoints. We fully acknowledge that endpoints like disease-free survival (DFS) and recurrence-free survival (RFS) can indeed vary across studies, and all endpoints in our study were chosen based on their broad usage in clinical oncology and alignment with public datasets, e.g. DFS for breast cancer and gastric cancer, and RFS for head and neck cancer. We believe this preserves the clinical utility that makes our findings meaningful for both researchers and practicing oncologists.

4. As for comment 8, I find it confusing to include pathology reports for tasks like staging, grading and subtyping. The final labels in those tasks are likely to be stated in the pathology report already.

Response:

Sorry for confusion, and we appreciate the reviewer’s attention to evaluation validity. We want to clarify that the data used for downstream evaluation were rigorously excluded from pretraining data, and pathology reports were only used during pretraining (for multimodal alignment with WSIs). All downstream tasks were evaluated using H&E diagnostic slides only as the input with no text data. This aligns with standard foundation model paradigms (e.g., CONCH [11], where text aids pretraining but isn’t available during inference). This clarification has been added in Methods section 4.3.

[11] Lu M Y, Chen B, Williamson D F K, et al. A visual-language foundation model for computational pathology[J]. Nature medicine, 2024, 30(3): 863-874.

5. Minor question:

What do the authors mean by “microscopic-level pathology slides” as an additional modality? I can see pathology reports and gene expression profiles being used to train the slide-level model. I can not link “microscopic-level pathology slides” to the training framework presented in the paper.

Response: We sincerely appreciate the reviewer's careful reading and this opportunity to clarify our terminology. We acknowledge that the phrase "microscopic-level pathology slides" in our original manuscript caused confusion, and we have removed this potentially ambiguous expression in our revised version. The term "microscopic-level" was intended to suggest that histopathology slides are examined under microscopes during clinical diagnosis, and we have removed it in the revised manuscript.

Our response to Reviewer#3:

1. We thank the authors for their detailed response and commend the improved clarity and expanded ablation studies in the revised manuscript. The new experiments systematically assess different combinations of pre-training modalities: pathology images, pathology reports, and RNA-Seq data. Notably, the results indicate that for mSTAR, pre-training on pathology alone performs comparably to combinations with either reports or RNA-Seq. A marginal performance gain of +0.01 in Avg C-Index is observed only when all three modalities are combined. The authors interpret this as evidence of scalable multimodal integration. But the tremendous pre-training effort that is required does not justify to use the mSTAR method compared to other models like UNI or CONCH.

Previous literature has already demonstrated that dual-modality pre-training (e.g., pathology + reports or pathology + RNA-Seq) can outperform unimodal pre-training. Thus, the added value of including a third modality with the specialized slide-level contrastive learning and fine-tuning applied in this study remains insufficiently substantiated.

Response:

We thank the reviewer for engaging deeply with our revised analyses. We address the concerns about multimodal scalability through new evidence that substantiates the unique advantages of our approach:

1) **The impact of multimodal data on performance gains differs significantly depending on the task type**. While the previous version only reported average overall survival (OS) results across 9 TCGA datasets, in this revised manuscript we have systematically evaluated proportional datasets for each task category (24 datasets), as shown in Figure 2a.

The results demonstrate consistent performance improvements across all types of tasks, with a notable average increase of 5.3% specifically observed in the 12 molecular prediction tasks compared to that of only 2 modalities. More interestingly, we found that text modalities outperform genomic data in diagnostic tasks, which intuitively corresponds to the diagnostic-rich nature of pathology reports. For molecular and prognostic predictions, textual and genomic data demonstrated comparable performance, consistent with prior findings [12] that both phenotype (text-derived) and genotype features contribute to stratification.

Given the substantial heterogeneity across cancer types and varying prediction difficulty among tasks, we found that average performance metrics were insufficient to reflect true improvements. Therefore, we also present detailed task-specific performance for each dataset in Figure 3. In 21 out of 24 datasets, the three-modality models demonstrated consistent performance improvements over dual-modality combinations.

2) **mSTAR's design is highly resource-efficient, while achieve comparable performance with SOTA pathology foundation models**.

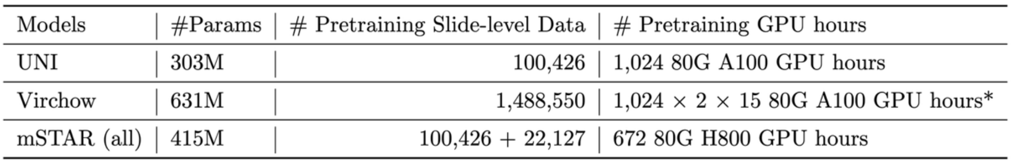
- Parameter-wise, compared to Virchow [13] of 631M, a SOTA pathology foundation model, mSTAR merely consists of 415M (34% reduction) parameters including the 303M visual encoder (the same as UNI [14]), 2.67M TransMIL module, 0.94M genomic Performer, and 108M BioBERT text encoder (smaller than CONCH [11]).
- Data-wise, when benchmarked against UNI's performance baseline, we compared mSTAR with Virchow as shown in Figure 2d. mSTAR achieves competitive improvements with only 22K additional slide-level pretraining samples—a far smaller

fraction of the 1.39M extra slides required by Virchow for better performance gains. This 53× reduction in additional data demonstrates that multimodal data brought higher efficiency per sample compared to vision-only brute-force data scaling, significantly reducing pretraining costs in GPU-hours while even enhancing clinical-grade accuracy. Our findings offer a remarkable advantage and a practical pathway in scaling pathology foundation models, especially for resource-constrained medical AI development where large-scale data collection is often impractical.

- Training-wise, while UNI requires 4x8 A100 80GB GPUs (4 nodes, each with 8 GPUs) for 32 GPU hours (1024 GPU hours in total), mSTAR merely needs 4 H800 80GB GPUs (1 node, each with 4 GPUs) for 7 days (672 GPU hours in total). While Virchow does not report exact training durations, its substantially larger model size (631M vs. our 414M parameters) and 53× greater pretraining data (1.39M vs. our 22K slides) inevitably require far greater computational resources. More comparison can be seen in Table 1 and 2.

The improvement is substantial in the current context. mSTAR achieves Virchow's performance with just 2% of its additional pretraining data—a clear indication that the achieved improvements are not only substantial but also highly efficient. This reveals that multimodal scaling yields substantially greater returns than large-scale vision-only expansion, effectively liberating the field of PFMs from reliance on massive slide collections.

Table 1, **Resources Comparison** between scaling slides only (Virchow) v.s. scaling modalities (mSTAR) for pretraining, with UNI as a baseline. * means that the pretraining GPU hours are coarsely estimated based on that of UNI, since it didn't report such pretraining details.



Models	#Params	# Pretraining Slide-level Data	# Pretraining GPU hours
UNI	303M	100,426	1,024 80G A100 GPU hours
Virchow	631M	1,488,550	1,024 × 2 × 15 80G A100 GPU hours*
mSTAR (all)	415M	100,426 + 22,127	672 80G H800 GPU hours

Table 2, **Performance Comparison** between scaling slides only (Virchow) v.s. scaling modalities (mSTAR) for pretraining, with UNI as a baseline. The best-performing model for each metric is bolded. Std is given by bootstrapping with 1,000 bootstraps. Both Virchow and mSTAR underwent one-sided Wilcoxon signed-rank tests against the baseline UNI. For results outperforming the baseline, all unmarked (*) ones exhibited statistically significant differences at $P < 0.001$.

Task	Dataset	UNI	Virchow	mSTAR
Pathlogical Diagnosis	CAMELYON (idpt)	0.9819±0.0184	0.9683±0.0248	**0.9935±0.0098**
Pathlogical Diagnosis	NFGC_Perineural (idpt)	0.9750±0.0349	0.9270±0.0685	**0.9776±0.0316**
Pathlogical Diagnosis	BRCA-PathSubtype (out)	0.9391±0.0489	**0.9449±0.0526**	0.9314±0.0510
Molecular Prediction	BRCA_PIK3CA (out)	0.6969±0.0315	0.6669±0.0492	**0.7215±0.0318**
Molecular Prediction	TCGA_BRCA_HER2 (out)	0.7636±0.0305	0.7287±0.0506	**0.7679±0.0310**
Molecular Prediction	TCGA_BRCA_PR (out)	0.5028±0.0195	0.5080±0.0258	**0.5254±0.0183**
Molecular Prediction	BRCA_MolSubtype (out)	0.7756±0.0219	0.7659±0.0301	**0.8057±0.0191**
Molecular Prediction	IHC_ZJ1_HER2_Level (idpt)	0.7758±0.0132	0.7628±0.0196	**0.7951±0.0125**
Molecular Prediction	IHC_ZJ1_ER_Level (idpt)	0.7892±0.0111	0.7960±0.0152	**0.8020±0.0107**
Molecular Prediction	IHC_ZJ1_HER2 (ext)	0.6153±0.0173	**0.6558±0.0226**	0.6429±0.0162
Molecular Prediction	IHC_ZJ1_PR (ext)	0.5490±0.0052	**0.5733±0.0063**	0.5673±0.0047
Molecular Prediction	ZJ1_Breast_MolSubtype (ext)	0.7844±0.0046	0.7582±0.0066	**0.7946±0.0045**
Survival Prediction	OS_BRCA (out)	0.6908±0.1048	0.6525±0.0506	**0.7076±0.0896**
Survival Prediction	OS_CRC (out)	**0.6906±0.0835**	0.6140±0.0577	0.6895±0.0836
Survival Prediction	OS_GBMLGG (out)	0.7905±0.0434	0.7790±0.0206	**0.7923±0.0426**
Survival Prediction	OS_HNSC (out)	0.6516±0.0798	0.6000±0.0381	**0.6604±0.0794**
Survival Prediction	OS_KIRC (out)	**0.7155±0.0659**	0.6967±0.0395	0.7027±0.0890
Survival Prediction	OS_LUAD (out)	0.6312±0.0996	0.6194±0.0457	**0.6329±0.0976**
Survival Prediction	OS_LUSC (out)	0.6273±0.0771	0.5382±0.0463	**0.6323±0.0785**
Survival Prediction	OS_SKCM (out)	0.6254±0.0776	0.6207±0.0423	**0.6281±0.0761**
Survival Prediction	OS_UCEC (out)	0.7845±0.1012	0.7477±0.0556	**0.8092±0.0865**

[12] Cuny M, Kramar A, Courjal F, et al. Relating genotype and phenotype in breast cancer: an analysis of the prognostic significance of amplification at eight different genes or loci and of p53 mutations[J]. Cancer research, 2000, 60(4): 1077-1083.

[13] Vorontsov E, Bozkurt A, Casson A, et al. A foundation model for clinical-grade computational pathology and rare cancers detection[J]. Nature medicine, 2024, 30(10): 2924-2935.

[14] Chen R J, Ding T, Lu M Y, et al. Towards a general-purpose foundation model for computational pathology[J]. Nature medicine, 2024, 30(3): 850-862.

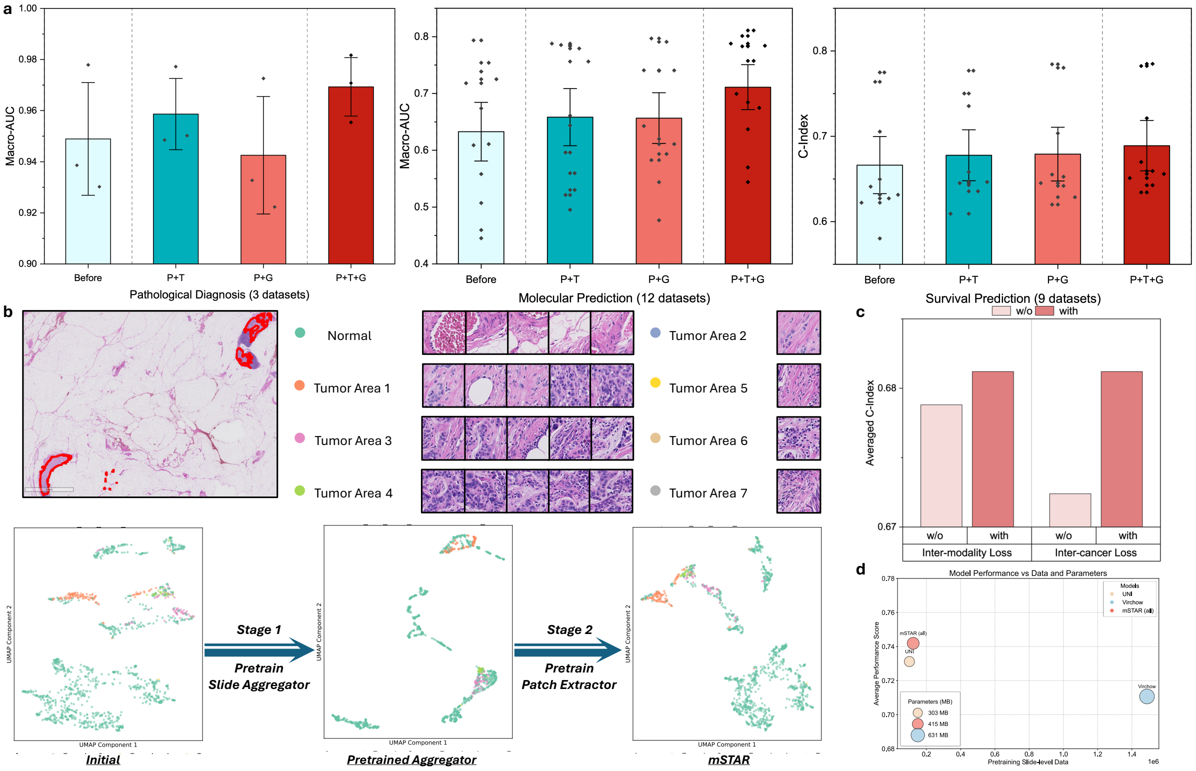


Figure 2. **Ablation Studies**. **a**, averaged performance on pathological diagnosis (3 datasets), molecular prediction (12 datasets) and survival prediction (9 datasets) , where `Before' refers to before pretraining, and `P', `T' and `G' indicate pathology slides, pathology reports and gene data, respectively. **b**, visualization of feature space evolution: from before pretraining (initial) to Stage 1 (pretrained aggregator) and Stage 2 (mSTAR), where the areas in red bounding box are multiple tumor regions (1-7) of the case of patient_042_node_3 of CAMELYON17 dataset. Note that different tumor areas correspond to different spatial positions. **c**, averaged performance (9 TCGA OS datasets) for ablating different pretraining objectives (Inter-modal Loss and Inter-cancer Loss) for survival prediction. **d**, averaged performance (24 datasets) and resources comparisons between scaling slides only (Virchow) v.s. scaling modalities (mSTAR) for pretraining, with UNI as a baseline. Detailed performances of every dataset are presented in Figure 3 and detailed comparisons are showcased in Table 1 and 2.

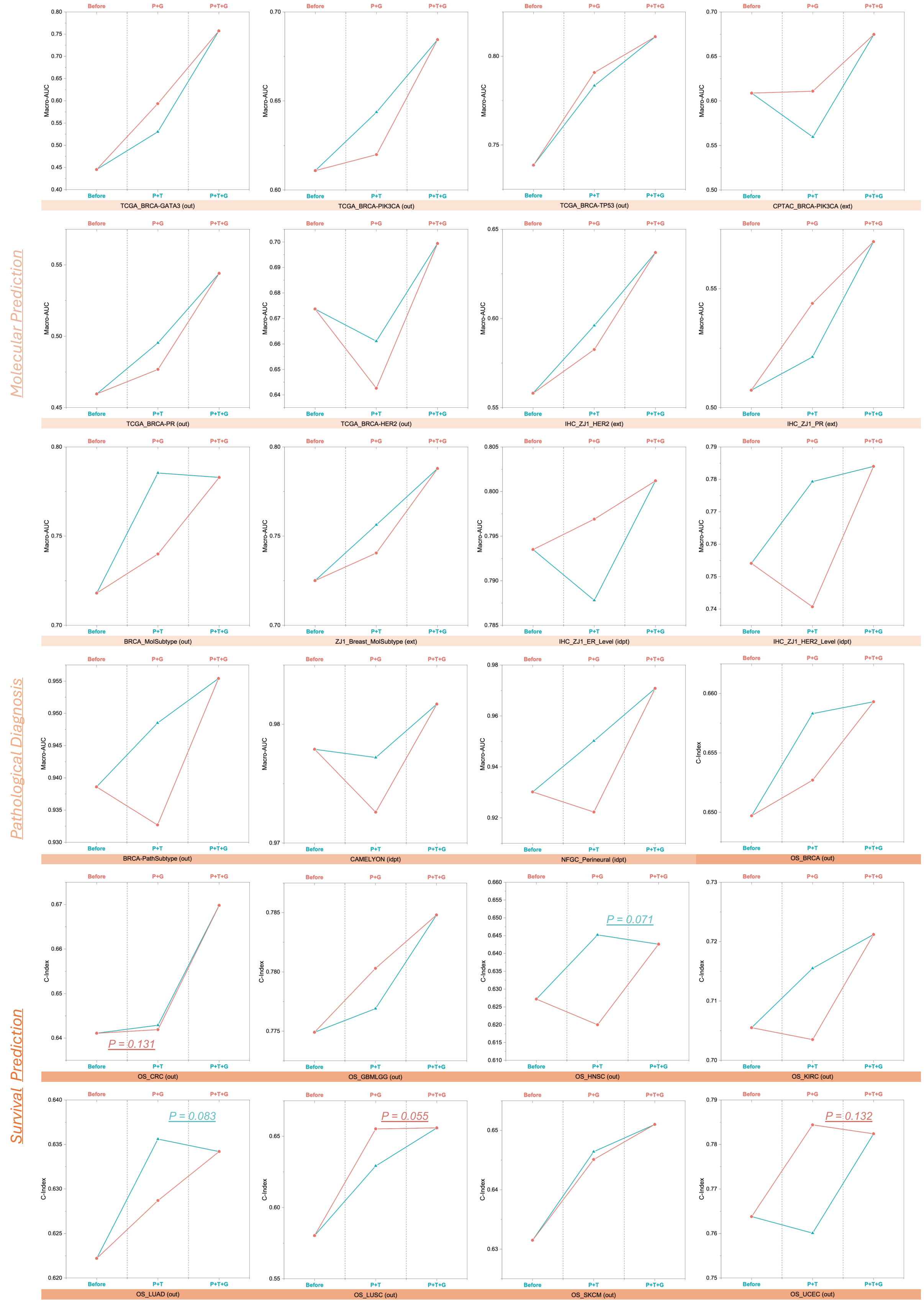


Figure 3. **Performance on each dataset in ablation studies (24 datasets)**, where the red curve indicates the trajectory of ‘before pretraining → + genes → + text’, while the green one represents the trajectory of ‘before pretraining → + text → + genes’. One-sided Wilcoxon

signed-rank tests were conducted for every pairwise comparisons along the trajectory. Unless otherwise specified, reported differences are statistically significant (*P* < 0.05). (Zoom in for details.)

2. A key limitation of this work is the entanglement of two experimental variables: the introduction of a third pre-training modality and a novel pre-training scheme involving fine-tuning of the image encoder (UNI). In contrast, models such as GigaPath, Tangle, and CHIEF rely on frozen feature extractors, allowing clearer attribution of performance gains to specific design choices. Without a controlled disentanglement of these factors, it becomes difficult to isolate the contribution of each component to the final performance.

Response:

We appreciate the reviewer's insightful observation regarding the need to disentangle the effects of our methodological innovations. We would like to clarify that our experimental design explicitly decouples these variables through staged evaluations and controlled comparisons:

1) **Disentangling the Third Modality Contribution**: The introduction of the third modality (gene data or pathology reports) is only involved in the pre-training of aggregator at the Stage 1. Therefore, to demonstrate its contribution, we only need to compare the performance of pretraining on three modalities with 'before aggregator pre-training' and 'after pre-training with two modalities'. As shown in Figures 2 and 3, and discussed in Question 1's answer (points 1 and 2), the three-modality approach demonstrates superior performance both in terms of average results and task-specific dataset.

2) **Disentangling the Novel Pre-training Scheme:** The novel pre-training scheme consists of two stages. Stage 1 focuses on training the aggregator while keeping the extractor frozen. Its effectiveness is evaluated by ablating the contribution of the third pre-training modality (as previously discussed). Stage 2 performs self-taught training of the extractor with the aggregator frozen. Quantitatively, one the one hand, as shown in Fig. 2a, the performance on P+T+G (three modalities) outperforms that of 'before pretraining', which validates the effectiveness of Stage 1. On the other hand, the extractor's improvement is validated by comparing its performance to the UNI, the baseline model (before self-taught training), with results verified across 97 tasks. This underscores the contribution of Stage 2. Qualitatively, from a feature-space perspective, we visualize the evolutionary dynamics of each stage in Fig. 2b. The clusters progressively coalesce, demonstrating clear separation between tumor and non-tumor regions as training advances. Both quantitative performance evaluation and qualitative feature-space visualization consistently verify the effectiveness of each training phase.

3. Furthermore, the study lacks an investigation into scaling laws with respect to model size, which are essential for understanding the cost-benefit trade-offs of integrating additional modalities. Given the minimal performance improvements reported, a detailed analysis of training efficiency and computational overhead is necessary to assess the practical utility of the proposed approach.

Response:

We appreciate the reviewer's suggestion regarding scaling laws and computational trade-offs, and we have discussed this in the Section 3. While we agree these are important considerations, our study specifically investigates modality scalability rather than model size scaling, as the latter has been extensively examined in prior foundation model research [13][15]. Our work builds upon these established scaling principles while focusing on the novel dimension of multimodal integration. A detailed analysis of training efficiency and computational overhead is provided in points 3 and 4 of the Question 1's answer, along with Table 1.

[15] Zimmermann E, Vorontsov E, Viret J, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology[J]. arXiv preprint arXiv:2408.00738, 2024.

5. Finally, if the goal is to evaluate modality scaling, we encourage the authors to consider broader avenues beyond RNA-Seq, such as including IHC or other specialized stainings (e.g., https://arxiv.org/abs/2408.02859), which have shown promising results in conjunction with scaled transformer architectures.

Response:

We sincerely appreciate this insightful suggestion regarding broader modality scaling avenues, including IHC and specialized stained slides. The proposed direction aligns perfectly with our long-term vision of extensible multimodal learning in computational pathology and we are working on that. In the current work, we focused on establishing a robust pretraining framework for integrating three fundamental data types. We thank the reviewer for highlighting this important research trajectory that advances our shared goal of comprehensive multimodal integration for advancing precision oncology. As noted in our Discussion (Section 3), we recognize this as a critical future direction.