nature portfolio

Peer Review File

A generalizable pathology foundation model using a unified knowledge distillation pretraining framework

Corresponding Author: Professor Hao Chen

Version 0:

Decision Letter:

Dear Hao,

Thank you again for submitting to *Nature Biomedical Engineering* your manuscript, "Towards A Generalizable Pathology Foundation Model via Unified Knowledge Distillation". As noted in previous correspondence, the manuscript has been seen by three experts, whose reports you will find at the end of this message (including the two reports I had already forwarded to you).

You will see that the reviewers appreciate the work. However, they express concerns about the degree of support for the claims, and provide useful suggestions for improvement. We hope that with substantial further work you can address the criticisms and convince the reviewers of the merits of the study. In particular, we would expect that a revised version of the manuscript provides systematic evaluations of the pathology foundation models across the full set of tasks as well as substantial additional external validations, as per the criticisms from Reviewers #2 and #3.

When you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.

Please follow the following recommendations:

* Clearly highlight any amendments to the text and figures to help the reviewers and editors find and understand the changes (yet keep in mind that excessive marking can hinder readability).

* If you and your co-authors disagree with a criticism, provide the arguments to the reviewer (optionally, indicate the relevant points in the cover letter).

* If a criticism or suggestion is not addressed, please indicate so in the rebuttal to the reviewer comments and explain the reason(s).

* Consider including responses to any criticisms raised by more than one reviewer at the beginning of the rebuttal, in a section addressed to all reviewers.

* The rebuttal should include the reviewer comments in point-by-point format (please note that we provide all reviewers will the reports as they appear at the end of this message).

* Provide the rebuttal to the reviewer comments and the cover letter as separate files.

We expect that you will be able to resubmit the manuscript within 20 weeks of receiving this message. If this is the case, you will be protected against potential scooping. Otherwise, we will be happy to consider a revised manuscript as long as the significance of the work is not compromised by work published elsewhere or accepted for publication at *Nature Biomedical Engineering*.

We hope that you will find the referee reports helpful when revising the work. Please do not hesitate to contact me should you have any questions.

Best wishes,

Pep

__
Pep Pàmies
Chief Editor, *Nature Biomedical Engineering*

_____
Reviewer #1 (Report for the authors (Required)):

This study establishes a comprehensive benchmark to evaluate the performance of pathology foundation models across 6 clinical types with 39 specific tasks. Then, this study proposes a self-supervised learning approach with a unified knowledge distillation framework consisting of both expert and self-knowledge distillation with pretraining 190 million images from 86K H&E WSIs, and introduced a vision foundation model for pathology “GPFM”.

On the model side, the study evaluated WSI classifier, ROI classifier, survival, retrieval, VQA, and report generation tasks. Then, this study used the expert models for “expert knowledge distillation”, by performing pretraining which includes mask image modeling, self-distillation, and expert knowledge distillation.

The result suggested that GPMF achieved great performance across 29 out of 39 tasks, which is substantially better than the second-best model UNI.

Overall I am impressed by the learning technique that this paper proposed. This paper is novel in terms of its methodology innovation and the comprehensiveness of training data and evaluation tasks. It builds on the similar DINOv2 student-teacher framework but innovatively lets three existing foundation models (UNI, Phikon, CONCH) to perform alignment and further guide the training of GPMF. The GPMF uses 33 public large WSI and patch dataset, trained upon existing off-the-shelf pathology models, and finally achieved decent results across a very comprehensive array of tasks.

Other than that, my main concern is that the figure panels don't follow the logical flow of the text. For example, I was unable to find Figure 6b discussion in main result section. Figure 6d also skipped. The other comment is that I hope the author can expand the technical detail in Methods section 4.1, specifically, further explain this part “To achieve distillation, we use the student network to encode the global views …”, especially the necessity of this alignment process, for both CLS and PATCH tokens.

Minor comments:
Figure 2c & figure 3d, etc.: The best performed model has the shortest bar which looks quite counter-intuitive to me. Suggest using box plot rather than using bar plot with error bar, and inverse the y axis (may be that can be better and more intuitive?)

Figure 2d. Suggest reorder the rows from the low-performance model to best-performance model from top to bottom based on average ranking.

Figure 6b: Suggest adding more detail on caption – what data? Also I believe Figure 6b was not discussed in the main result section.

Figure 6: Instead of using color to represent each model, consider also using unique symbols (star, asterisk, square, etc.). Readers may be color-blinded and may use black/white printer.

Code:

In Github source code training part: https://github.com/birkhoffkiki/GPFM/blob/master/train_scripts/UBC-OCEAN.sh

It seems the main.py training code is missing.

Reviewer #2 (Report for the authors (Required)):

This is timely, interesting and technically new, but has a few issues

major issues:

(1) However, the core claim of the paper regarding generalizability is questionable as out of the 39 tasks, only 7 are external validation, and for 6 of these, the model was still trained on the same cohort, with only a subset held out for testing. The only truly external validation was CPTAC-LUAD, but even here, CPTAC was part of the model's pretraining and I did not find Information on whether they excluded the LUAD slides in pretraining. Additionally, 11 of the 39 tasks are TCGA-based, and the model was heavily trained on TCGA images, and one expert model used (Phikon) was trained exclusively on TCGA data. This represents data leakage and is a major red flag.

(2) also, please make the codes and data accessible now and not just at acceptance

minor issues:

- there are inconsistencies in the formatting and organization of affiliations and other information
- typos such as missing punctuation and inconsistent spacing
- some sections redundantly restate similar content ... this creates unnecessary length
- poor wrtiing with complex and lengthy sentences make the text difficult to follow at times

Reviewer #3 (Report for the authors (Required)):

The authors evaluated the generalizability of foundation models in computational pathology. They found that while existing foundation models excel in specific areas, their performance varies across a broader range of applications. The authors propose a knowledge distillation framework combining expert knowledge distillation, which integrates insights from multiple models, and self-knowledge distillation, which enhances image representation through local-global alignment. Using this framework, they developed the Generalizable Pathology Foundation Model (GPFM) and evaluated its performance on tasks including cancer classification and pathology report generation. Results show that GPFM had an average rank of 1.36 among the models they compared with.

General comments:
1. The criteria for task selection are unclear. For example, The Cancer Genome Atlas (TCGA) datasets they used can support approximately 20 whole-slide image classification tasks and another 20 survival prediction tasks, but only a subset was selected and presented. Different subset selections could alter comparative results.

2. Related to the previous point, the authors did not stratify the pathology visual question answering, report generation, and image retrieval tasks by cancer type in their current analysis. If these results were stratified by cancer type, would the results change?

3. Some of the selected foundation models were not fully evaluated. For example, the PLIP feature can be used for whole-slide pathology image classification and survival analyses. In addition, CONCH, Ctranspath, and UNI features can be used for survival analyses. However, these foundation models were not evaluated for these tasks.

4. Several new foundation models showed better performance in the tasks presented in the manuscript. These models include GigaPath, CHIEF, and Virchow V2. However, these models were not included in the current analyses.

5. The GPFM model for survival prediction showed substantial performance decay when applied to the publicly available Clinical Proteomic Tumor Analysis Consortium (CPTAC) external validation dataset. The authors could discuss the potential implications of this finding.

6. Glioblastoma (GBM) and low-grade glioma (LGG) have distinct pathology imaging profiles and very different prognoses. They should be separated in the survival outcome prediction. The pooled analyses shown in the current manuscript do not have clinical significance.

7. The pathology visual question answering dataset appears to be very noisy, and the meaning of the labels is unclear. For example, what does “polycystic disease infant” mean? A more precise term might be “polycystic kidney disease of the infant.” Similarly, the example of “What is present? Answer: cardiovascular” is also unclear. A better description could be, “What is the tissue type shown in this pathology image? Answer: Blood vessels.”

8. The BLEU scores presented in the pathology report generation task are low (<0.4). In addition, Phikon has performed better than the method proposed by the authors across all evaluation metrics. The authors could further investigate the performance of Phikon in non-TCGA datasets for this task.

9. In addition to the BLEU scores and related metrics, blinded human-based evaluation of the generated pathology reports will provide better insights into the quality of the generated texts.

10. It is interesting to see that DINOv2 without expert knowledge distillation performs much better than the proposed methods in the BreakHis dataset. The authors could discuss the potential reasons behind this.

11. The authors did not compute the p-values for the tasks where GPFM performs worse. Adding these statistical analyses will help readers better understand the differences between GPFM and the better-performing models in these instances.

Additional comment:
1. The figure legend of Figure 6d is incomplete.

Code:

The code provides a README file with sufficient instructions for installing and running the application.

Version 1:

Decision Letter:

Dear Professor Chen,

Thank you for your revised manuscript, "Towards A Generalizable Pathology Foundation Model via Unified Knowledge Distillation". Having consulted with two of the original reviewers (whose comments you will find at the end of this message), I am pleased to write that we shall be happy to publish the manuscript in *Nature Biomedical Engineering*.

We will be performing detailed checks on your manuscript, and in due course will send you a checklist detailing our editorial and formatting requirements. You will need to follow these instructions before you upload the final manuscript files.

Please do not hesitate to contact me if you have any questions.

Best wishes,

Barbara Cheifet
Editor
Nature Biomedical Engineering

_____
Reviewer #1 (Report for the authors (Required)):

The revised manuscript shows substantial improvement over the initial submission. It addresses my previous concerns regarding the figure order and provides greater clarity on the methodological details. The updated figure also enhances readability. Reviewing the responses to other reviewers' comments, I believe the authors have adequately addressed the raised issues. The proposed method is novel, and the authors have conducted various of additional experiments to strengthen their findings.

Major technical criticisms:
None.

Minor technical criticisms or questions:
None.

Missing or unclear details about statistics, protocols or materials:
N/A

Stylistic issues or recommendations:
I recommend that the authors ensure each figure caption is self-contained. For example, in Fig. 1, I recommend use "foundation model" instead of "FM" when first appear.

Reviewer #1 (Remarks on code availability):

The code is well-documented and runs smoothly. However, I have one suggestion: consider including a requirements.txt file for pip installation or a Conda environment YAML file to facilitate reproducibility.

Reviewer #2 (Report for the authors (Required)):

All comments have been addressed. Thank you.

Reviewer #3 (Report for the authors (Required)):

The authors have addressed the comments raised previously. Thank you.

Reviewer #3 (Remarks on code availability):

The code provides a README file with instructions for running the application.

Version 2:

Decision Letter:

Dear Professor Chen,

I am happy to inform you that your manuscript, "A generalizable pathology foundation model using a unified knowledge distillation pretraining framework", has now been accepted for publication in *Nature Biomedical Engineering*.

Over the next few weeks, the figures will be checked for production quality, the text edited to ensure that it conforms to house style, and the manuscript typeset.

Our Articles are published about 40 days after the acceptance date (we recommend that you inform your institutional press office of this timeframe), and you will be notified of the actual publication date a few days in advance. Articles can be published any working day of the week, and are pushed live shortly after 10 am London time.

Publishing agreement. You will be asked to digitally sign a publishing agreement (grant of rights). After the signed publishing agreement has been received, the proofs of the article will be sent to you for review. If you have any queries during the production process, or you cannot meet the requested deadline for returning the proofs, please contact rjsproduction@springernature.com.

Nature Biomedical Engineering is a Transformative Journal. Authors may publish their research with us through the traditional subscription access route, or make their paper immediately open access through payment of an article-processing charge. More information about publication options is available.

You may need to take specific actions to comply with funder and institutional open-access mandates. If the work described in the accepted manuscript is supported by a funder that requires immediate open access (as outlined, for example, by Plan S) and your manuscript was originally submitted on or after January 1st 2021, then you should select the gold OA route. Authors selecting subscription publication will need to accept our standard licensing terms (including our self-archiving policies), and these will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Acceptance of your manuscript is conditional on agreement, by all authors, with both our media embargo and confidentiality and pre-publicity policies. In particular, you may arrange your own publicity of the Article (for instance, through your institutional press office), as long as you ensure that journalists strictly adhere to the media embargo.

To assist you in disseminating the work, as soon as the Article is published you will be able to take advantage of the Springer Nature SharedIt initiative to generate a unique shareable link to the Article that will allow anyone (with or without a subscription) to read it. Recipients of the link who are subscribers will also be able to download and print the PDF.

Thank you for having submitted this work to *Nature Biomedical Engineering*.

Best wishes,

Barbara Cheifet
Editor
Nature Biomedical Engineering

Response Letter

Dear Editor and Reviewers,

We sincerely appreciate your valuable feedback and constructive comments, which have greatly contributed to improving the quality of our work. We have carefully addressed each of the points raised and provided detailed responses to all comments. For your convenience, we have organized our responses by each reviewer:

- **Reviewer #1**
- **Reviewer #2**
- **Reviewer #3**

Please feel free to navigate to the respective sections using the links above. Should you require any additional information or clarification, we would be happy to provide it.

Once again, we thank you for your time and effort in reviewing our manuscript.

Best regards,

Hao Chen

Reviewer #1:

Overall Comments: This study establishes a comprehensive benchmark to evaluate the performance of pathology foundation models across 6 clinical types with 39 specific tasks. Then, this study proposes a self-supervised learning approach with a unified knowledge distillation framework consisting of both expert and self-knowledge distillation with pretraining 190 million images from 86K H&E WSIs, and introduced a vision foundation model for pathology "GPFM".
On the model side, the study evaluated WSI classifier, ROI classifier, survival, retrieval, VQA, and report generation tasks. Then, this study used the expert models for "expert knowledge distillation", by performing pretraining which includes mask image modeling, self-distillation, and expert knowledge distillation.
The result suggested that GPMF achieved great performance across 29 out of 39 tasks, which is substantially better than the second-best model UNI.
Overall I am impressed by the learning technique that this paper proposed. This paper is novel in terms of its methodology innovation and the comprehensiveness of training data and evaluation tasks. It builds on the similar DINOv2 student-teacher framework but innovatively lets three existing foundation models (UNI, Phikon, CONCH) to perform alignment and further guide the training of GPMF. The GPMF uses 33 public large WSI and patch dataset, trained upon existing off-the-shelf pathology models, and finally achieved decent results across a very comprehensive array of tasks.

Response:

We sincerely appreciate your thorough review and positive feedback on our study. We are glad to hear that you found our methodology innovative and the comprehensive evaluation of our model, GPFM. We will address your comments and suggestions one by one.

Comment 1: My main concern is that the figure panels don't follow the logical flow of the text. For example, I was unable to find Figure 6b discussion in main result section. Figure 6d also skipped.

Response:

Thank you for pointing out the potential readability problem. We have carefully revised the manuscript to ensure that all figure panels, including Figure 6b and Figure 6d, are now discussed in the main results section. These changes ensure that the logical flow between the text and figures is now consistent and improves the overall readability of the manuscript. It is worth noting that to improve the readability of the manuscript, we replot Figure 6. The previous Figure 6 and revised Figure 6 are shown below.

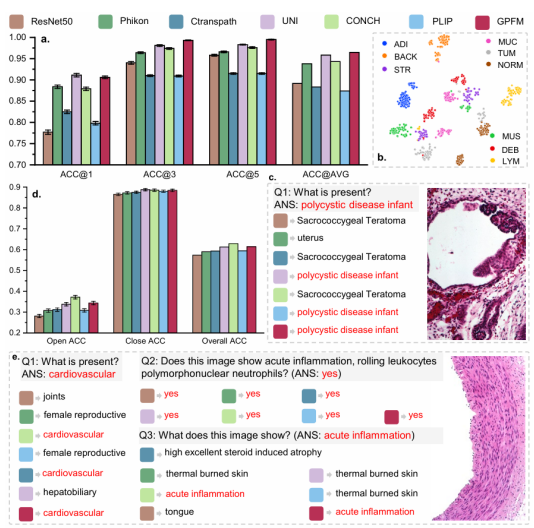


Fig. 6 Overview of Pathology Tissue Retrieval and VQA. a. The top-1, top-3, top-5, and average accuracy of different foundation models on pathology tissue retrieval tasks. **b.** The distribution of features extracted by GPFM. For each class, 100 samples from the test set were used, and a total of 900 samples were subjected to t-SNE dimensionality reduction to 2D. The different classes are distinctly colored in the 2D t-SNE plot. **c.** An open-ended question along with the answers generated by various foundation models. **d.** The performance of VQA, measured by open-ended accuracy, closed-ended accuracy, and overall accuracy, for different foundation models. The error bars **e.** Three questions and the answers generated by foundation models related to the query image. The error bars indicate standard deviation.

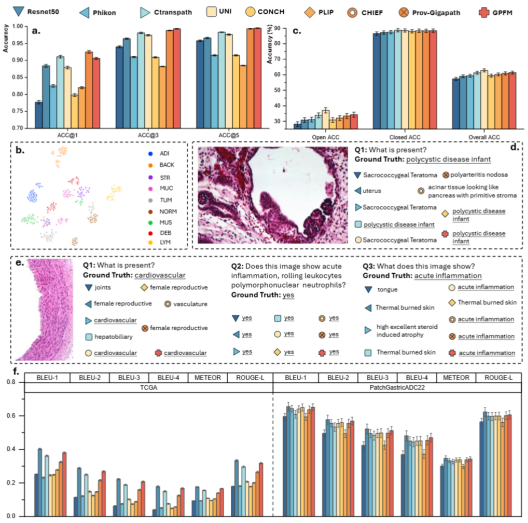


Fig. 6 Overview of Pathology Tissue Retrieval, VQA, and Report Generation. a. The top-1, top-3, top-5, and average accuracy of different FMs on pathology tissue retrieval tasks. **b.** The distribution of features extracted by GPFM. For each class, 100 samples from the test set were used, and a total of 900 samples were subjected to t-SNE dimensionality reduction to 2D. **c.** The performance of VQA on PathVQA dataset, measured by open-ended accuracy, closed-ended accuracy, and overall accuracy, for different FMs. **d.** An open-ended question along with the answers generated by various FMs. **e.** Three questions and the answers generated by FMs related to the query image. **f.**The performance of report generation on TCGA and PatchGastricADC22 data. The models are measured by six different metrics. In all subfigures, the error bars indicate standard deviation.

Previous Figure 6 (left) and reivesed Figure 6 (right).

The revised figures can be found in **Section 2.6 on page 12** in the manuscript.

For the figure 6, the discussion in the main section is listed below:

Fig 6a:

The experimental results (Fig. 6a, Extended Data Table A37) show that the GPFM model achieved the second-best Top-1 accuracy with a value of 0.906 (-1.9%, Prov-Gigapath). However, GPFM outperforms other models in terms of Top-3 and Top-5 accuracy, achieving values of 0.993 (+0.5%, Prov-Gigapath) and 0.995 (+0.2%, Prov-Gigapath), respectively.

Fig 6b:

To further explore the clustering effect and feature representation ability, we utilized t-Distributed Stochastic Neighbor Embedding (t-SNE) [50] to project the features extracted by GPFM into a 2D embedding space. The categories are well clustered, further illustrating that the features are highly discriminative (Fig. 6b).

Fig 6c:

For the patch-level VQA task, our model achieved the second-best performance, with results only slightly lower than those of CONCH (Fig. 6c, Extended Data Table A38). It is important to note that CONCH is a vision-language FM trained on millions of image-text pairs, which inherently provides it with an advantage in VQA tasks.

Fig 6d and Fig 6e:

Despite this, our results highlight the substantial potential of our approach compared to other pure vision FMs. To further illustrate the capabilities of our model, we visualized the query images, questions, and answers generated by different FMs (Fig. 6d and 6e).

Fig 6f:

The experimental results demonstrate that Phikon achieved the best performance across all six metrics, while GPFM achieved comparable performance and ranked as the second-best model on both tasks (Fig. 6f, Extended Data Table A40 and A42).

Comment 2: The other comment is that I hope the author can expand the technical detail in Methods section 4.1, specifically, further explain this part "To achieve distillation, we use the student network to encode the global views …", especially the necessity of this alignment process, for both CLS and PATCH tokens.

Response:

Thank you for the suggestions. To avoid potential confusion of our distillation method, we add more details in the revised manuscript in **section 4.1.** The previous version and the revised text are listed below.

Previous version: "To achieve distillation, we use the student network to encode the global views *u* and *v* and extract the [CLS] and [PATCH] tokens. Additionally, we employ the off-the-shelf foundation models (UNI, Phikon, CONCH) to obtain their respective [CLS] and [PATCH] tokens. For aligning the class tokens, we utilize cosine similarity. As for the patch token alignment, we employ both cosine similarity and smooth L1 distance.

Revised version:

"To maximize the generalizability of the pretrained model, it is crucial to balance the performance and diversity of expert models. We evaluated several existing models across six different tasks, selecting those that excelled in classification (UNI), survival analysis (Phikon), and visual question answering (CONCH) as expert models (see Fig. 1c). The [CLS] token, which represents the overall information of a patch for downstream tasks, serves as a critical component in our approach. If the [CLS] token of our model aligns well with those of the expert models, it indicates that our model can effectively assimilate the knowledge from selected experts. Similarly, the [PATCH] token also contains rich information. For example, some methods use mean pooling to perform downstream tasks [76]. Therefore, aligning the [PATCH] token can further improve the effect of knowledge transfer. To achieve above alignments, we use the student network to encode the global views u and v and extract the [CLS] and [PATCH] tokens. Additionally, we employ the adopted experts to obtain their [CLS] and [PATCH] tokens, respectively."

Comment 3: Figure 2c & figure 3d, etc.: The best performed model has the shortest bar which looks quite counter-intuitive to me. Suggest using box plot rather than using bar plot with error bar, and inverse the y axis (may be that can be better and more intuitive?)

Response:

Thank you for your suggestions. To make it clearer, we used box plot and inverse the y axis. The previous figures and revised figures are shown in the following lines.

Fig 2:

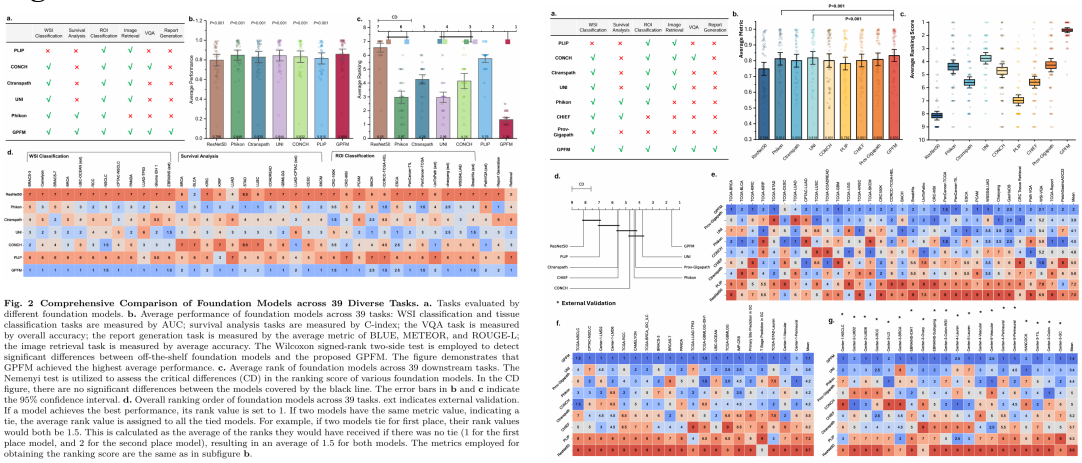
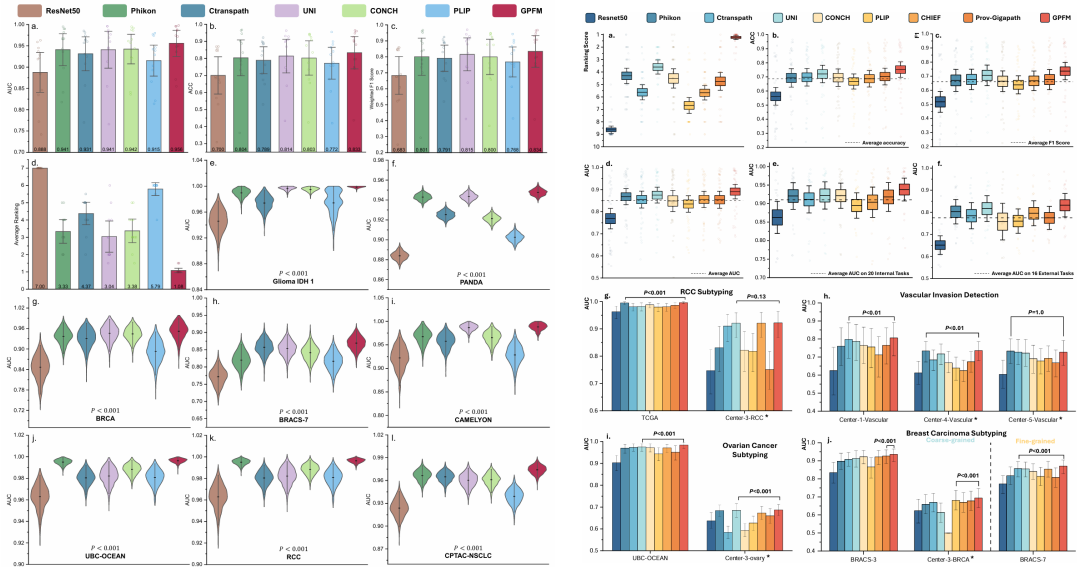


Fig. 2 Comprehensive Comparison of Foundation Models across 39 Diverse Tasks. a. Tasks evaluated by different foundation models. **b.** Average performance of foundation models across 39 tasks: WSI classification and tissue classification tasks are measured by AUC; survival analysis tasks are measured by C-index; the VQA task is measured by overall accuracy; the report generation task is measured by the average metric of BLUE, METEOR, and ROUGE-L; the image retrieval task is measured by average accuracy. The Wilcoxon signed-rank two-side test is employed to detect significant differences between off-the-shelf foundation models and the proposed GPFM. The figure demonstrates that GPFM achieved the highest average performance. **c.** Average rank of foundation models across 39 downstream tasks. The Nemenyi test is utilized to assess the critical differences (CD) in the ranking score of various foundation models. In the CD figure, there are no significant differences between the models covered by the black line. The error bars in **b** and **c** indicate the 95% confidence interval. **d.** Overall ranking order of foundation models across 39 tasks. ext indicates external validation. If a model achieves the best performance, its rank value is set to 1. If two models have the same metric value, indicating a tie, the average rank value is assigned to all the tied models. For example, if two models tie for first place, their rank values would both be 1.5. This is calculated as the average of the ranks they would have received if there was no tie (1 for the first place model, and 2 for the second place model), resulting in an average of 1.5 for both models. The metrics employed for obtaining the ranking score are the same as in subfigure **b**.

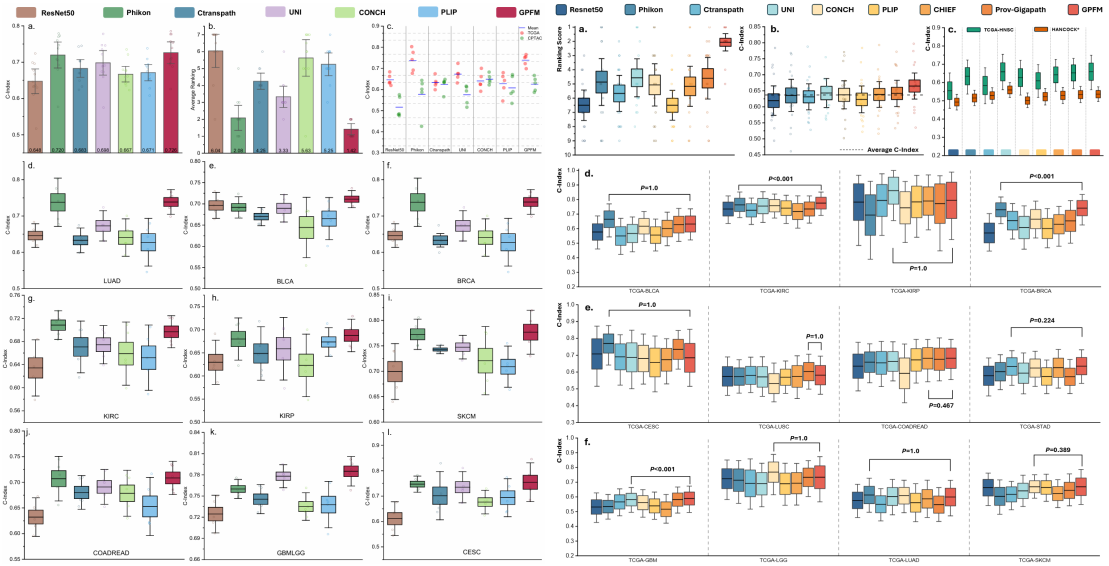
The previous Fig. 2 (left) and the revised Fig. 2 (right) in the manuscript. The bar plot Fig. 2c is replaced with a box plot. The revised figure can be found at **Section 2, Page 5**.

Fig 3:



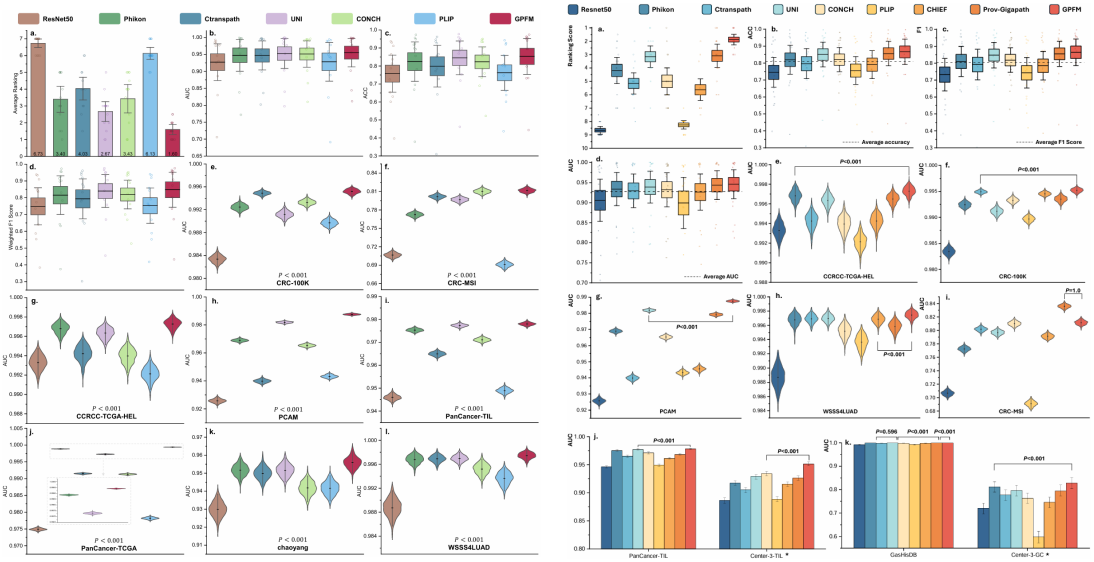
The previous Fig. 3 (left) and the revised Fig. 3 (right) in the manuscript. The bar plot **Fig. 3d** is replaced with a box plot **Fig. 3a**. For the revised figure, please refer to **Section 2.1, Page 6**.

Fig 4:



The previous Fig. 4 (left) and the revised Fig. 4 (right) in the manuscript. The bar plot **Fig. 4a** is replaced with a box plot **Fig. 4a**. The revised figure 4 can be found at **Section 2.2, Page 7**.

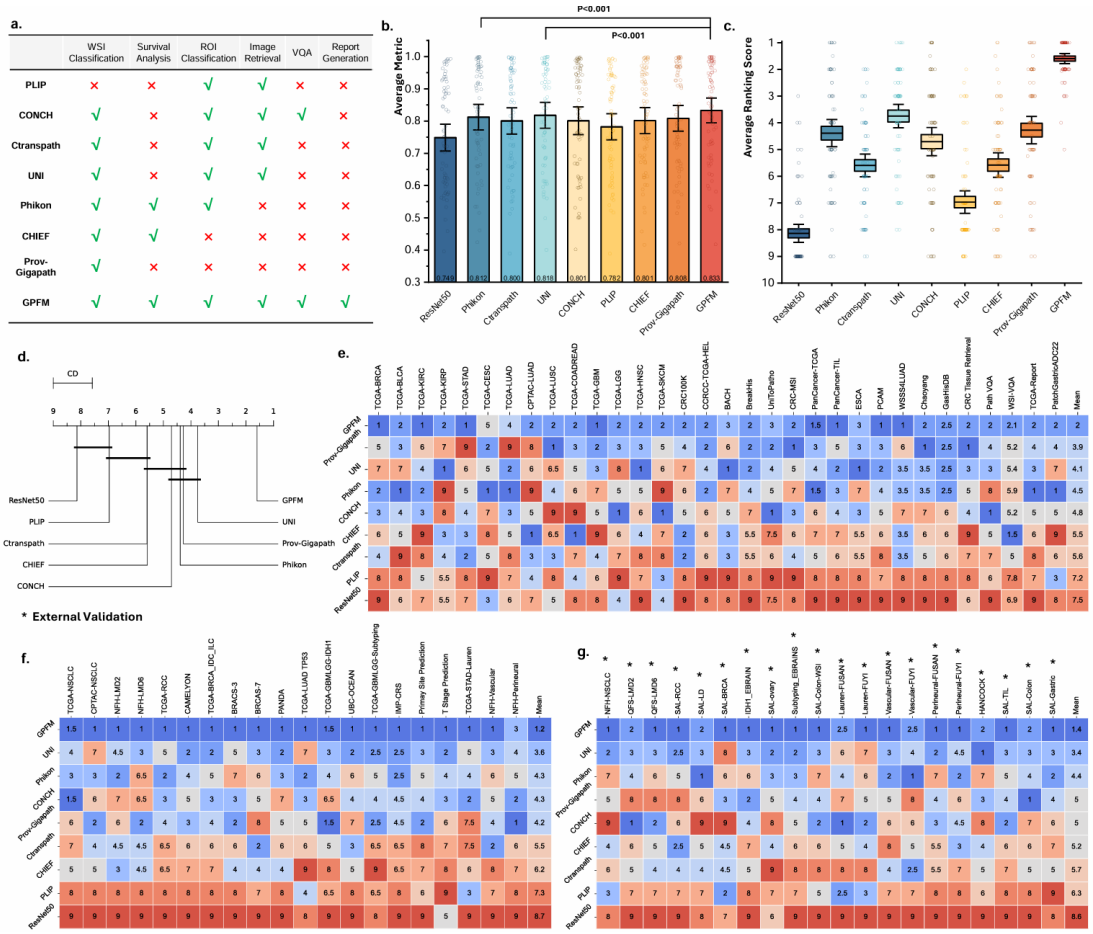
Fig 5:



The previous Fig. 5 (left) and the revised Fig. 5 (right) in the manuscript. The bar plot **Fig. 5a** is replaced with a box plot **Fig. 5a**. The revised figure 5 can be found in **Section 2.3, Page 9**.

Comment 4: Figure 2d. Suggest reorder the rows from the low-performance model to best-performance model from top to bottom based on average ranking.

Response: Thank you for your suggestions. That's a good suggestion which is much clearer. The revised Figure 2 is shown below. It is worth noting that we add another 33 tasks, to make it clearer. We split the table into 3 parts, internal WSI classification tasks (Fig. 2f), external tasks (Fig. 2g), and other tasks (Fig. 2e). The order is based on their average rank. The revised figure can be found at **Section 2, Page 5**.



The Revised Fig. 2. We reordered the models based on their average ranking score.

Comment 5: Figure 6b: Suggest adding more detail on caption – what data? Also I believe Figure 6b was not discussed in the main result section.

Response:

Thank you for your suggestions. The data used for visualization is CRC-100K, which is also used for the tissue retrieval task (Figure 6a). To avoid any potential confusion, we add more details in the caption. The revised caption is shown in below lines:

Previous version:

"b. The distribution of features extracted by GPFM. For each class, 100 samples from the test set were used, and a total of 900 samples were subjected to t-SNE dimensionality reduction to 2D."

Revised version:

"b. Distribution of features extracted by GPFM on the CRC-100K dataset. For each

class, 100 samples from the test set are used, and a total of 900 samples are reduced to 2D embeddings using t-SNE. Different classes are represented by different colors in the 2D t-SNE plot."

Regarding your second question, we appreciate you pointing this out and helping us improve the quality of this paper. To make it more complete, we discussed Figure 6b in the main result section as shown in the following lines:
"To further explore the clustering effect and feature representation ability, we utilized t-Distributed Stochastic Neighbor Embedding (t-SNE) [50] to project the features extracted by GPFM into a 2D embedding space. The categories are well clustered, further illustrating that the features are highly discriminative (Fig. 6b)."

For the above modifications, we have updated them in the manuscript in **Section 2.4, Page 10**.

Comment 6: Figure 6: Instead of using color to represent each model, consider also using unique symbols (star, asterisk, square, etc.). Readers may be color-blinded and may use black/white printer.

Response: We are grateful for the suggestions. To make the manuscript easier to read for all people, we adopted unique symbols to represent different models. The modified figure is as below. The revised figures can be found in **Section 2.6 on page 12** in the manuscript.

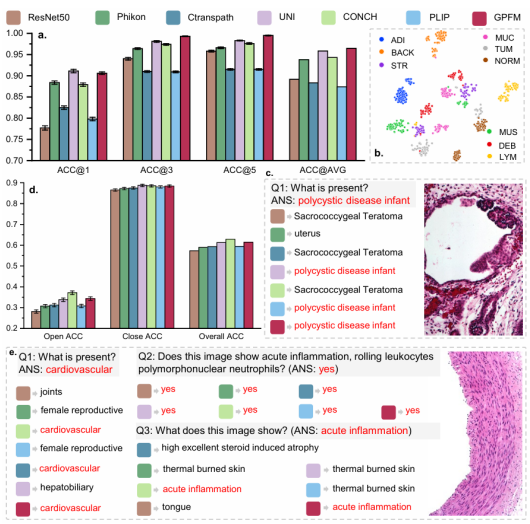


Fig. 6 Overview of Pathology Tissue Retrieval and VQA. a. The top-1, top-3, top-5, and average accuracy of different foundation models on pathology tissue retrieval tasks. **b.** The distribution of features extracted by GPFM. For each class, 100 samples from the test set were used, and a total of 900 samples were subjected to t-SNE dimensionality reduction to 2D. The different classes are distinctly colored in the 2D t-SNE plot. **c.** An open-ended question along with the answers generated by various foundation models. **d.** The performance of VQA, measured by open-ended accuracy, closed-ended accuracy, and overall accuracy, for different foundation models. The error bars **e.** Three questions and the answers generated by foundation models related to the query image. The error bars indicate standard deviation.

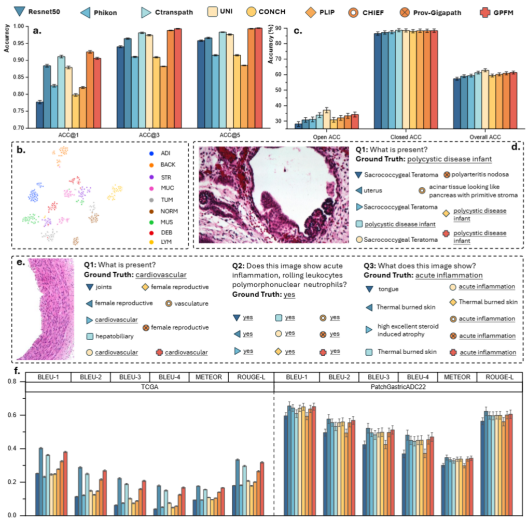


Fig. 6 Overview of Pathology Tissue Retrieval, VQA, and Report Generation. a. The top-1, top-3, top-5, and average accuracy of different FMs on pathology tissue retrieval tasks. **b.** The distribution of features extracted by GPFM. For each class, 100 samples from the test set were used, and a total of 900 samples were subjected to t-SNE dimensionality reduction to 2D. **c.** The performance of VQA on PathVQA dataset, measured by open-ended accuracy, closed-ended accuracy, and overall accuracy, for different FMs. **d.** An open-ended question along with the answers generated by various FMs. **e.** Three questions and the answers generated by FMs related to the query image. **f.**The performance of report generation on TCGA and PatchGastricAD22 data. The models are measured by six different metrics. In all subfigures, the error bars indicate standard deviation.

Previous Figure 6 (left) and reivesed Figure 6 (right).

Comment 7: In Github source code training part:
https://github.com/birkhoffkiki/GPFM/blob/master/train_scripts/UBC-OCEAN.sh
It seems the main.py training code is missing.

Response:

Thanks for your reminder. We uploaded main.py. We are actively maintaining this

project and continuously updating it to address feedback from the research community. If you encounter any further issues or have additional suggestions, please feel free to reach out. We appreciate your valuable feedback!

Reviewer #2:

Overall comment: This is timely, interesting and technically new, but has a few issues.
Response:
Thank you for your valuable feedback. We appreciate your recognition of the timeliness and technical novelty of our work. We are committed to addressing the issues you've identified to improve the manuscript. The response to your comments is listed below one by one. Thank you again for your insights!

Comment 1: However, the core claim of the paper regarding generalizability is questionable as out of the 39 tasks, only 7 are external validation, and for 6 of these, the model was still trained on the same cohort, with only a subset held out for testing. The only truly external validation was CPTAC-LUAD, but even here, CPTAC was part of the model's pretraining and I did not find Information on whether they excluded the LUAD slides in pretraining. Additionally, 11 of the 39 tasks are TCGA-based, and the model was heavily trained on TCGA images, and one expert model used (Phikon) was trained exclusively on TCGA data. This represents data leakage and is a major red flag.
Response:
Thank you for your insightful comment. For your first concern, we acknowledge that the initial number of external validation tasks was limited. To address this, we have expanded our evaluation to include an additional 33 tasks (28 of which are non-TCGA), bringing the total number of tasks to 72. Among these, 20 are external validation tasks, and 13 are internal tasks. These additions significantly strengthen the generalizability assessment of our model. Regarding the concern about data leakage, particularly with the use of the Phikon model, we have taken care to ensure that the newly added tasks do not overlap with the pretraining data of GPFM or the expert models (CONCH, UNI, Phikon). In Table 1 (below), the newly added tasks are highlighted in red, and underlined entries indicate datasets that were not used in the pretraining of GPFM or any of the expert models. This ensures a rigorous evaluation without data leakage.

Table 1. Evaluated extra downstream tasks.

Task Name	**Internal data (Data Number)**	**External Data (Data Number)**
NSCLC Subtyping	TCGA-NSCLC (1053 WSIs)	CENTER-1-NSCLC (210 WSIs)
Metastatic Detection (Lung Cancer)	CENTER-1-LMD2 (1198 WSIs)	CENTER-2-LMD2 (530 WSIs)
Primary Site Prediction (Lung Cancer)	CENTER-1-LMD6 (1198 WSIs)	CENTER-2-LMD6 (530 WSIs)

RCC Subtyping	TCGA-RCC (937 WSIs)	CENTER-3-RCC (88 slides)
ILC & IDC classification	TCGA-BRCA (985 WSIs)	CENTER-3-LD (383 WSIs)
Breast Carcinoma Subtyping	BRACS (545 WSIs)	CENTER-3-BRCA (467 WSIs)
IDH Mutation Prediction in Glioma[1]	TCGA-GBMLGG-IDH1 (979 WSIs)	EBRAIN-IDH1 (852)
Ovarian Cancer Subtyping	UBC-OCEAN (527 WSIs)	CENTER-3-Ovary (370 WSIs)
Brain Tumor Subtyping[1]	TCGA-GBMLGG-Subtyping (1276 WSIs)	EBRAIN-Subtyping (732)
Lesion Grade Classification (Colon Cancer)	IMP-CRS (5332 WSIs)	CENTER-3-Colon-WSI (297 WSIs)
Primary Site Prediction (Head & Neck Cancer)	HANCOCK (708 WSIs)	-
T stage Prediction (Head & Neck Cancer)	HANCOCK (705 WSIs)	-
Lauren Subtyping (Gastric Cancer)	TCGA-STAD (390 WSIs)	☐ CENTER-5-Lauren (141 WSIs) ☐ CENTER-4-Lauren (319 WSIs)
Vascular Invasion Detection (Gastric Cancer)	CENTER-1-Vascular (396 WSIs)	☐ CENTER-5-Vascular (230 WSIs) ☐ CENTER-4-Vascular (319 WSIs)
Perineural Invasion Detection (Gastric Cancer)	CENTER-1-Perineural (397 WSIs)	☐ CENTER-5-Perineural (232 WSIs) ☐ CENTER-4- Perineural (319 WSIs)

Survival Analysis (Head & Neck Cancer)	TCGA-HNSC (443 cases)	HANCOCK (749 cases)
Survival Analysis (Brain Cancer)(2)	☐ TCGA-GBM (372 cases) ☐ TCGA-LGG (462 cases)	-
Tumor infiltrating lymphocyte (TIL) classification	Pancancer-TIL (304,097 patches)	CENTER-3-TIL (18,492 patches)
Colon Tissue Classification	Chaoyang (6,160 patches)	CENTER-3-Colon (21,068 patches)
Gastic Tissue Classification	GasHisDB (13,124 patches)	CENTER-3-GC (2,537 patches)
VQA for WSI	WSI-VQA (977 WSIs)	-
Report Generation (Gastric Cancer)	PatchGastricADC22 (991 WSIs)	-

(1) In the previous manuscript, this task only performed internal training and testing on EBRAIN dataset. In the revised manuscript, we use the TCGA-GBMLGG as internal data and use the EBRAIN as the external data.

(2) In the previous manuscript, this task is performed on TCGA-GBMLGG data. However, the comment 6 raised by reviewer #3 suggests that GBM and LGG should predict separately due to distinct pathology imaging profiles and very different prognoses. Therefore, we reformulate this task.

The ranking scores of various foundation models across the expanded set of tasks are presented in the accompanying Figure 1. As shown, GPFM achieves the best average performance on both the 52 internal tasks and the 20 external validation tasks (see more results at the end of this comment). This robust performance across a diverse set of tasks, including 28 non-TCGA and 20 external validation tasks, underscores the generalizability of GPFM and the effectiveness of our proposed expert knowledge distillation pretraining strategy. We believe these additions and clarifications address the reviewer's concerns and further validate the generalizability and robustness of our model.

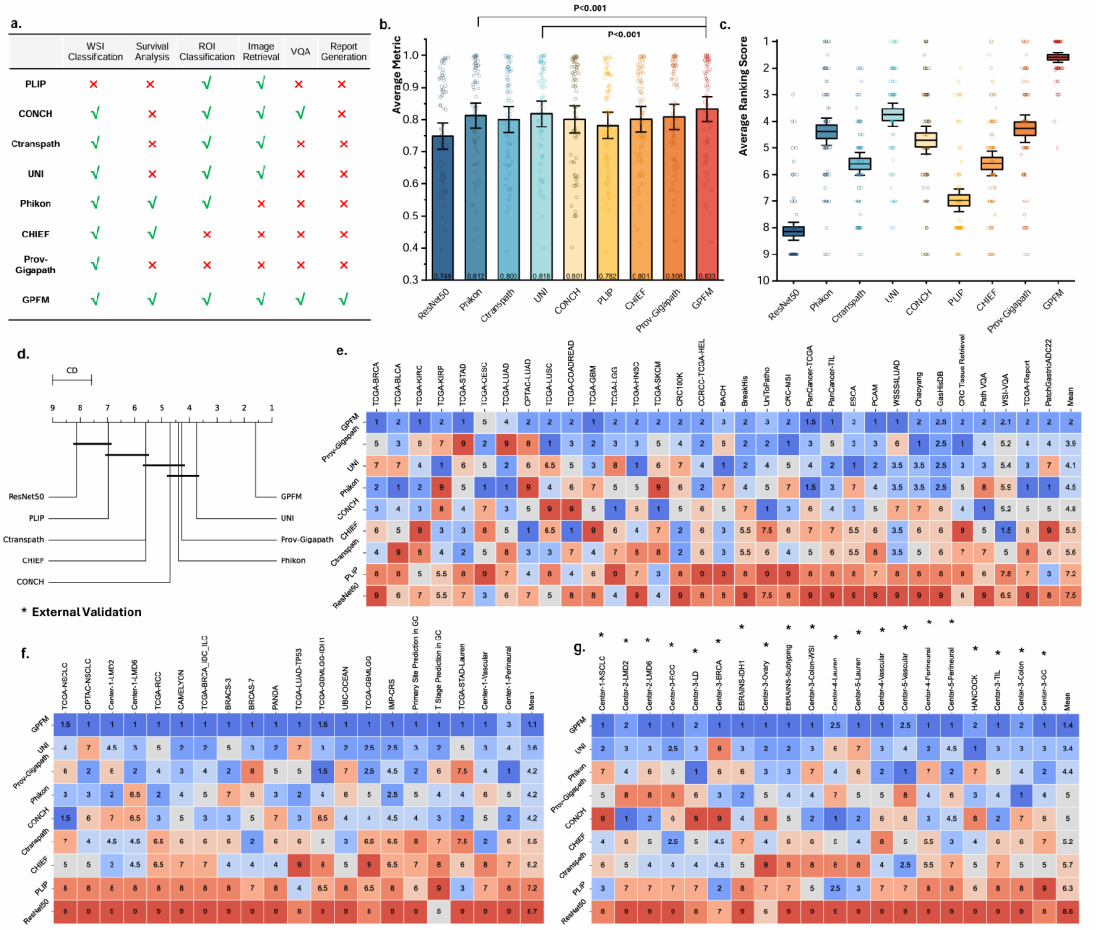


Figure 1. The ranking score of foundation models across 72 tasks. The revised figure is presented in **Section 2, Page 5** in the manuscript.

Additional Experimental Results:

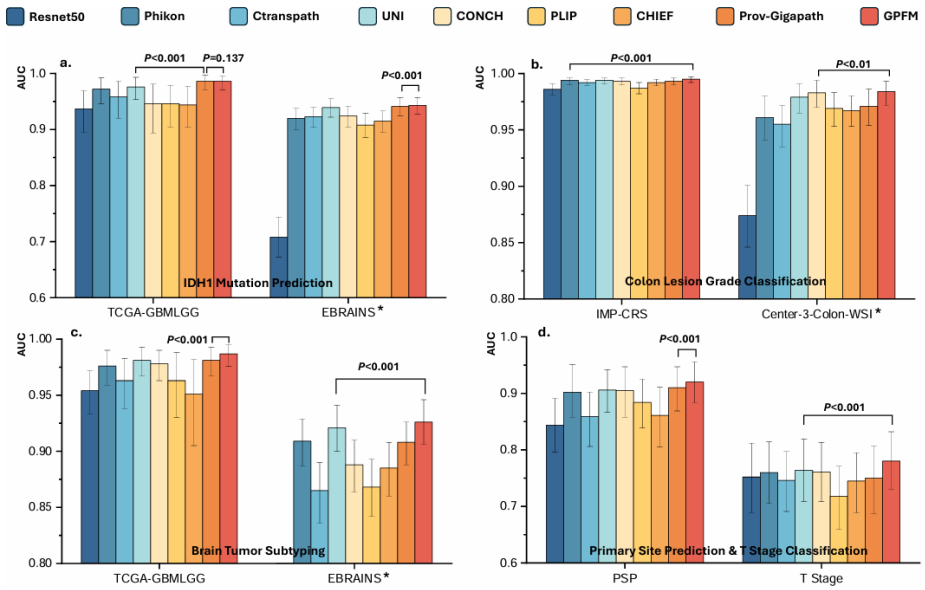


Fig. A2 Extended Results of WSI Classification. a. IDH-1 mutation prediction in brain tumors. **b.** Lesion grading in colon cancer. **c.** Brain tumor subtyping performance. **d.** Dual-task evaluation: primary site prediction and T-stage classification in head & neck cancer. Error bars represent 95% CI. External validation cohorts are marked with *.

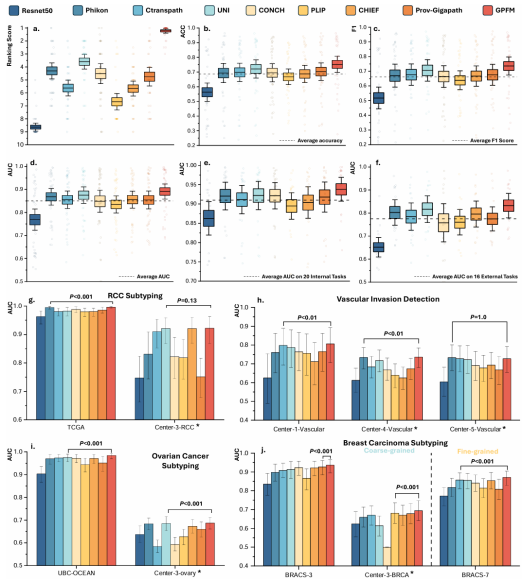


Fig. 3 Performance of FMs on WSI Classification Tasks. a. Average ranking of FMs based on AUC across 36 WSI classification tasks. **b-d.** Average balanced accuracy (ACC), and weighted F1 score (F1), and AUC of FMs across 36 WSI classification tasks. **e.** Average AUC of FMs on 20 internal WSI classification tasks. **f.** Average AUC of FMs on 16 external validation cohorts. **g-h.** Model performance on specific tasks: RCC subtyping, vascular invasion detection, ovarian cancer subtyping, and breast carcinoma subtyping. * represents external validation cohorts. Error bars represent 95% CI. Additional results are shown in Extended Data **Fig.** A1 and **Fig.** A2.

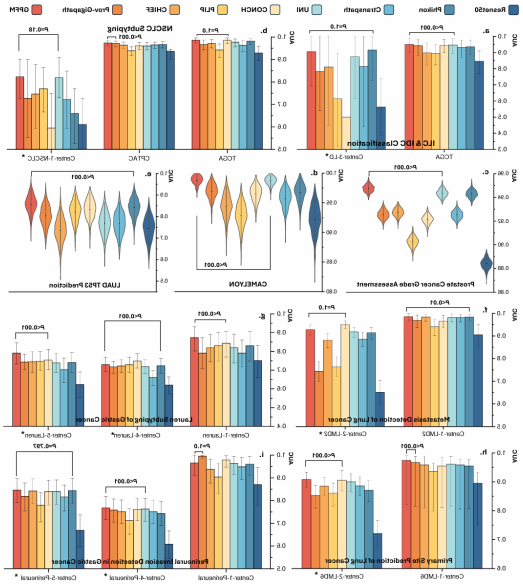


Fig. A1 Extended Results of WSI Classification. a. Performance comparison of foundation models in ILC and IDC classification. **b.** NSCLC subtyping performance across models. **c-e.** Model performance in prostate cancer grading, breast cancer metastasis detection, and LUAD TP53 mutation prediction, respectively. **f-i.** Extended evaluation including lung cancer metastasis detection, gastric cancer Lauren subtyping, lung cancer primary site prediction, and gastric cancer peritoneal invasion detection. Violin plots show the distribution of 1,000 bootstrap replicates. Error bars represent 95% CI. External validation cohorts are marked with *.

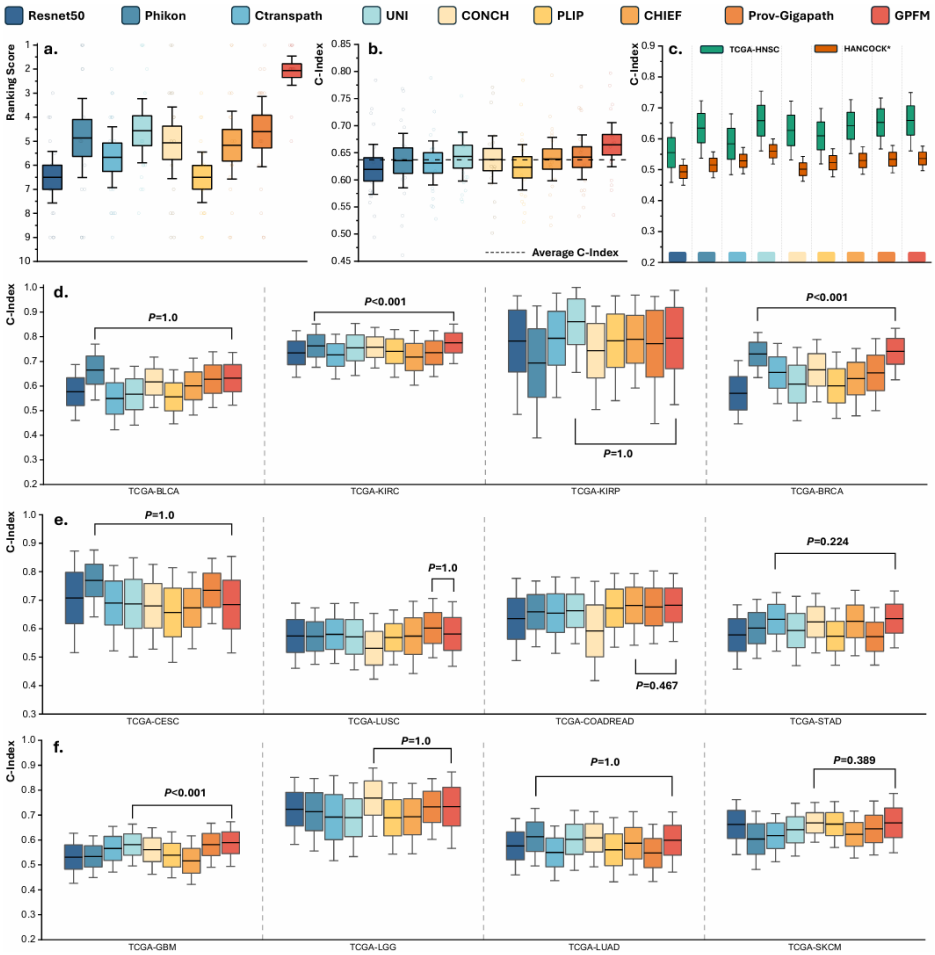


Fig. 4 Performance of FMs across 15 Survival Analysis Tasks. a. Average ranking of FMs in 15 survival analysis tasks. **b.** Average C-Index of various FMs across 15 tasks. **c.** Results on TCGA-HNSC data and the HANCOCK cohort. The survival prediction model was trained on the TCGA-HNSC cohort and subsequently tested on the HANCOCK cohort. **d-f.** C-Index of FMs across 12 survival analysis tasks. In all subfigures, error bars indicate 95% CI. For box plots, the center line represents the mean, and the box limits represent the standard error.

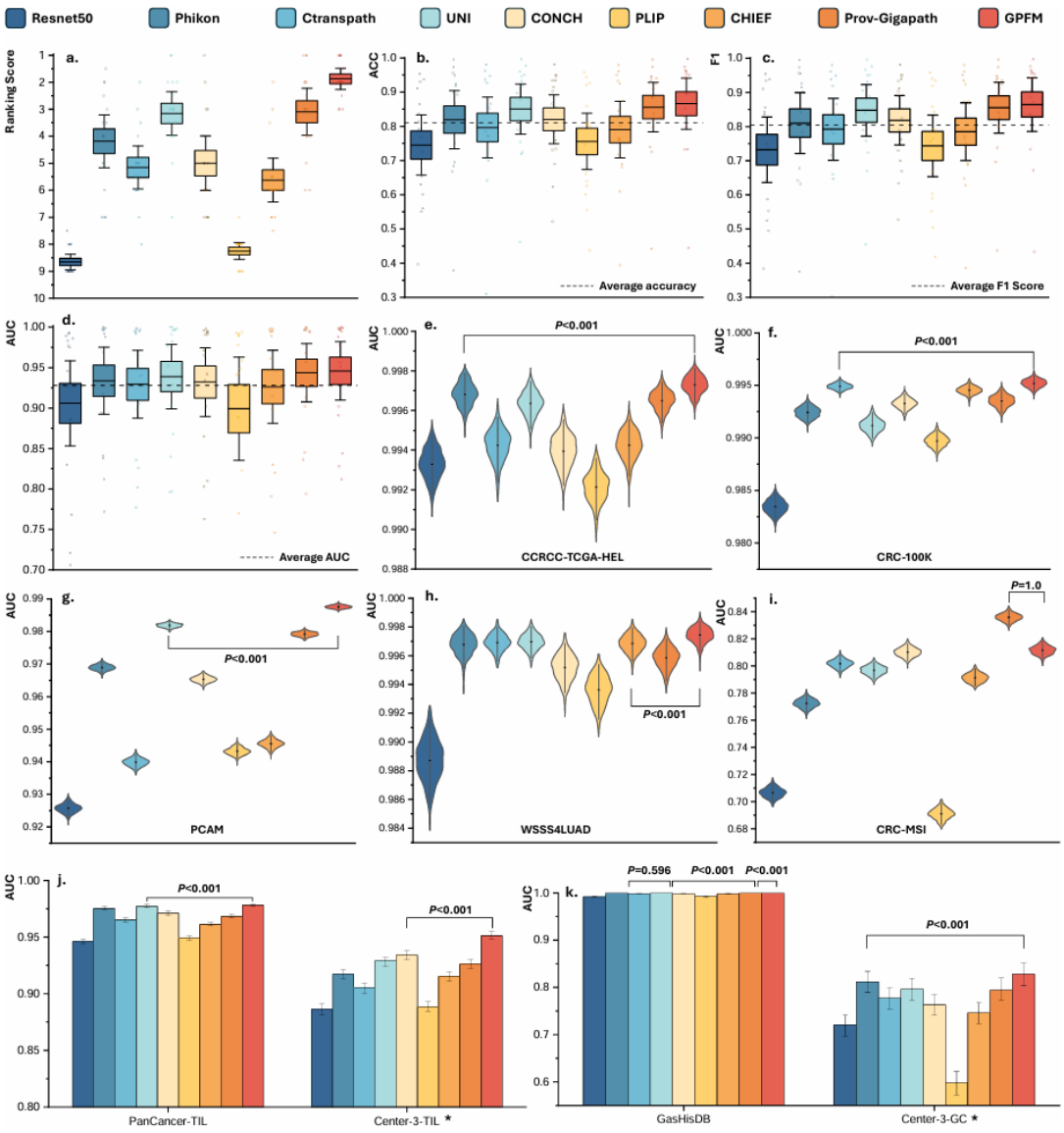


Fig. 5 Performance of FMs on Tissue Classification Tasks. a. Average ranking order of FMs based on AUC across 16 tasks. **b-d.** Average balanced accuracy (ACC), and weighted F1 score (F1), and AUC of FMs across 16 tasks. The center line represents mean and the box limits represents the standard error. **e-i.** AUC of FMs across 5 tissue classification tasks. The Wilcoxon signed-rank one-side test is adopted to detect significant difference. Then center black line in the violin plot represents the mean AUC. **j.** Tumor infiltrating lymphocytes classification based on the PanCancer-TIL (internal) and Center-3-TIL data (external). **k.** Gastric cancer tissue classificaiton with GasHisDB (internal) and Center-3-GC data (external). In all subfigures, the error bars indicate 95% CI. More results are presented in Extended Data **Fig.** A3.

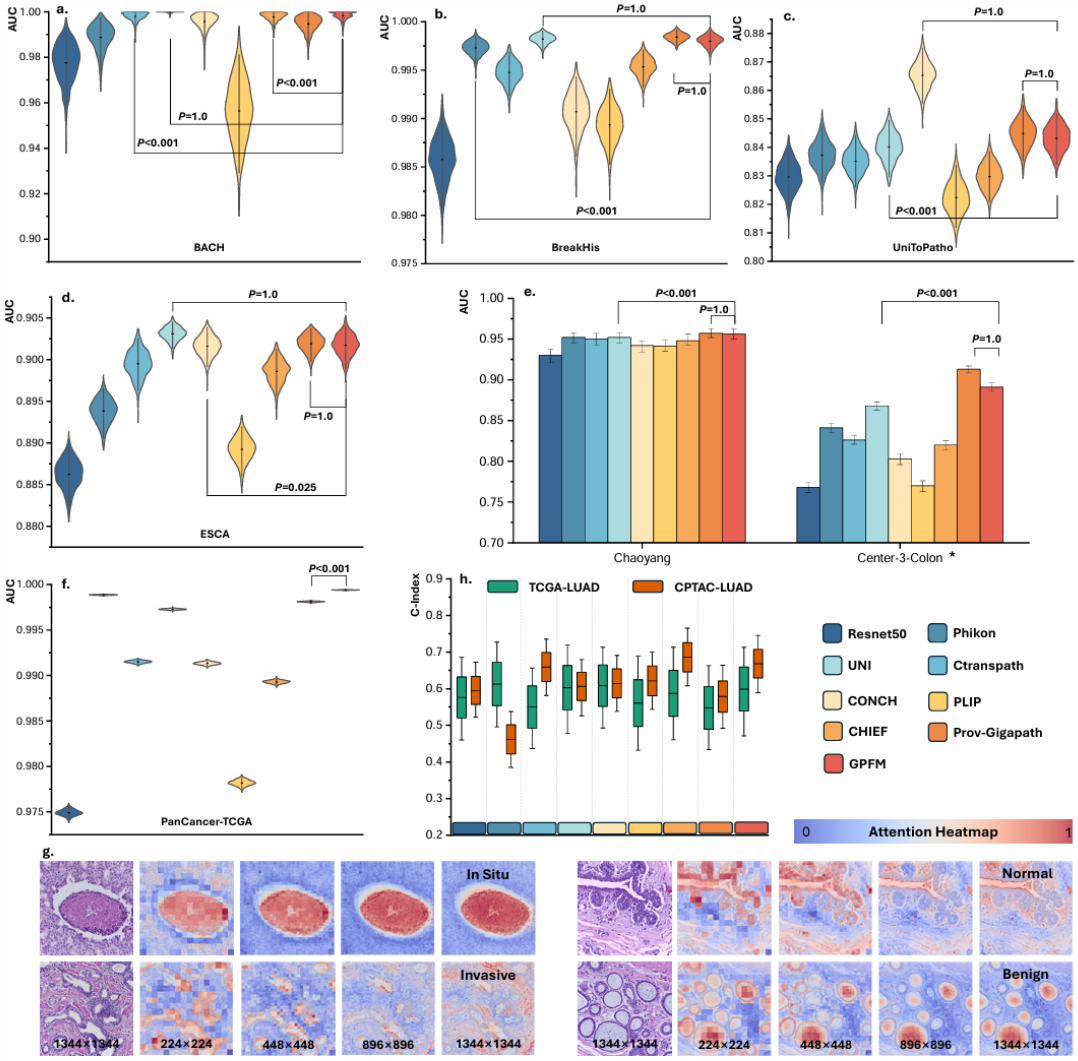


Fig. A3 Extended Result of ROI Classification Tasks. a-d. The AUC of foundation models on BACH, BreakHis, UniToPatho, and ESCA, respectively. **e.** The colon tissue classification performance. The Chaoyang and Center-3-Colon serve as internal and external, respectively. **f.** The performance of pancancer classificaiton of different foundation models. **g.**Attention heatmap of GPFM across various image resolutions for BRCA subtyping in BACH dataset. The colored squares represent the 14×14 [PATCH] tokens encoded by the GPFM model. The heatmap values indicate the similarity between each [PATCH] token and the [CLS] token generated by the last layer of GPFM, measured using Euclidean distance. The consistent attention patterns observed across varying image resolutions and tissue types underscore the robust capabilities of the GPFM model. **h.** Results on TCGA-LUAD data and the CPTAC-LUAD cohort. The survival prediction model was trained on the TCGA-LUAD cohort and subsequently tested on the CPTAC-LUAD cohort.

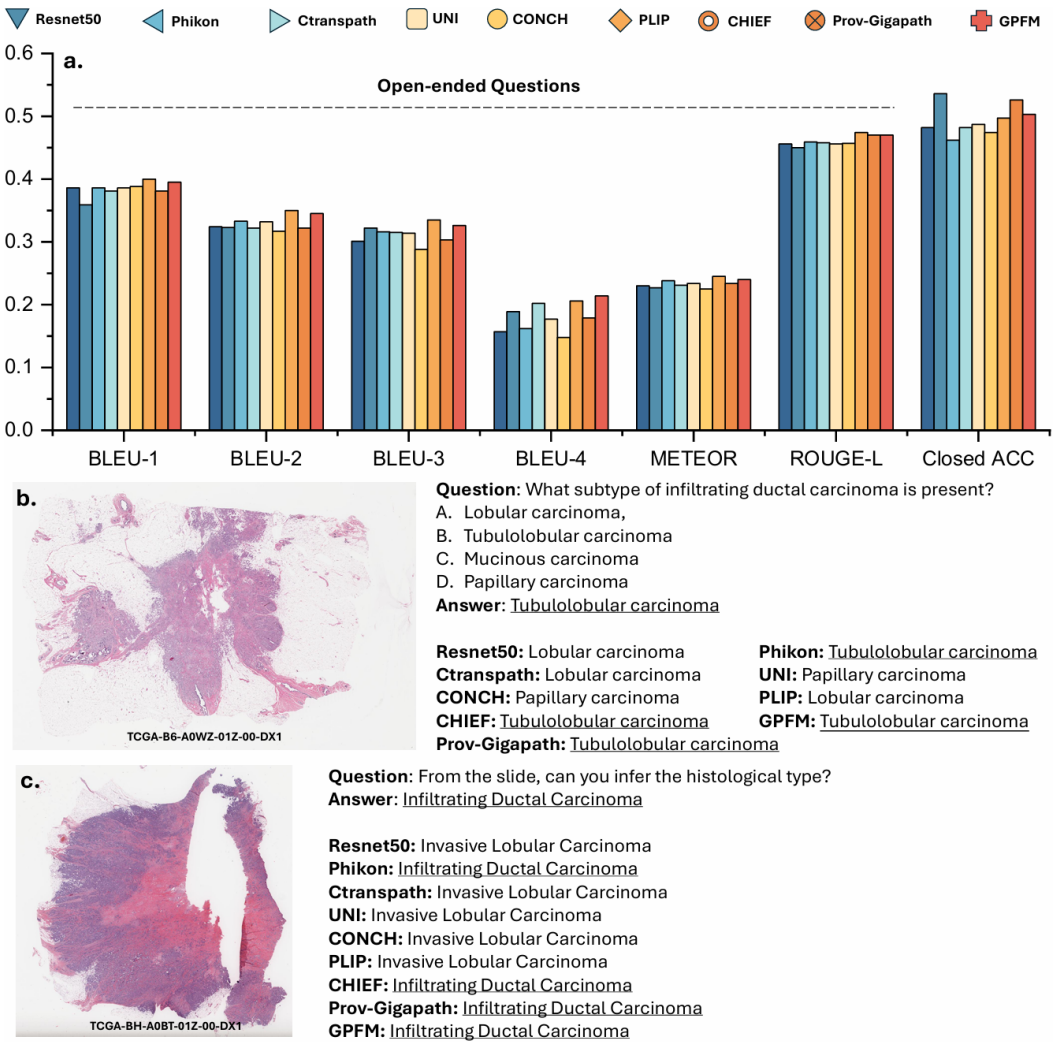


Fig. A6 VQA results on WSI-VQA dataset. a. Open-ended and close-ended statistical results. **b.** A close-ended question and corresponding answers. **c.**An open-ended question and corresponding answers.

Comment 2: also, please make the codes and data accessible now and not just at acceptance.

Response:

Thank you for your reminder regarding code and data accessibility. We are committed to promoting transparency and reproducibility in our work. To this end, we have made the following resources publicly available at https://github.com/birkhoffkiki/GPFM:

1. **Public Data and Splits:** All publicly available datasets and the corresponding splits used for experiments are provided (see below Table 2).
2. **Pretraining Code:** The code for pretraining the GPFM model is included.
3. **Downstream Task Code:** The code for evaluating the model on downstream tasks is also provided.

Regarding the data from medical centers, these datasets are not publicly available due to patient privacy obligations, institutional review board requirements, and data use agreements. However, researchers interested in accessing de-identified data may submit a reasonable request directly to corresponding authors, subject to obtaining the

necessary ethical approvals and complying with institutional policies. We hope this addresses your concerns and facilitates further research in the community. Please let us know if additional clarification or resources are needed.

Table 2. The link of public data used in this work.

Dataset	Link or Source
1. TCGA [68]	https://portal.gdc.cancer.gov/
2. CPTAC [69]	https://proteomic.datacommons.cancer.gov/pdc/
3. PANDA [73]	https://www.kaggle.com/c/prostate-cancer-grade-assessment/data
4. NADT-Prostate [116]	https://www.cancerimagingarchive.net/collection/nadt-prostate/
5. BCNB [117]	https://bcnb.grand-challenge.org/
6. CAMELYON16 [70]	https://camelyon16.grand-challenge.org/Data/
7. CAMELYON17 [71]	https://camelyon17.grand-challenge.org/Data/
8. BRACS [72]	https://www.bracs.icar.cnr.it/download/
9. TIGER2021 [118]	https://tiger.grand-challenge.org/
10. MIDOG2022 [119]	https://midog.deepmicroscopy.org/download-dataset/
11. AGGC2022 [120]	https://aggc22.grand-challenge.org/
12. O.B.R. [121, 122]	https://www.cancerimagingarchive.net/collection/ovarian-bevacizumab-response/
13. ACROBAT2023 [123]	https://acrobat.grand-challenge.org/
14. AML-C-LMU [124]	https://www.cancerimagingarchive.net/collection/aml-cytomorphology_lmu/
15. ARCH [125]	https://warwick.ac.uk/fac/cross_fac/tia/data/arch
16. BACH [88]	https://zenodo.org/records/3632035
17. CAMEL [126]	https://drive.google.com/open?id=1brr8CnU6ddzAYT157wkdXjbSzoiIDF9y
18. DiagSet [127]	https://ai-econsilio.diag.pl/
19. DLBCL [128]	https://github.com/stanfordmlgroup/DLBCL-Morph
20. GTEx [129]	https://gtexportal.org/home/histologyPage
21. HunCRC [130]	https://www.cancerimagingarchive.net/collection/hungarian-colorectal-screening/
22. Janowczyk [131]	https://andrewjanowczyk.com/use-case-1-nuclei-segmentation/
23. LC25000 [132]	https://academictorrents.com/details/7a638ed187a6180fd6e464b3666a6ea0499af4af
24. MIDOG2021 [119]	https://imig.science/midog2021/download-dataset/
25. OCELOT [133]	https://zenodo.org/record/7844149
26. Oste. Tumor [134]	https://www.cancerimagingarchive.net/collection/osteosarcoma-tumor-assessment/
27. PAIP2019 [135]	https://paip2019.grand-challenge.org/
28. PAIP2020 [136]	https://paip2020.grand-challenge.org/
29. PAIP2021	https://paip2021.grand-challenge.org/
30. Post-NAT-BRCA [137]	https://www.cancerimagingarchive.net/collection/post-nat-brca/
31. SICAPv2 [138]	https://data.mendeley.com/datasets/9xxm58dvs3/1
32. SLN-Breast [139]	https://www.cancerimagingarchive.net/collection/sln-breast/
33. SPIE2019 [140]	https://breastpathq.grand-challenge.org/
34. TUPAC [141]	https://tupac.grand-challenge.org/
35. UBC-OCEAN [75]	https://www.kaggle.com/competitions/UBC-OCEAN/data
36. RCC-DHMC [142]	https://bmirds.github.io/KidneyCancer/
37. CRC-100K [49]	https://zenodo.org/records/1214456
38. CRC-MSI [91]	https://zenodo.org/records/3832231
39. CCRCC-TCGA-HEL [87]	https://zenodo.org/records/7898308
40. PanCancer-TCGA [92]	https://zenodo.org/records/5889558
41. PanCancer-TIL [93]	https://zenodo.org/records/6604094
42. ESCA[95]	https://zenodo.org/records/7548828
43. PCAM[96]	https://github.com/basveeling/pcam
44. BreakHis [89]	https://www.kaggle.com/datasets/ambarish/breakhis
45. UniToPatho [90]	https://ieee-dataport.org/open-access/unitopatho
46. Chaoyang [99]	https://github.com/bupt-ai-cz/HSA-NRL
47. PathVQA [101]	https://github.com/UCSD-AI4H/PathVQA
48. HistGen [51]	https://github.com/dddavid4real/HistGen
49. IMP-CRS [77–79]	https://rdm.inesctec.pt/dataset/nis-2023-008
50. HANCOCK [80]	https://github.com/ankilab/HANCOCK_MultimodalDataset
51. GasHisDB [143]	https://figshare.com/ndownloader/files/28969725
51. PatchGastricADC22 [54]	https://zenodo.org/records/6550925
52. WSI-VQA [102]	https://github.com/cpystan/WSI-VQA

Comment 3: there are inconsistencies in the formatting and organization of affiliations and other information.

Response:

Thank you for your reminder. We have now standardized the formatting to ensure consistency throughout the manuscript. Specifically, we have applied the following rules:

1. University Affiliations:

Format: Department, University, City, Country

Example: Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China.

2. Hospital Affiliations:

Format: Department, Hospital, Affiliated University, City, Country

Example: Department of Pathology, Nanfang Hospital and School of Basic Medical Sciences, Southern Medical University, Guangzhou, China.

3. Laboratory Affiliations:

Format: Laboratory, City, Country

Example: Shanghai Artificial Intelligence Laboratory, Shanghai, China

We have carefully reviewed and updated all affiliations in the manuscript to adhere to these formats. We hope this revision resolves the issue and improves the overall readability and professionalism of the paper.

Towards A Generalizable Pathology Foundation Model via Unified Knowledge Distillation

Jiabo Ma[1†], Zhengrui Guo[1†], Fengtao Zhou[1], Yihui Wang[1], Yingxue Xu[1], Jinbang Li[2,3], Fang Yan[4], Yu Cai[5], Zhengjie Zhu[6], Cheng Jin[1], Yi Lin[1], Xinrui Jiang[1], Chenglong Zhao[2,3,7], Danyi Li[2,3], Anjia Han[8], Zhenhui Li[9], Ronald Cheong Kin Chan[10], Jiguang Wang[11,12], Peng Fei[13], Kwang-Ting Cheng[1,5], Shaoting Zhang[4,14*], Li Liang[2,3,15*], Hao Chen[1,11,12,16,17*]

[1]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China.

[2]Department of Pathology, Nanfang Hospital and School of Basic Medical Sciences, Southern Medical University, Guangzhou, China.

[3]Guangdong Provincial Key Laboratory of Molecular Tumor Pathology, Guangzhou, China.

[4]Shanghai Artificial Intelligence Laboratory, Shanghai, China.

[5]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China.

[6]Information Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China.

[7]Department of Pathology, The First Affiliated Hospital of Shandong First Medical University and Shandong Provincial Qianfoshan Hospital, Jinan, China.

[8]Department of Pathology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China.

[9]Department of Radiology, The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming, China.

[10]Department of Anatomical and Cellular Pathology, The Chinese University of Hong Kong, Hong Kong SAR, China.

[11]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China.

[12]Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China.

[13]School of Optical Electronic Information, Huazhong University of Science and Technology, Wuhan, China.

[14]Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China.

[15]Jinfeng Laboratory, Chongqing, China.

[16]State Key Laboratory of Molecular Neuroscience, The Hong Kong University of Science and Technology, Hong Kong SAR, China.

[17]Shenzhen-Hong Kong Collaborative Innovation Research Institute, The Hong Kong University of Science and Technology, Shenzhen, China.

Comment 4: typos such as missing punctuation and inconsistent spacing

Response:

Thank you for your valuable feedback regarding the typos, missing punctuation, and inconsistent spacing in our manuscript. We have carefully reviewed the document and made the necessary corrections to improve clarity and consistency. Your attention to detail is greatly appreciated, and we believe these changes enhance the overall quality of the work. To make the response clear, we provide **some examples** listed below:

Example for missing punctuation:

Previous: "Across the three external validation datasets, the GPFM achieved the best average rank of 1.5, while the second best performing model, UNI, obtained an average rank of 2.3, illustrating the generalization capability of GPFM."

Revised: *"In contrast, by integrating knowledge from all adopted expert models, the unified knowledge distillation enables GPFM to surpass the performance of individual models, achieving a significantly lower average ranking score of 1.88, outperforming the next-best model by more than one point. This underscores GPFM's strength as a highly generalizable FM."* (Please note that we add more tasks, therefore, we rewrite text related to the task and results.)

Example for inconsistent spacing:

Previous: *"*This dataset was adopted for the subtyping of NSCLC using lung squamous cell carcinoma (LSCC) and LUAD WSIs sourced from the CPTAC data portal[79].*"*

Revised: *"To perform subtyping of non-small cell lung cancer (NSCLC), we utilized data from the TCGA [68], CPTAC [69], and Center-1."* (Please note that we adopted WSIs from medical Center-1 to conduct external validation. To reduce the redundancy, we introduce NSCLC related data in one paragraph.)

The previous manuscript with marks is included in the attached files, which indicates the modifications.

Comment 5: some sections redundantly restate similar content ... this creates unnecessary length

Response:

We are grateful for the suggestions. We added 33 new tasks and rewrote several parts to avoid redundancy. We updated the manuscript based on your suggestion, and **some examples** are listed below:

Examples:

- In section 2.3, **Previous:** *"*The performance of WSI classification is influenced by both the feature extractor and the MIL method. In the WSI classification tasks, we use the ABMIL to evaluate whether the features extracted by the foundation models are discriminative. In Region of Interest (ROI) classification tasks, we can directly evaluate the feature representation abilities of the foundation models without using any MIL method.*"*

Revised: *"The performance of WSI classification is influenced by both the feature extractor (i.e., FM) and the MIL method. Unlike WSI classification, Region-of-Interest (ROI) classification tasks allow for a direct assessment of the FMs' feature representation capabilities, independent of MIL methods."*

- **Removed:** *"The comparison of various foundation models is illustrated in Figure 5."*
- In section 2.4, **Removed:** *"The experimental results for ROI retrieval are depicted in Fig. 6a, and the detailed results can be found in Extended Data Table A33, showcasing the Top-1, Top-3, and Top-5 accuracy achieved by different foundation models."*
- In section 2.4, **move the dataset introduction into section 4.**
- In section 2.5, **Removed:** *"The performance of different foundation models on open-ended and close-ended VQA problems is presented in Figure 6, and detailed results can be found in Extended Data Table A34."*
- In section 2.6, **Removed**: *"The performance of each foundation model in report generation is presented in Figure 7 and the detailed results are reported in Extended Data Table A35."*
- In section 2.7, **Removed**: *"The experimental results are presented in Figure 8. More details can be found in Extended Data Table A36."*
- In section 4, **we removed some sentences** that have been mentioned in section 2 to reduce redundancy. For example, in section 4.2, we removed *"WSI classification holds significant importance in pathology diagnosis. It plays a crucial role in accurately analyzing and interpreting WSI, enabling pathologists to make informed diagnostic decisions."*.
- In section 4, **we removed the redundant description** of survival analysis data and **used a table** to show the details of data.

The previous manuscript with marks is included in the attached files, which indicates the modifications.

Comment 6: poor writing with complex and lengthy sentences make the text difficult to follow at times

Response:

Thank you for your valuable feedback regarding the writing style of our manuscript. We are sorry for the issues caused by complex and lengthy sentences, which made the text difficult to follow. In response to your comment, we have thoroughly revised the manuscript to improve clarity and readability. Specifically, we have:

1. **Simplified Sentence Structures:** Broken down lengthy and complex sentences into shorter, more concise statements to enhance readability.
2. **Improved Flow:** Reorganized paragraphs to ensure a logical and smooth progression of ideas.
3. **Removed Redundancies:** Eliminated unnecessary words and phrases to make the text more direct and accessible.

Below, we provide ***some examples*** of specific modifications made to address these issues:

In section 1:
Previous: "The ability of GPFM to consistently perform well across a diverse type of clinical tasks highlights the benefits of leveraging knowledge distillation to combine the strengths of expert models, ultimately leading to more robust and versatile foundation models for supporting clinicians and improving patient care."
Revised: "*The consistent performance of GPFM across a diverse range of clinical tasks underscores the advantages of employing knowledge distillation to integrate the strengths of specialized expert models. This approach facilitates the development of more robust and versatile foundation models (FMs), thereby enhancing their utility in supporting clinical decision-making and advancing patient care outcomes.*"
Please refer to Page 4 in the manuscript.

In section 2.2:
Previous: "In the context of clinical trials for oncology, survival analysis is commonly employed, with the time to an event, such as death or disease progression, serving as the primary outcome under investigation [51–54]."
Revision: We removed this sentence.

In section 2.7:
Previous: "Not only did the individual task performances improve significantly, but the average performance also exhibited enhancement, with notable improvements in all three metrics."
Revised: "*The experimental results demonstrated significant improvements not only in the performance of individual tasks but also in the overall average performance, with substantial enhancements observed across all three evaluation metrics.*"
Please refer to Page 13 in the manuscript.

In section 3:
Previous: "To further maximize the diversity of data used for pretraining, we gathered 190 million images sourced from 47 sources, spanning 34 major tissue types. This rich dataset, combined with our advanced pretraining methodology, empowers GPFM to surpass current foundation models in performance across six major categories (WSI classification, survival analysis, ROI classification, image retrieval, VQA, and report generation) of CPath tasks, comprising a total of 39 specific tasks."
Revised: "*To further maximize the diversity of data used for pretraining, we gathered 190 million images sourced from 56 sources, spanning 34 major tissue types. This rich dataset, combined with our advanced pretraining methodology, empowers GPFM to surpass current FMs in performance across 72 CPath tasks.*"
Please refer to Page 13 in the manuscript.

In section 4.1:

Previous: "It is important to note that, for all images used in the downstream tasks, the feature extraction operations were performed on images resized to 224×224, unless specified otherwise."
Revised: "*For all downstream tasks, it should be emphasized that feature extraction was consistently performed on images resized to 224×224 resolution, except where explicitly stated otherwise in the experimental protocol.*"
Please refer to Page 15 in the manuscript.

In section 4.2:
Previous: "Our experiments encompass 12 pathology WSI classification tasks, including (1) breast cancer metastasis detection, (2) coarse-grained breast carcinoma subtyping, (3) fine-grained breast carcinoma subtyping, (4) lobular \& ductal carcinoma subtyping, (5) ovarian cancer subtyping, (6) renal cell carcinoma (RCC) subtyping based on TCGA, (7) non-small cell lung cancer subtyping based on TCGA, (8) non-small cell lung cancer subtyping based on CPTAC, (9) prostate cancer grade assessment, (10) TP53 mutation prediction on LUAD (TCGA), (11) Brain tumor subtyping, and (12) glioma IDH1 mutation prediction. "
Revised: "*To evaluate the performance of the MIL model, we assess the balanced accuracy, weighted F1 score, and AUC, which consider the class imbalance present in the dataset. Our experiments encompass 36 pathology WSI classification tasks, including 20 internal and 16 external validation datasets.*"
Please refer to Page 16 in the manuscript.

We believe these revisions significantly improve the overall clarity and comprehension of the manuscript. We appreciate your constructive feedback and hope that the revised version meets your expectations. Please let us know if further improvements are needed.

Reviewer #3:

The authors evaluated the generalizability of foundation models in computational pathology. They found that while existing foundation models excel in specific areas, their performance varies across a broader range of applications. The authors propose a knowledge distillation framework combining expert knowledge distillation, which integrates insights from multiple models, and self-knowledge distillation, which enhances image representation through local-global alignment. Using this framework, they developed the Generalizable Pathology Foundation Model (GPFM) and evaluated its performance on tasks including cancer classification and pathology report generation. Results show that GPFM had an average rank of 1.36 among the models they compared with.

Response:

Thank you for sharing your insights about our model GPFM. Our team has worked diligently to develop a framework that not only combines expert knowledge distillation but also leverages self-knowledge distillation to create a more robust and adaptable model. We believe this work represents an important step forward in making AI technologies more reliable and applicable across different pathological contexts. Your comments are really helpful for improving the quality of this manuscript, we responded to your comments one by one in below.

Comment 1. The criteria for task selection are unclear. For example, The Cancer Genome Atlas (TCGA) datasets they used can support approximately 20 whole-slide image classification tasks and another 20 survival prediction tasks, but only a subset was selected and presented. Different subset selections could alter comparative results.

Response:

Thank you for your comment. We appreciate the opportunity to clarify the task selection process for the TCGA datasets. For the TCGA-related tasks, we followed the tasks commonly used in previous foundation models (UNI [1] and Phikon [2]) and our previous works [3][4]. The selection criteria were based on clinical relevance, dataset size, and diversity of cancer types to ensure a robust evaluation of foundation models.

We acknowledge the importance of extensive task validation to verify the effectiveness of GPFM. Based on your comments and those of Reviewer #2, we have added an additional 33 tasks (28 of which are non-TCGA), bringing the total number of tasks to 72. These new tasks can provide broader coverage of cancer types and to address potential limitations in the original task set. Among these, 20 are external validation tasks, and 13 are internal tasks, ensuring a rigorous evaluation of GPFM's generalizability and robustness. The new tasks are detailed in the following table.

Task Name	Internal data (Data Number)	External Data (Data Number)

NSCLC Subtyping	TCGA-NSCLC (1053 WSIs)	CENTER-1-NSCLC (210 WSIs)
Metastatic Detection (Lung Cancer)	CENTER-1-LMD2 (1198 WSIs)	CENTER-2-LMD2 (530 WSIs)
Primary Site Prediction (Lung Cancer)	CENTER-1-LMD6 (1198 WSIs)	CENTER-2-LMD6 (530 WSIs)
RCC Subtyping	TCGA-RCC (937 WSIs)	CENTER-3-RCC (88 slides)
ILC & IDC classification	TCGA-BRCA (985 WSIs)	CENTER-3-LD (383 WSIs)
Breast Carcinoma Subtyping	BRACS (545 WSIs)	CENTER-3-BRCA (467 WSIs)
IDH Mutation Prediction in Glioma[(1)]	TCGA-GBMLGG-IDH1 (979 WSIs)	EBRAIN-IDH1 (852)
Ovarian Cancer Subtyping	UBC-OCEAN (527 WSIs)	CENTER-3-Ovary (370 WSIs)
Brain Tumor Subtyping[(1)]	TCGA-GBMLGG-Subtyping (1276 WSIs)	EBRAIN-Subtyping (732)
Lesion Grade Classification (Colon Cancer)	IMP-CRS (5332 WSIs)	CENTER-3-Colon-WSI (297 WSIs)
Primary Site Prediction (Head & Neck Cancer)	HANCOCK (708 WSIs)	-
T stage Prediction (Head & Neck Cancer)	HANCOCK (705 WSIs)	-
Lauren Subtyping (Gastric Cancer)	TCGA-STAD (390 WSIs)	☐ CENTER-5-Lauren (141 WSIs) ☐ CENTER-4-Lauren (319 WSIs)
Vascular Invasion Detection (Gastric Cancer)	CENTER-1-Vascular (396 WSIs)	☐ CENTER-5-Vascular (230 WSIs) ☐ CENTER-4-

		Vascular (319 WSIs)
Perineural Invasion Detection (Gastric Cancer)	CENTER-1-Perineural (397 WSIs)	☐ CENTER-5-Perineural (232 WSIs) ☐ CENTER-4- Perineural (319 WSIs)
Survival Analysis (Head & Neck Cancer)	TCGA-HNSC (443 cases)	HANCOCK (749 cases)
Survival Analysis (Brain Cancer)(2)	☐ TCGA-GBM (372 cases) ☐ TCGA-LGG (462 cases)	-
Tumor infiltrating lymphocyte (TIL) classification	Pancancer-TIL (304,097 patches)	CENTER-3-TIL (18,492 patches)
Colon Tissue Classification	Chaoyang (6,160 patches)	CENTER-3-Colon (21,068 patches)
Gastic Tissue Classification	GasHisDB (13,124 patches)	CENTER-3-GC (2,537 patches)
VQA for WSI	WSI-VQA (977 WSIs)	-
Report Generation (Gastric Cancer)	PatchGastricADC22 (991 WSIs)	-

(1) In the previous manuscript, this task only performed internal training and testing on EBRAIN dataset. In the revised manuscript, we use the TCGA-GBMLGG as internal data and use the EBRAIN as the external data.

(2) In the previous manuscript, this task is performed on TCGA-GBMLGG data. However, the comment 6 raised by reviewer #3 suggests that GBM and LGG should predict separately due to distinct pathology imaging profiles and very different prognoses. Therefore, we reformulate this task.

The ranking scores of various foundation models across the expanded set of tasks are presented in the accompanying Figure 1. As shown, GPFM achieves the best average performance on both the 52 internal tasks and the 20 external validation tasks. This robust performance across a diverse set of tasks, including 28 non-TCGA and 20 external validation tasks, underscores the generalizability of GPFM and the effectiveness of our proposed expert knowledge distillation pretraining strategy. We believe these additions and clarifications address the reviewer's concerns and further validate the generalizability and robustness of our model.

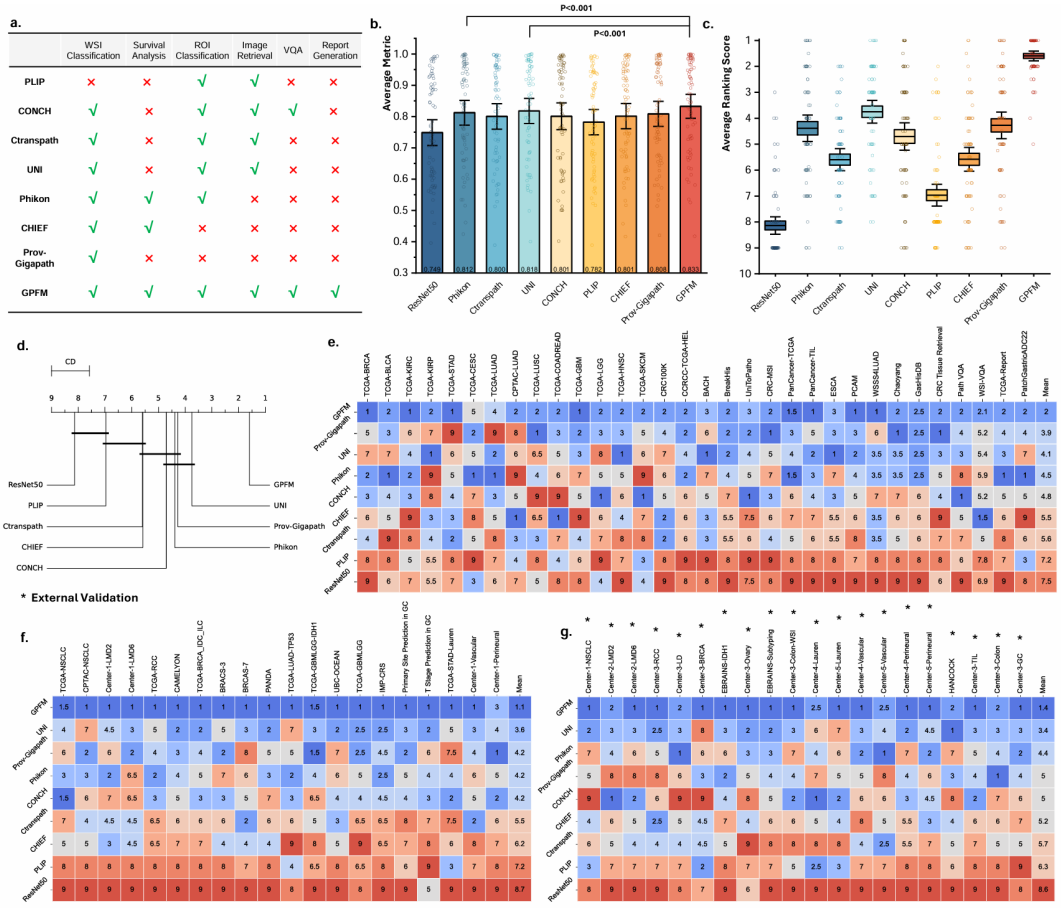


Figure 1. The ranking score of foundation models across 72 tasks.

[1] 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.

[2] Filiot A, Ghermi R, Olivier A, et al. Scaling self-supervised learning for histopathology with masked image modeling[J]. medRxiv, 2023: 2023.07. 21.23292757.

[3] Zhou F, Chen H. Cross-modal translation and alignment for survival analysis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 21485-21494.

[4] Zhou, Fengtao, et al. "iMD4GC: Incomplete Multimodal Data Integration to Advance Precise Treatment Response Prediction and Survival Analysis for Gastric Cancer." *arXiv preprint arXiv:2404.01192* (2024).

Comment 2. Related to the previous point, the authors did not stratify the pathology visual question answering, report generation, and image retrieval tasks by cancer type in their current analysis. If these results were stratified by cancer type, would the results change?

Response: We appreciate the reviewer's insightful suggestion to analyze performance across different types of cancer. Following this recommendation, we stratified our evaluation by major organs (breast, lung, and kidney). The number of test cases for each organ is 188, 157, and 175 respectively. The analysis revealed that our GPFM model demonstrated superior performance for kidney-related report generation, achieving a BLEU-4 score of 0.255 compared to Phikon's 0.247. For breast and lung cancer reports,

Phikon maintained its leading position with BLEU-4 scores of 0.194 and 0.179 respectively, compared to GPFM's scores of 0.182 and 0.152. Based on the experimental results, the conclusion remains same stratified by cancer type. Detailed results for all models are shown in the following Table. Please refer to **Extended Data Table A41** in the revised manuscript.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
			Breast			
ResNet50	0.228±0.007	0.079±0.004	0.034±0.003	0.016±0.002	0.081±0.002	0.157±0.004
Phikon	**0.416±0.006**	**0.312±0.005**	**0.251±0.004**	**0.208±0.004**	**0.194±0.003**	**0.364±0.005**
Ctranspath	0.254±0.007	0.131±0.004	0.078±0.002	0.047±0.002	0.097±0.004	0.192±0.004
UNI	0.361±0.006	0.255±0.005	0.198±0.004	0.161±0.003	0.162±0.004	0.306±0.005
CONCH	0.265±0.006	0.163±0.005	0.113±0.004	0.084±0.003	0.117±0.003	0.226±0.005
PLIP	0.269±0.005	0.148±0.003	0.093±0.003	0.061±0.002	0.106±0.003	0.201±0.002
CHIEF	0.272±0.010	0.151±0.006	0.093±0.004	0.060±0.003	0.117±0.004	0.209±0.005
Prov-Gigapath	0.334±0.007	0.230±0.004	0.175±0.005	0.139±0.003	0.148±0.004	0.278±0.005
GPFM	0.390±0.007	0.289±0.004	0.231±0.005	0.192±0.004	0.182±0.003	0.346±0.006
			Lung			
ResNet50	0.224±0.008	0.085±0.005	0.035±0.003	0.016±0.002	0.078±0.002	0.159±0.004
Phikon	**0.405±0.009**	**0.284±0.006**	**0.211±0.005**	**0.162±0.004**	**0.179±0.004**	**0.346±0.007**
Ctranspath	0.150±0.011	0.066±0.006	0.032±0.004	0.015±0.002	0.058±0.004	0.131±0.006
UNI	0.329±0.009	0.220±0.006	0.158±0.004	0.119±0.003	0.140±0.004	0.279±0.006
CONCH	0.229±0.008	0.134±0.005	0.088±0.004	0.061±0.003	0.091±0.003	0.188±0.005
PLIP	0.198±0.007	0.079±0.005	0.035±0.004	0.014±0.003	0.072±0.003	0.139±0.004
CHIEF	0.209±0.012	0.098±0.007	0.044±0.004	0.019±0.003	0.079±0.004	0.167±0.007
Prov-Gigapath	0.308±0.009	0.199±0.006	0.140±0.004	0.102±0.003	0.132±0.004	0.260±0.006
GPFM	0.349±0.008	0.235±0.006	0.173±0.005	0.132±0.004	0.152±0.004	0.300±0.007
			Kidney			
ResNet50	0.426±0.006	0.281±0.004	0.202±0.003	0.153±0.003	0.187±0.002	0.320±0.004
Phikon	0.500±0.007	0.375±0.006	0.300±0.005	0.247±0.005	0.225±0.004	0.406±0.005
Ctranspath	0.415±0.011	0.269±0.007	0.193±0.005	0.147±0.004	0.184±0.005	0.318±0.007
UNI	0.450±0.006	0.333±0.005	0.267±0.005	0.222±0.004	0.201±0.004	0.364±0.005
CONCH	0.420±0.006	0.280±0.005	0.203±0.004	0.156±0.004	0.185±0.003	0.318±0.004
PLIP	0.400±0.006	0.259±0.004	0.185±0.003	0.141±0.002	0.171±0.002	0.303±0.003
CHIEF	0.384±0.004	0.233±0.005	0.153±0.004	0.106±0.003	0.153±0.004	0.280±0.005
Prov-Gigapath	0.416±0.006	0.292±0.005	0.224±0.005	0.179±0.004	0.184±0.004	0.329±0.005
GPFM	**0.504±0.008**	**0.381±0.006**	**0.307±0.005**	**0.255±0.004**	**0.226±0.004**	**0.407±0.005**

Comment 3. Some of the selected foundation models were not fully evaluated. For example, the PLIP feature can be used for whole-slide pathology image classification and survival analyses. In addition, CONCH, Ctranspath, and UNI features can be used for survival analyses. However, these foundation models were not evaluated for these tasks.

Response:

We appreciate the opportunity to clarify this point. All foundation models, including PLIP, CONCH, CTransPath, and UNI, were evaluated across all 72 tasks, encompassing WSI classification, survival analysis, and other relevant tasks. The potential confusion may stem from **Fig. 1b, 1c, and 1d**. In **Fig. 1b and 1c**, we only plotted the top 4 models based on their average ranking scores to ensure clarity in the visualization. In **Fig. 1d**, we evaluated all six models across the 72 tasks (including WSI classification, image retrieval, ROI classification, VQA, report generation, and survival analysis) and calculated the average performance for each task type to assist in expert selection. The full evaluation results, including the performance of all models across all tasks, are detailed in **Fig. 2e-g**. This comprehensive evaluation demonstrates that all foundation models were rigorously assessed for their capabilities in the specified tasks.

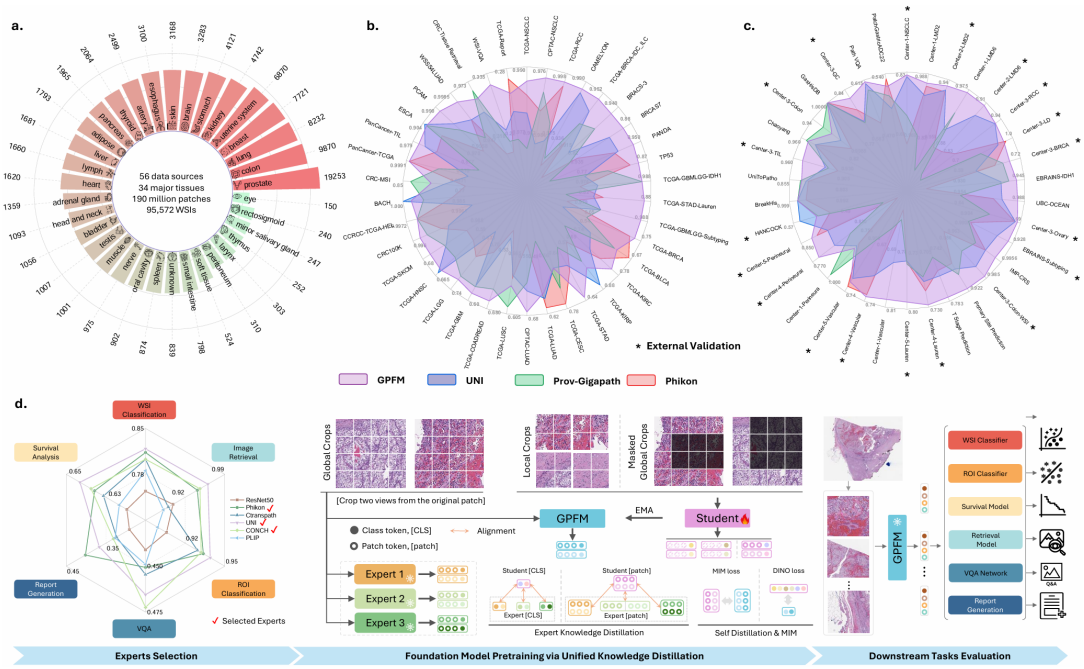


Fig. 1 The overview of GPFM.

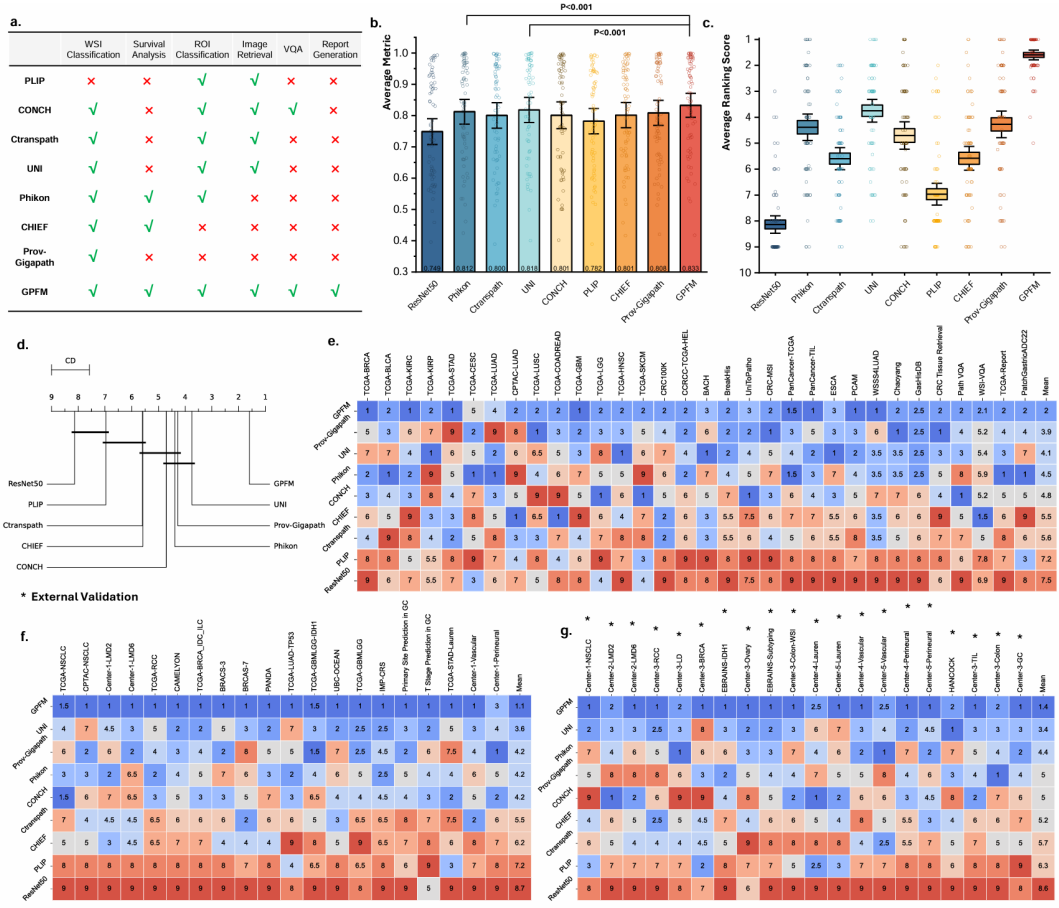


Fig 2. Overall performance.

Comment 4. Several new foundation models showed better performance in the tasks presented in the manuscript. These models include GigaPath, CHIEF, and Virchow V2.

However, these models were not included in the current analyses.

Response:

We appreciate your comment regarding the latest foundation models. In response, we have conducted additional experiments to compare our method with two publicly available slide-level foundation models, CHIEF [1] and Prov-GigaPath [2]. As shown in Fig. 2c-d, from a ranking perspective, GPFM, UNI, and Prov-GigaPath achieved the top 3 positions, respectively. From an average metric perspective, GPFM, UNI, and Phikon attained the top 3 positions, respectively. GPFM and UNI are still the most robust pathology foundation models. Regarding Virchow V2, while we acknowledge its potential, it is currently excluded from our analysis due to the lack of peer-reviewed validation and restricted access to its implementation. We would like to emphasize that the core contribution of our work lies in proposing a novel knowledge distillation framework capable of integrating information from multiple state-of-the-art foundation models. This framework remains valid and applicable even as newer and more powerful models emerge in the future.

[1] Wang, Xiyue, et al. "A pathology foundation model for cancer diagnosis and prognosis prediction." Nature 634.8035 (2024): 970-978.

[2] Xu, Hanwen, et al. "A whole-slide foundation model for digital pathology from real-world data." *Nature* (2024): 1-8.

Comment 5. The GPFM model for survival prediction showed substantial performance decay when applied to the publicly available Clinical Proteomic Tumor Analysis Consortium (CPTAC) external validation dataset. The authors could discuss the potential implications of this finding.

Response:

Thank you for highlighting this issue. Upon revisiting our experiments, we identified a bug related to flipped censorship labels in the survival analysis tasks. After fixing this issue and re-evaluating the results across 15 survival analysis tasks (shown in Fig. 4 and Fig. A3), we observed the following:

1. For the LUAD survival analysis on TCGA and CPTAC data (Fig. A3h), while Phikon achieved the best performance on TCGA, it performed the worst on CPTAC. In contrast, GPFM ranked second on CPTAC, with performance only slightly lower than CHIEF.
2. In another experiment using TCGA-HNSC as the internal dataset and HANCOCK as the external dataset, GPFM achieved the second-best performance, slightly lower than UNI.

These results highlight that no single model consistently dominates all survival analysis tasks, underscoring the importance of developing robust models that generalize well across diverse cancer types. We have discussed these findings in Section 2.2.

"It is noteworthy that survival analysis tasks are inherently more challenging than WSI classification, and no single model has been able to dominate these tasks (Fig. 2e). The experimental results from both WSI classification and survival analysis highlight the

limited generalization capability of existing FMs. This limitation is likely attributable to the data distribution of their training sets and the pretraining methods they employ. While existing FMs exhibit limited generalization, they demonstrate exceptional performance on specific types of tasks. By leveraging their individual strengths, it is possible to construct a more powerful and versatile model. This is precisely what we have achieved in this study: we propose a unified distillation framework to distill the capabilities of existing models--particularly in tasks where they excel--into the GPFM, thereby significantly enhancing its generalization ability."

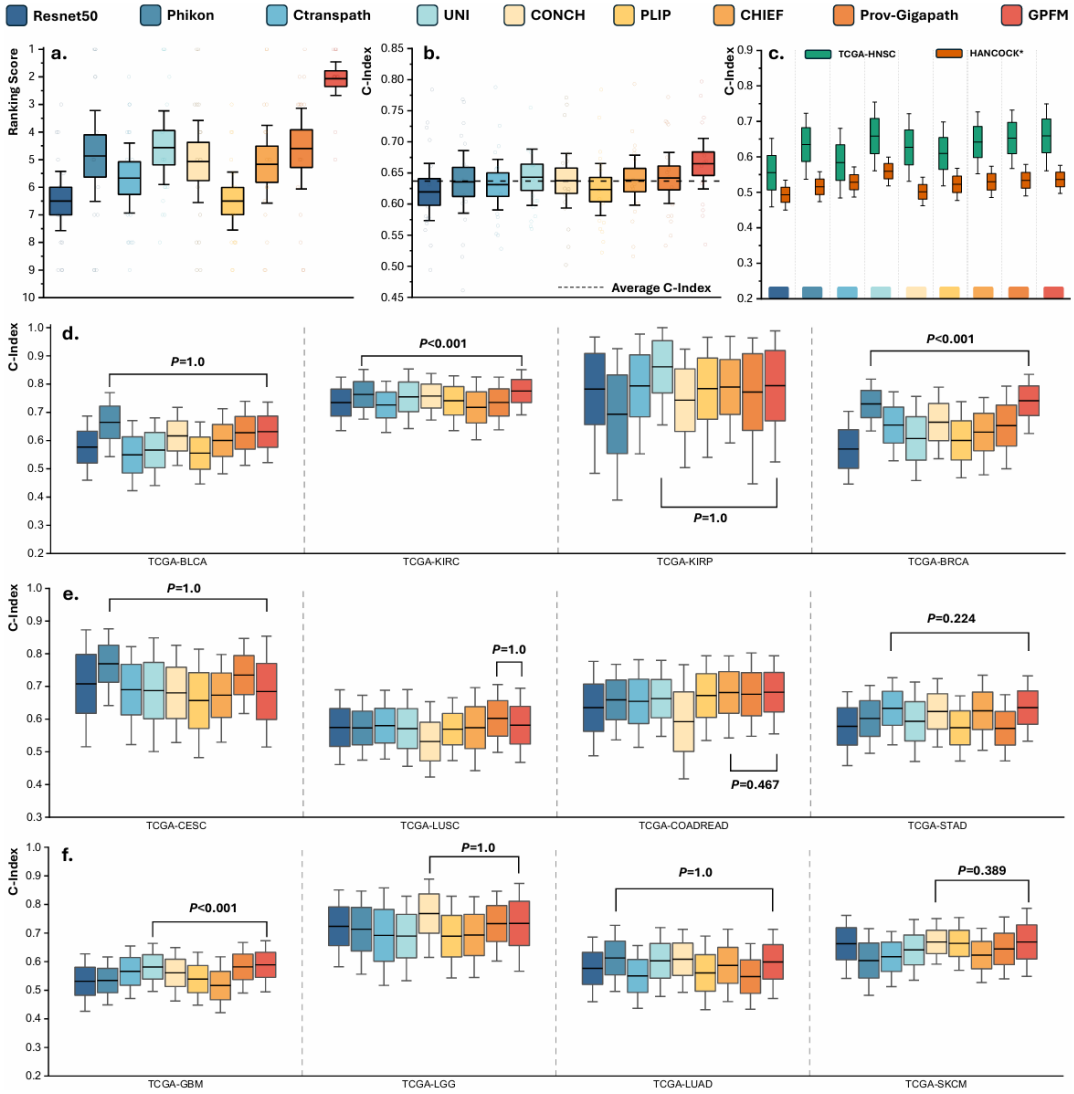


Fig. 4 Performance of FMs across 15 Survival Analysis Tasks. a. Average ranking of FMs in 15 survival analysis tasks. **b.** Average C-Index of various FMs across 15 tasks. **c.** Results on TCGA-HNSC data and the HANCOCK cohort. The survival prediction model was trained on the TCGA-HNSC cohort and subsequently tested on the HANCOCK cohort. **d-f.** C-Index of FMs across 12 survival analysis tasks. In all subfigures, error bars indicate 95% CI. For box plots, the center line represents the mean, and the box limits represent the standard error.

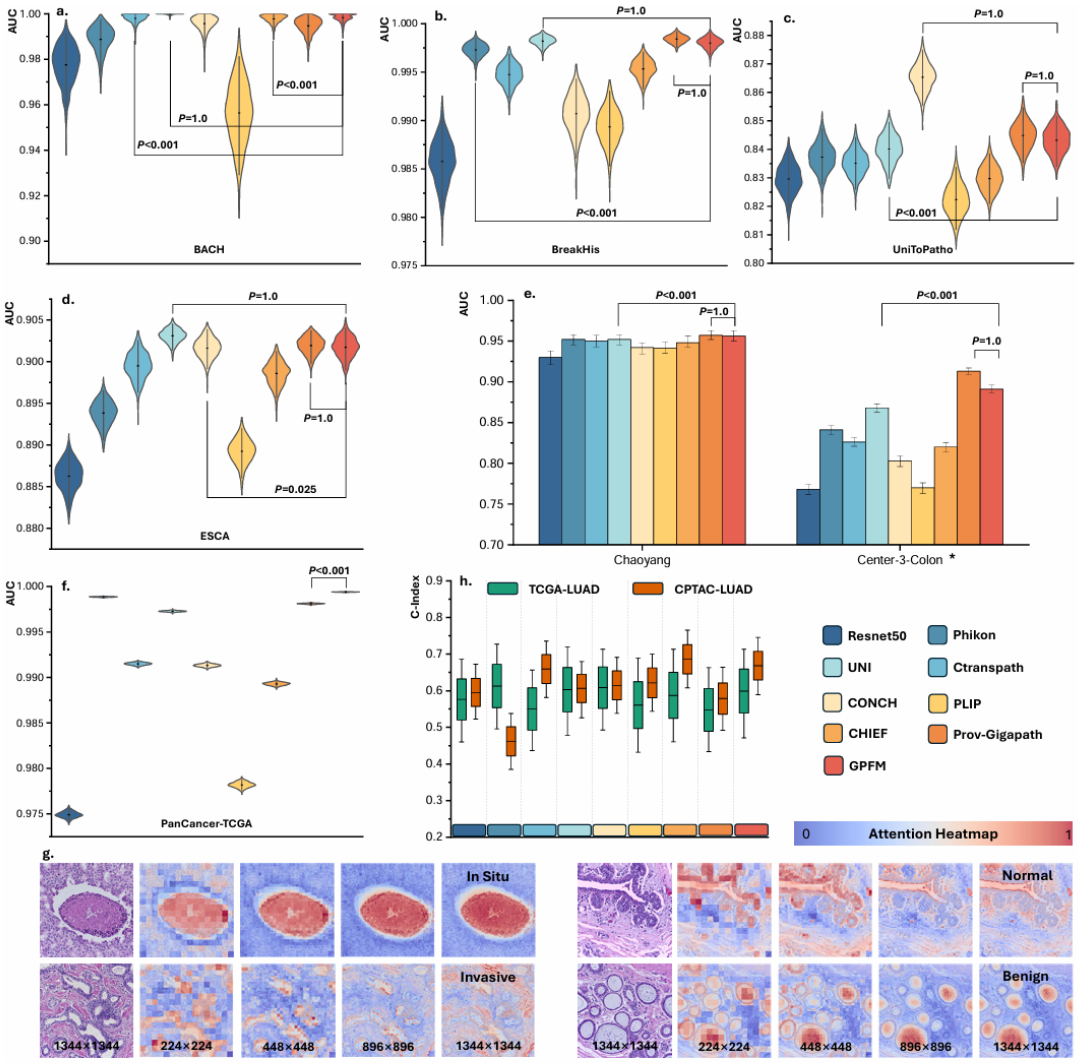


Fig. A3 Extended Result of ROI Classification Tasks. a-d. The AUC of foundation models on BACH, BreakHis, UniToPatho, and ESCA, respectively. **e.** The colon tissue classification performance. The Chaoyang and Center-3-Colon serve as internal and external, respectively. **f.** The performance of pancancer classificaiton of different foundation models. **g.**Attention heatmap of GPFM across various image resolutions for BRCA subtyping in BACH dataset. The colored squares represent the 14×14 [PATCH] tokens encoded by the GPFM model. The heatmap values indicate the similarity between each [PATCH] token and the [CLS] token generated by the last layer of GPFM, measured using Euclidean distance. The consistent attention patterns observed across varying image resolutions and tissue types underscore the robust capabilities of the GPFM model. **h.** Results on TCGA-LUAD data and the CPTAC-LUAD cohort. The survival prediction model was trained on the TCGA-LUAD cohort and subsequently tested on the CPTAC-LUAD cohort.

Comment 6. Glioblastoma (GBM) and low-grade glioma (LGG) have distinct pathology imaging profiles and very different prognoses. They should be separated in the survival outcome prediction. The pooled analyses shown in the current manuscript do not have clinical significance.

Response:

We thank the reviewer for this insightful comment regarding the distinct nature of GBM and LGG. We fully agree that these two types of gliomas have distinct pathological characteristics and prognostic profiles. To address this important point, we conducted separate analyses by splitting the TCGA-GBMLGG dataset into TCGA-GBM and TCGA-LGG cohorts. As shown in **Table A22**, our GPFM model demonstrates robust performance across both individual cohorts:

1. For GBM patients, GPFM achieved the highest performance (C-index: 0.590, 95% CI: 0.500-0.676) among all compared methods, suggesting its superior capability in predicting outcomes for this aggressive tumor type.
2. For LGG patients, GPFM achieved the second-highest performance (C-index: 0.731, 95% CI: 0.562-0.872) behind CONCH (C-index: 0.771).
3. Notably, GPFM maintained consistent performance across both subtypes, with relatively narrow confidence intervals, indicating its robust generalizability.

Table A22 Performance of Survival Analysis on TCGA-COADREAD, TCGA-GBM, and TCGA-LGG datasets. 5-fold cross validation is adopted for training and evaluation. The 95% CI is included in parentheses. The best and second-best performed model are **bolded** and underlined.

	TCGA-COADREAD	TCGA-GBM	TCGA-LGG
ResNet50	0.632 (0.496-0.767)	0.533 (0.434-0.629)	0.723 (0.566-0.860)
Phikon	0.660 (0.534-0.762)	0.534 (0.452-0.613)	0.710 (0.555-0.850)
Ctranspath	0.656 (0.522-0.784)	0.565 (0.475-0.647)	0.693 (0.493-0.868)
UNI	0.662 (0.543-0.779)	0.580 (0.498-0.659)	0.687 (0.527-0.830)
CONCH	0.593 (0.434-0.766)	0.559 (0.474-0.651)	**0.771 (0.635-0.893)**
PLIP	0.673 (0.540-0.805)	0.539 (0.447-0.629)	0.684 (0.538-0.816)
CHIEF	**0.682 (0.557-0.803)**	0.516 (0.422-0.605)	0.694 (0.551-0.824)
Prov-Gigapath	0.675 (0.537-0.802)	0.581 (0.496-0.663)	0.728 (0.599-0.845)
GPFM	0.678 (0.552-0.788)	**0.590 (0.500-0.676)**	0.731 (0.562-0.872)

Comment 7. The pathology visual question answering dataset appears to be very noisy, and the meaning of the labels is unclear. For example, what does "polycystic disease infant" mean? A more precise term might be "polycystic kidney disease of the infant." Similarly, the example of "What is present? Answer: cardiovascular" is also unclear. A better description could be, "What is the tissue type shown in this pathology image? Answer: Blood vessels."

Response:

Thank you for your insightful comment regarding the PathVQA dataset. We acknowledge that some labels in the PathVQA dataset, such as "polycystic disease infant" and "What is present? Answer: cardiovascular," may appear unclear. The PathVQA dataset was constructed from two publicly available pathology textbooks ("Textbook of Pathology" and "Basic Pathology") and the Pathology Education Informational Resource (PEIR) digital library. During dataset construction, the original sentences were simplified to avoid overly complex syntactic structures, which may have introduced some noise and ambiguity in the labels. However, the key information in the question-answer pairs remains intact, ensuring that the evaluation of different models on this dataset is still fair, as demonstrated in **Fig. 6**.

To further validate our findings, we also evaluated the models on a WSI-level VQA dataset [2], as shown in **Fig. A6**. Across both datasets, GPFM achieved the second-best performance, demonstrating its robustness and strong overall performance.

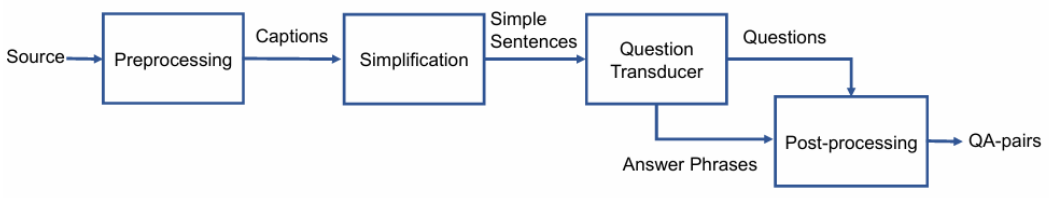


Figure 3: The framework of generating questions from captions

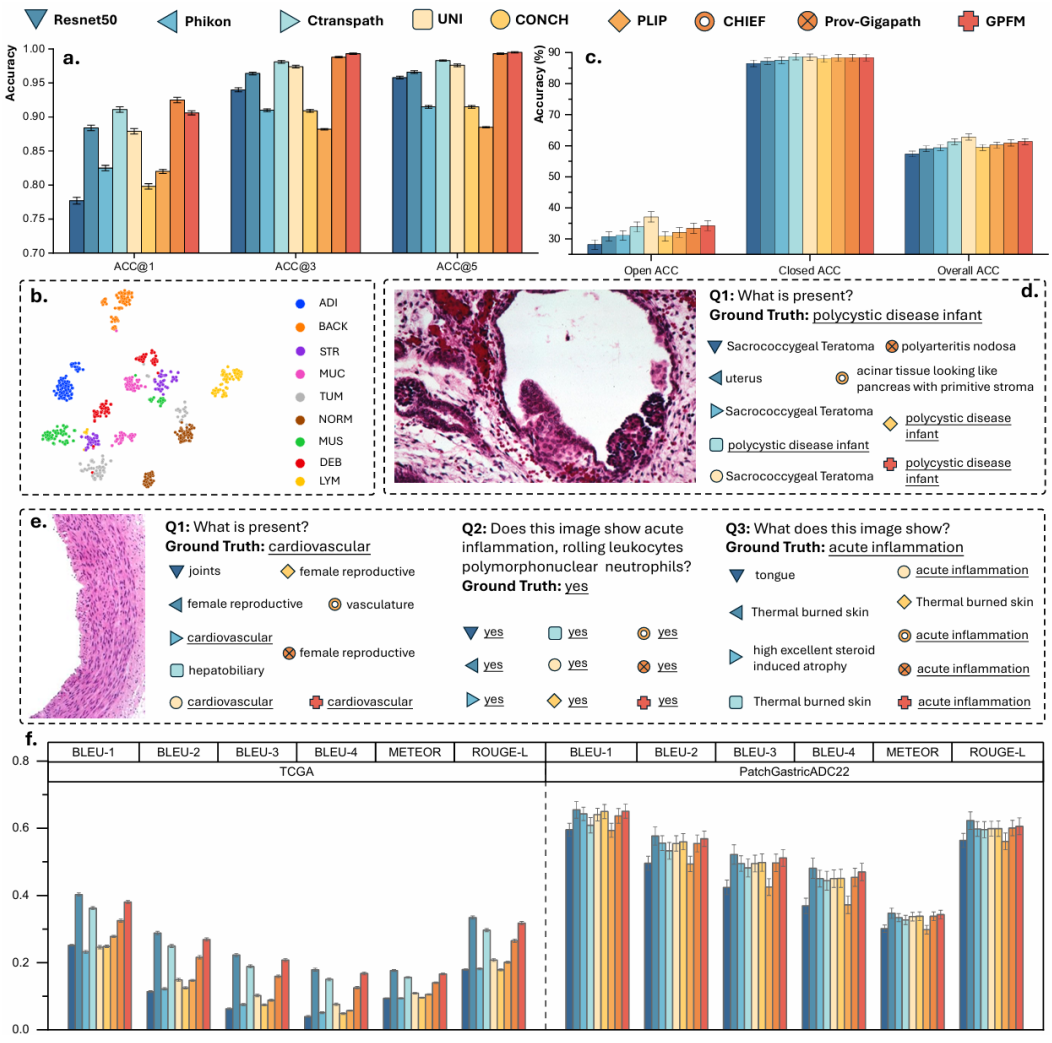


Fig. 6 Overview of Pathology Tissue Retrieval, VQA, and Report Generation. a. The top-1, top-3, top-5, and average accuracy of different FMs on pathology tissue retrieval tasks. **b.** The distribution of features extracted by GPFM. For each class, 100 samples from the test set were used, and a total of 900 samples were subjected to t-SNE dimensionality reduction to 2D. **c.** The performance of VQA on PathVQA dataset, measured by open-ended accuracy, closed-ended accuracy, and overall accuracy, for different FMs. **d.** An open-ended question along with the answers generated by various FMs. **e.** Three questions and the answers generated by FMs related to the query image. **f.**The performance of report generation on TCGA and PatchGastricAD22 data. The models are measured by six different metrics. In all subfigures, the error bars indicate standard deviation.

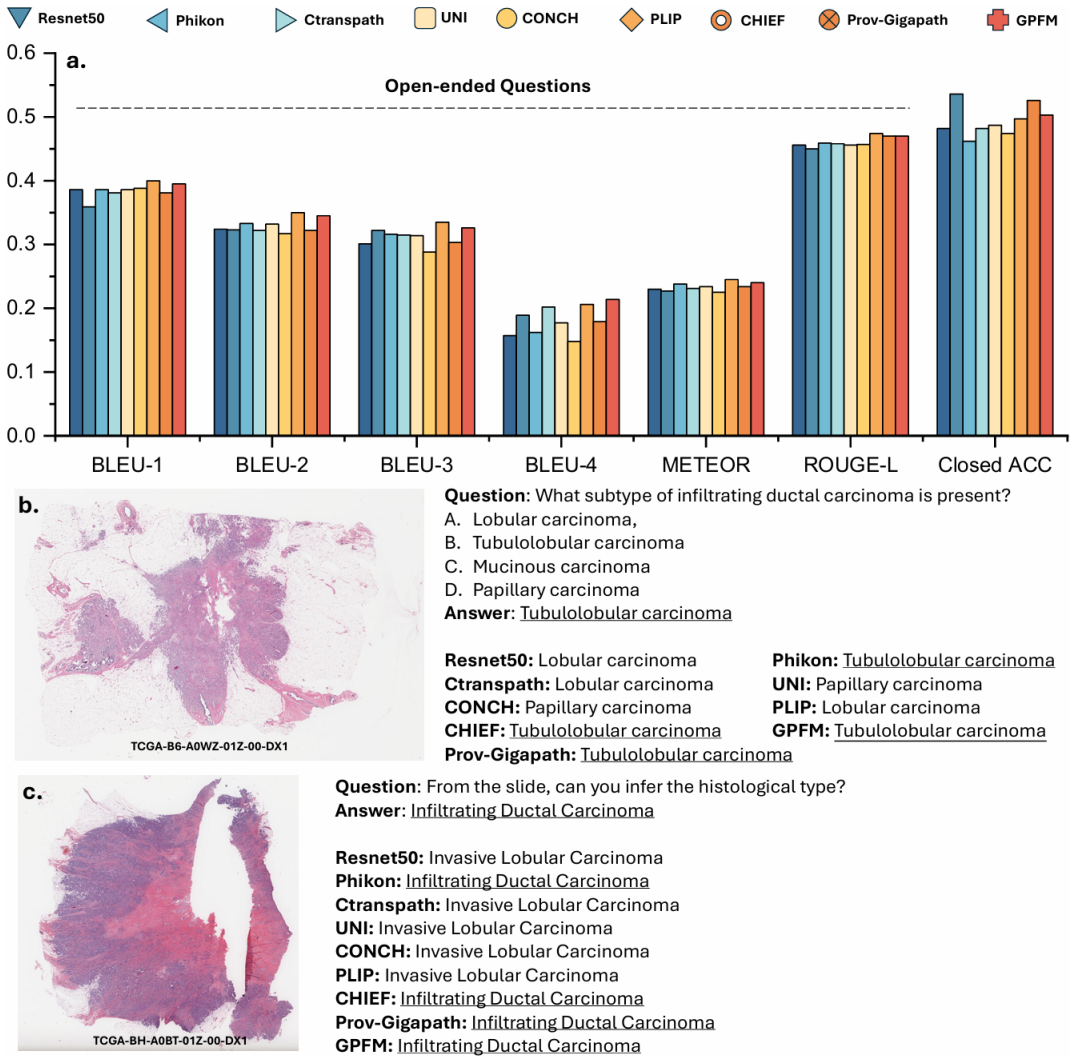


Fig. A6 VQA results on WSI-VQA dataset. a. Open-ended and close-ended statistical results. **b.** A close-ended question and corresponding answers. **c.**An open-ended question and corresponding answers.

[1] He X, Zhang Y, Mou L, et al. Pathvqa: 30000+ questions for medical visual question answering[J]. arXiv preprint arXiv:2003.10286, 2020.
[2] Chen, Pingyi, et al. "Wsi-vqa: Interpreting whole slide images by generative visual question answering." *ECCV*. Springer, Cham, 2025.

Comment 8. The BLEU scores presented in the pathology report generation task are low (<0.4). In addition, Phikon has performed better than the method proposed by the authors across all evaluation metrics. The authors could further investigate the performance of Phikon in non-TCGA datasets for this task.

Response: We thank the reviewer for this insightful comment regarding model performance evaluation on report generation. Regarding the low BLEU scores in the TCGA WSI-report dataset, this is attributable to several key factors:

1. The pathology reports in the TCGA WSI-report dataset are comprehensive and highly detailed, often containing multiple sections describing various aspects like

tumor characteristics, tissue architecture, cellular features, and molecular markers.
2. Pathologists frequently use different but equally valid medical terminology and sentence structures to describe the same pathological findings, making exact phrase matching (which BLEU measures) less likely.
3. The sequential nature of report writing means that similar observations might be documented in different orders, further impacting n-gram based metrics like BLEU.

To rigorously assess the generalizability of our approach, we conducted additional evaluations on the PatchGastricADC22 Dataset [1], which comprises 991 paired diagnostic captions of stomach adenocarcinoma endoscopic biopsy specimens. This dataset provides an important validation as it represents a different institutional source and cancer type from the TCGA WSI-report dataset.

As shown in Table A41, our analysis reveals several key findings:

1. While Phikon still manages to achieve the highest scores (BLEU-1: 0.655±0.025, ROUGE-L: 0.623±0.026), our GPFM demonstrates strong competitive performance:
    a. BLEU-1: 0.651±0.021 (second-highest, only 0.004 difference)
    b. BLEU-2: 0.569±0.023 (second-highest)
    c. BLEU-3: 0.512±0.025 (second-highest)
    d. BLEU-4: 0.470±0.026 (second-highest)
2. When considering the complete evaluation metrics suite, GPFM consistently ranks among the top two performers across all metrics, demonstrating robust and stable performance.

These findings are consistent with the results in TCGA WSI-report Dataset, as shown in Table A40.

Table A41 Performance of foundation models in WSI report generation on PatchGastricADC22 dataset. The best performing model for each metric is **bolded** and the second-best performing model is underlined.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
ResNet50	0.596±0.019	0.496±0.021	0.424±0.022	0.369±0.023	0.301±0.011	0.564±0.021
Phikon	**0.655±0.025**	**0.577±0.027**	**0.522±0.029**	**0.481±0.030**	**0.347±0.015**	**0.623±0.026**
Ctranspath	0.643±0.020	0.556±0.022	0.495±0.023	0.450±0.025	0.334±0.012	0.598±0.022
UNI	0.609±0.023	0.533±0.025	0.482±0.027	0.444±0.028	0.327±0.014	0.596±0.024
CONCH	0.641±0.019	0.555±0.023	0.495±0.025	0.450±0.026	0.337±0.013	0.599±0.022
PLIP	0.650±0.021	0.560±0.024	0.498±0.026	0.451±0.027	0.338±0.013	0.599±0.023
CHIEF	0.594±0.021	0.494±0.023	0.425±0.025	0.372±0.026	0.298±0.013	0.561±0.025
Prov-Gigapath	0.637±0.022	0.555±0.025	0.497±0.026	0.454±0.027	0.338±0.013	0.601±0.023
GPFM	0.651±0.021	0.569±0.023	0.512±0.025	0.470±0.026	0.343±0.013	0.606±0.025

Table A40 Performance of foundation models in WSI report generation on TCGA WSI-Report dataset. The best performing model for each metric is **bolded** and the second-best performing model is underlined.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
ResNet50	0.252±0.003	0.113±0.003	0.062±0.003	0.039±0.003	0.093±0.001	0.179±0.002
Phikon	**0.404±0.005**	**0.290±0.005**	**0.225±0.005**	**0.181±0.005**	**0.178±0.003**	**0.336±0.005**
Ctranspath	0.254±0.004	0.131±0.003	0.079±0.003	0.052±0.003	0.097±0.002	0.189±0.003
UNI	0.363±0.005	0.250±0.005	0.189±0.005	0.151±0.004	0.156±0.003	0.298±0.005
CONCH	0.246±0.005	0.149±0.004	0.104±0.004	0.077±0.003	0.110±0.002	0.208±0.004
PLIP	0.265±0.004	0.135±0.003	0.080±0.003	0.053±0.003	0.102±0.002	0.188±0.003
CHIEF	0.278±0.003	0.147±0.003	0.088±0.003	0.057±0.002	0.105±0.002	0.201±0.003
Prov-Gigapath	0.325±0.005	0.216±0.005	0.159±0.004	0.125±0.004	0.140±0.002	0.265±0.005
GPFM	0.384±0.005	0.271±0.005	0.210±0.005	0.169±0.005	0.168±0.003	0.320±0.005

[1] Tsuneki, Masayuki, and Fahdi Kanavati. "Inference of captions from histopathological patches." International Conference on Medical Imaging with Deep Learning. PMLR, 2022.

Comment 9. In addition to the BLEU scores and related metrics, blinded human-based evaluation of the generated pathology reports will provide better insights into the quality of the generated texts.

Response:

We appreciate the reviewer's valuable suggestion regarding human evaluation. To address this, we conducted a rigorous evaluation in collaboration with an experienced pathologist. The assessment used a four-tier scoring system as described in **Fig. A5**.

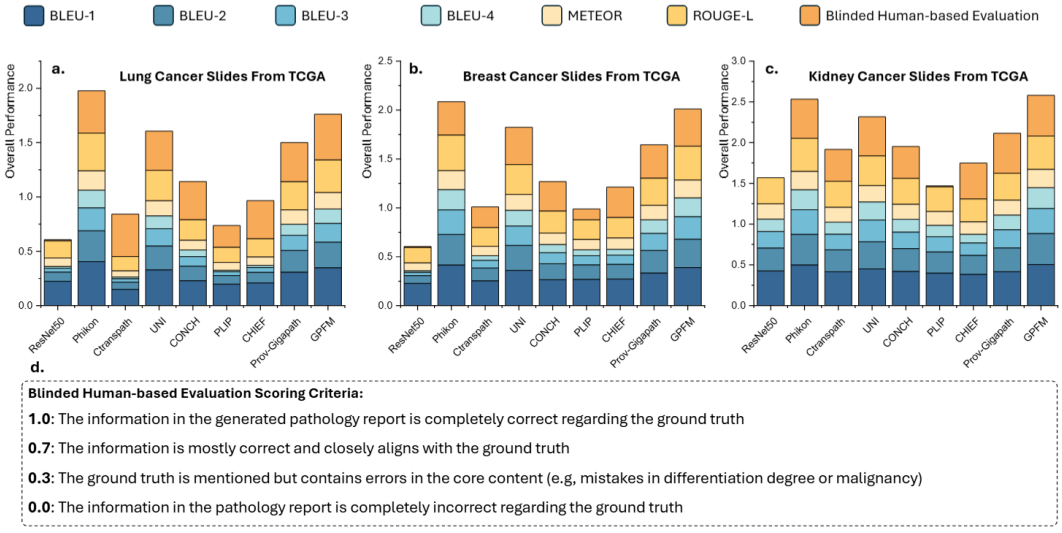


Fig. A5 Evaluation of Report Quality Based on Organ-Specific Analysis. a-c. Performance assessment of generated pathology reports for lung cancer, breast cancer, and kidney cancer, respectively. **d.** Scoring criteria for human-based blind evaluation of foundation-model-generated pathology reports. The scoring system ranges from 0.0 to 1.0, where 1.0 indicates complete accuracy with ground truth, 0.7 represents mostly correct information, 0.3 indicates presence of core content errors, and 0.0 denotes completely incorrect information.

We evaluated nine models across three distinct cancer types: breast (n=188), lung (n=157), and kidney (n=175). The results in Table A42 demonstrate several key findings:

1. GPFM consistently achieved superior performance:
    a. Breast cancer: tied for highest average score (0.38) with UNI
    b. Lung cancer: highest average score (0.42)

    c. Kidney cancer: highest average score (0.50)
2. Qualitative analysis reveals GPFM's strengths:
    a. Lowest number of completely incorrect reports (Score: 0) across all cancer types
    b. Highest proportion of largely accurate reports (Score: 0.7) for kidney cancer (90 reports)
    c. Strong performance in maintaining report accuracy for lung cancer (51 reports with Score: 0.7)
3. Cross-cancer consistency:
    a. GPFM showed progressive improvement from breast (0.38) to lung (0.42) to kidney (0.50) cancers
    b. Maintained consistently low error rates across all cancer types compared to baseline models like ResNet50

These human evaluation results complement our automated metrics and provide strong evidence for GPFM's capability to generate clinically relevant and accurate pathology reports across different cancer types.

Table A42 Human-based blind evaluation of foundation models in WSI report generation on TCGA WSI-report dataset, where the generated reports of breast, lung, and kidney cancers are used for evaluation. The number of reports in each score rated by the pathologist is listed and the average score is reported. The best performing model for each metric is **bolded** and the second-best performing model is underlined.

	Score: 0	Score: 0.3	Score: 0.7	Score: 1	Avg.
Breast					
ResNet50	184	4	0	0	0.01
Phikon	18	136	34	0	0.34
Ctranspath	80	88	20	0	0.21
UNI	9	134	45	0	**0.38**
CONCH	36	125	27	0	0.30
PLIP	118	69	1	0	0.11
CHIEF	25	139	24	0	0.31
Prov-Gigapath	24	127	37	0	0.34
GPFM	15	126	47	0	**0.38**
Lung					
ResNet50	153	3	1	0	0.01
Phikon	7	109	41	0	0.39
Ctranspath	19	87	51	0	0.39
UNI	13	109	35	0	0.36
CONCH	11	120	26	0	0.35
PLIP	53	104	0	0	0.20
CHIEF	15	113	29	0	0.35
Prov-Gigapath	13	111	33	0	0.36
GPFM	6	100	51	0	**0.42**
Kidney					
ResNet50	175	0	0	0	0.00
Phikon	5	86	84	0	0.48
Ctranspath	32	76	67	0	0.39
UNI	2	91	82	0	0.48
CONCH	11	118	46	0	0.39
PLIP	172	3	0	0	0.01
CHIEF	9	100	66	0	0.44
Prov-Gigapath	5	86	82	2	0.49
GPFM	2	83	90	0	**0.50**

Comment 10. It is interesting to see that DINOv2 without expert knowledge distillation

performs much better than the proposed methods in the BreakHis dataset. The authors could discuss the potential reasons behind this.

Response:

Thank you for your thoughtful review and observation about the performance of DINOv2 on the BreakHis dataset. The key contribution of our paper is developing a foundation model by expert and self knowledge distillation. Balancing the contributions of different experts is crucial. For instance, CONCH's performance (AUC) on the BreakHis dataset is only slightly better than ResNet50, indicating that this "less effective knowledge" may dominate the distillation process, thereby reducing the performance of GPFM in this task. Our experiments in other tasks (e.g., survival analysis) also illustrate that even when we distill the most powerful models into a single framework, it may not outperform experts in all tasks. There is an inherent trade-off in the distillation process. Finding the right balance among different experts and developing more effective distillation methods may be a promising direction for future research. To avoid any potential confusion of this part, we discussed this phenomenon in the ablation study section. The discussion is shown below:

"However, even with the distillation, GPFM still can not beat vanilla DINOv2 in all tasks such as Chaoyang and BreakHis, illustrating that there is still room for improving the distillation strategy."

Comment11. The authors did not compute the p-values for the tasks where GPFM performs worse. Adding these statistical analyses will help readers better understand the differences between GPFM and the better-performing models in these instances.

Response:

Thanks for your reminder. We adopted the Wilcoxon signed-rank one-side test to detect significant differences for all tasks. We computed the significance of our method and the best alternative method. The results are shown in the following figures.

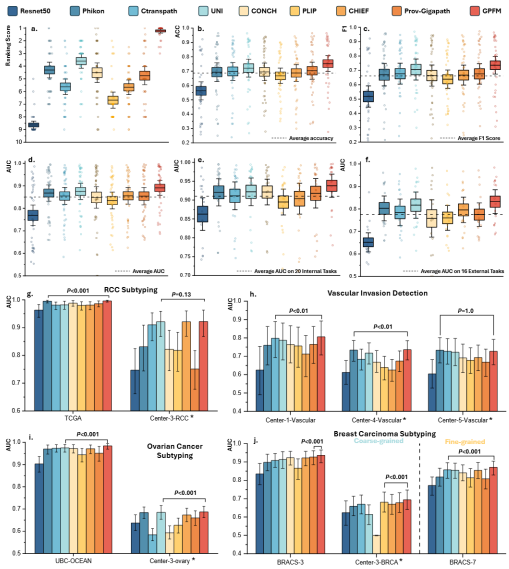


Fig. 3 Performance of FMs on WSI Classification Tasks. a. Average ranking of FMs based on AUC across 36 WSI classification tasks. **b-d.** Average balanced accuracy (ACC), and weighted F1 score (F1), and AUC of FMs across 36 WSI classification tasks. **e.** Average AUC of FMs on 20 internal WSI classification tasks. **f.** Average AUC of FMs on 16 external validation cohorts. **g-h.** Model performance on specific tasks: RCC subtyping, vascular invasion detection, ovarian cancer subtyping, and breast carcinoma subtyping. * represents external validation cohorts. Error bars represent 95% CI. Additional results are shown in Extended Data Fig. A1 and Fig. A2.

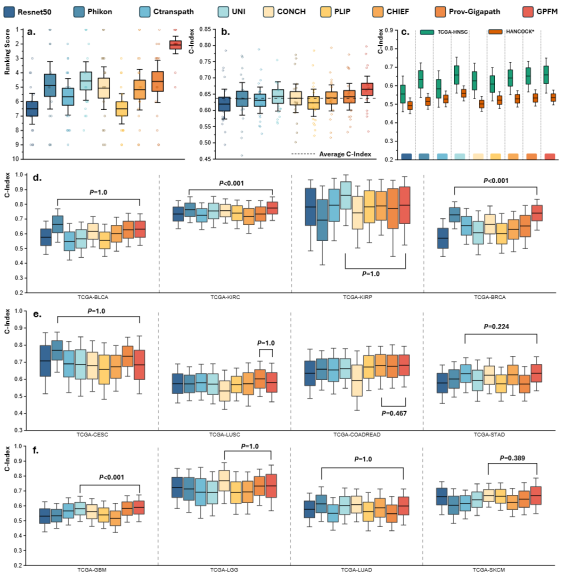


Fig. 4 Performance of FMs across 15 Survival Analysis Tasks. a. Average ranking of FMs in 15 survival analysis tasks. **b.** Average C-Index of various FMs across 15 tasks. **c.** Results on TCGA-HNSC data and the HANCOCK cohort. The survival prediction model was trained on the TCGA-HNSC cohort and subsequently tested on the HANCOCK cohort. **d-f.** C-Index of FMs across 12 survival analysis tasks. In all subfigures, error bars indicate 95% CI. For box plots, the center line represents the mean, and the box limits represent the standard error.

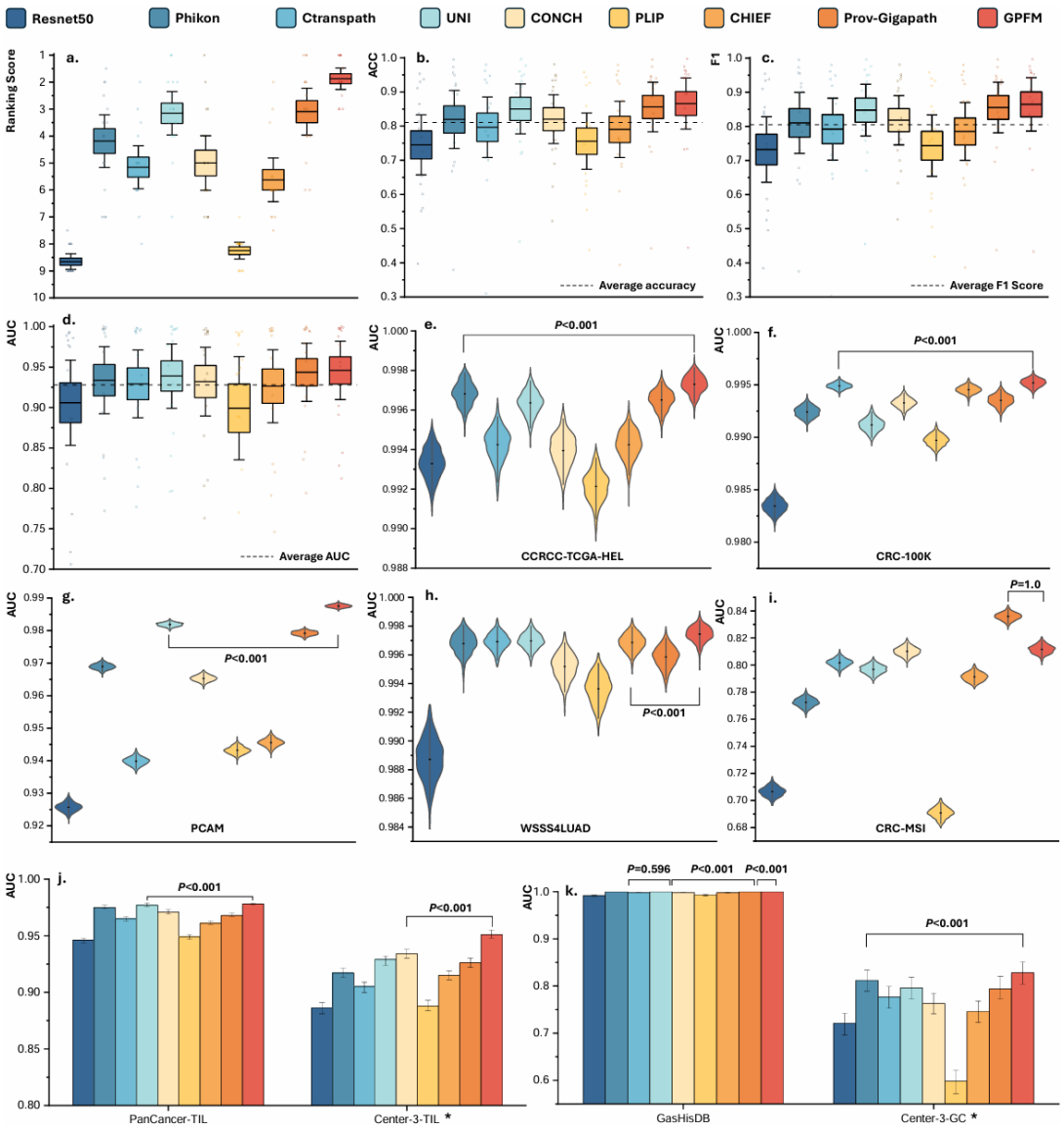


Fig. 5 Performance of FMs on Tissue Classification Tasks. a. Average ranking order of FMs based on AUC across 16 tasks. **b-d.** Average balanced accuracy (ACC), and weighted F1 score (F1), and AUC of FMs across 16 tasks. The center line represents mean and the box limits represents the standard error. **e-i.** AUC of FMs across 5 tissue classification tasks. The Wilcoxon signed-rank one-side test is adopted to detect significant difference. Then center black line in the violin plot represents the mean AUC. **j.** Tumor infiltrating lymphocytes classification based on the PanCancer-TIL (internal) and Center-3-TIL data (external). **k.** Gastric cancer tissue classificaiton with GasHisDB (internal) and Center-3-GC data (external). In all subfigures, the error bars indicate 95% CI. More results are presented in Extended Data **Fig.** A3.

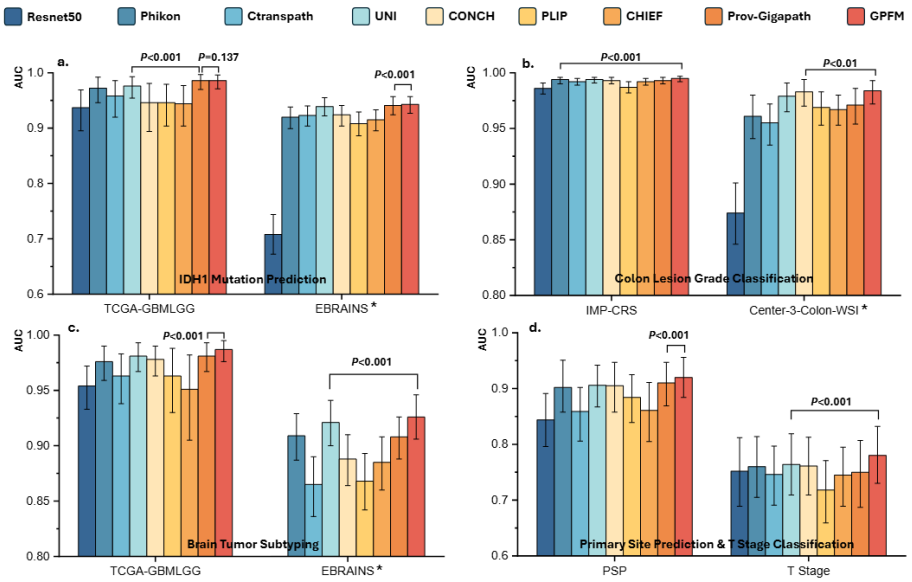


Fig. A2 Extended Results of WSI Classification. **a.** IDH-1 mutation prediction in brain tumors. **b.** Lesion grading in colon cancer. **c.** Brain tumor subtyping performance. **d.** Dual-task evaluation: primary site prediction and T-stage classification in head & neck cancer. Error bars represent 95% CI. External validation cohorts are marked with *.

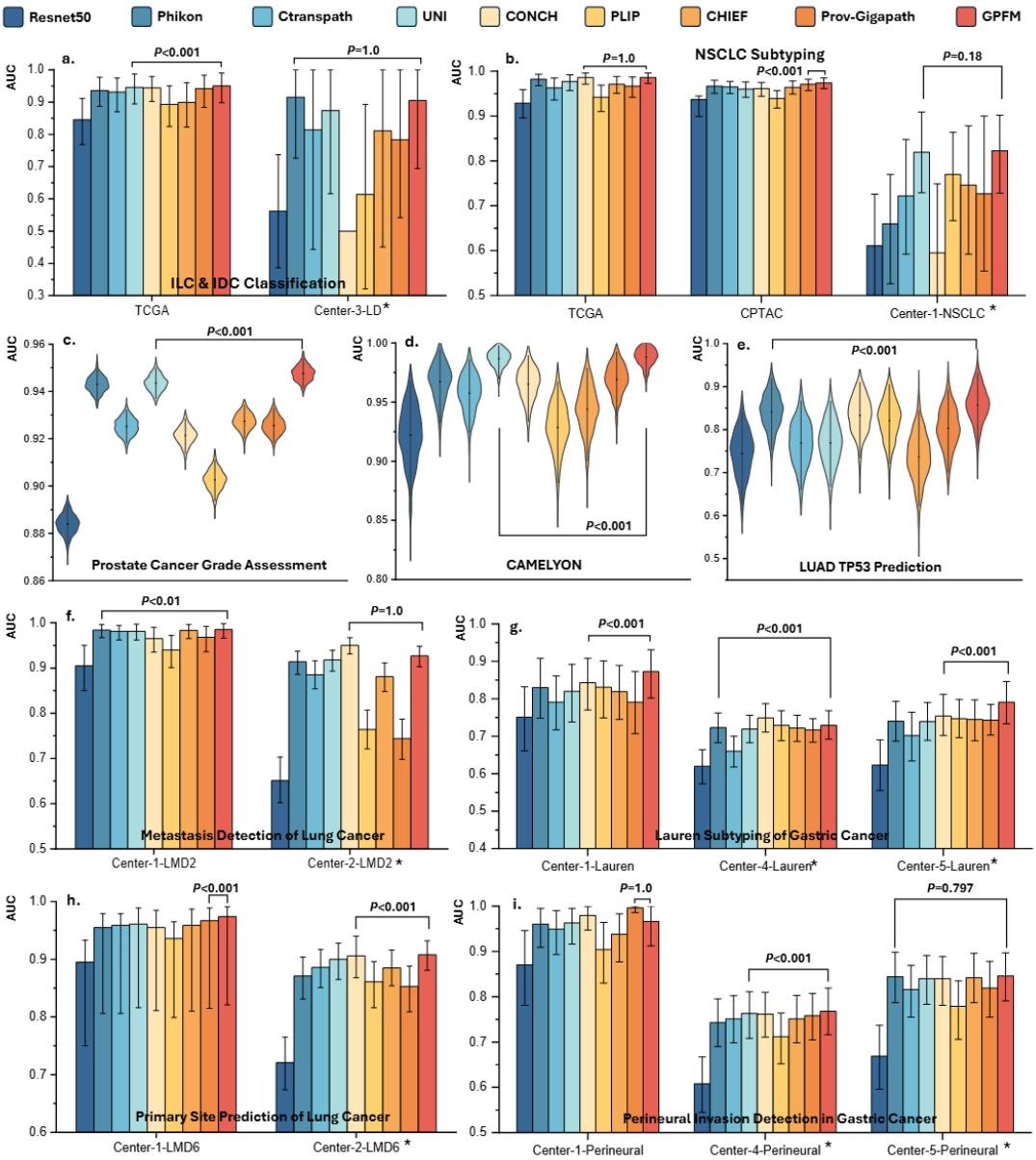


Fig. A1 Extended Results of WSI Classification. **a.** Performance comparison of foundation models in ILC and IDC classification. **b.** NSCLC subtyping performance across models. **c-e.** Model performance in prostate cancer grading, breast cancer metastasis detection, and LUAD TP53 mutation prediction, respectively. **f-i.** Extended evaluation including lung cancer metastasis detection, gastric cancer Lauren subtyping, lung cancer primary site prediction, and gastric cancer perineural invasion detection. Violin plots show the distribution of 1,000 bootstrap replicates. Error bars represent 95% CI. External validation cohorts are marked with *.

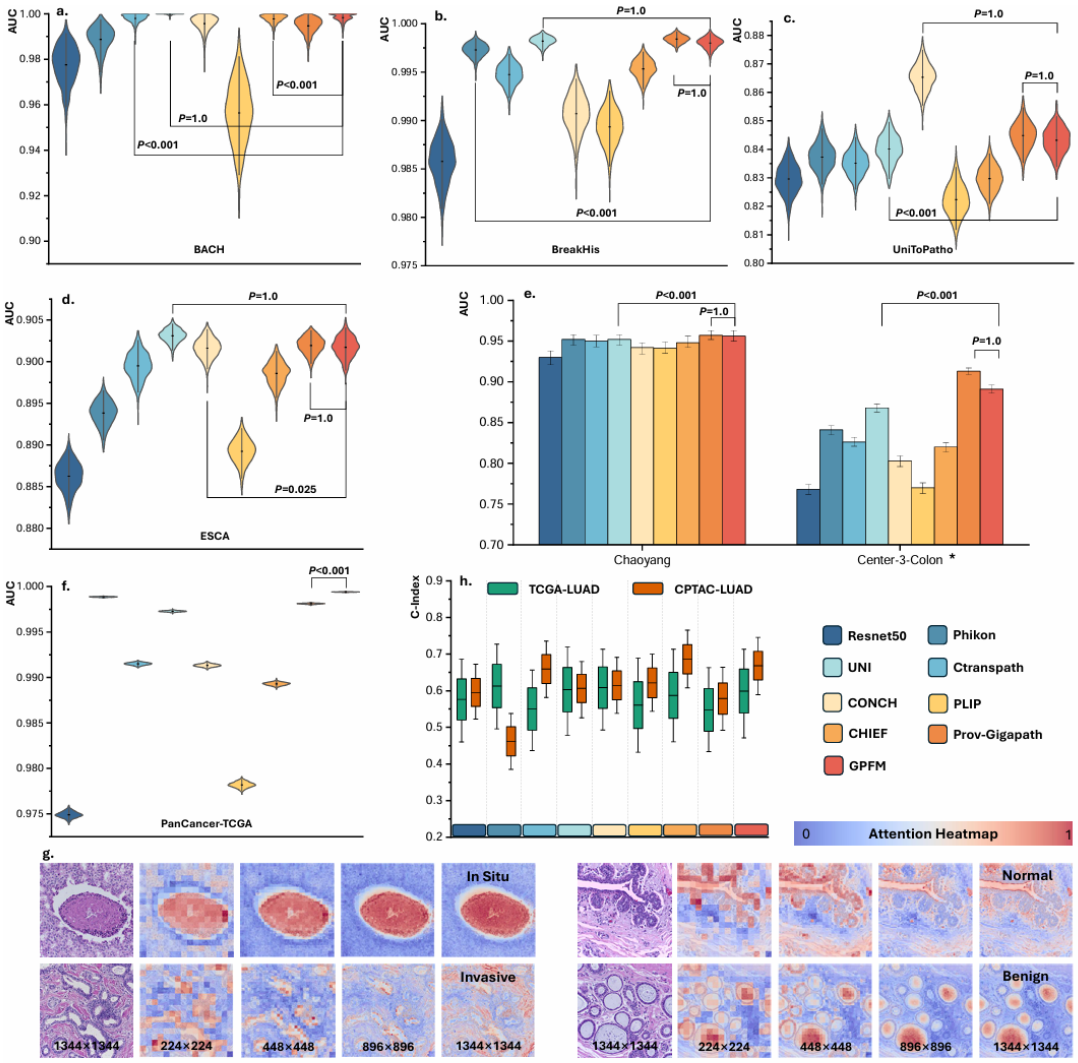


Fig. A3 Extended Result of ROI Classification Tasks. a-d. The AUC of foundation models on BACH, BreakHis, UniToPatho, and ESCA, respectively. **e.** The colon tissue classification performance. The Chaoyang and Center-3-Colon serve as internal and external, respectively. **f.** The performance of pancancer classificaiton of different foundation models. **g.**Attention heatmap of GPFM across various image resolutions for BRCA subtyping in BACH dataset. The colored squares represent the 14×14 [PATCH] tokens encoded by the GPFM model. The heatmap values indicate the similarity between each [PATCH] token and the [CLS] token generated by the last layer of GPFM, measured using Euclidean distance. The consistent attention patterns observed across varying image resolutions and tissue types underscore the robust capabilities of the GPFM model. **h.** Results on TCGA-LUAD data and the CPTAC-LUAD cohort. The survival prediction model was trained on the TCGA-LUAD cohort and subsequently tested on the CPTAC-LUAD cohort.

Additional comment 1: The figure legend of Figure 6d is incomplete.

Response:

Thank you for pointing out the problem of Figure 6d. We actually use different colors to represent different models for all subfigures. The legend is placed at the top of Figure 6, which may exist potential confusion. To avoid misleading, we adjusted the layout of Figure 6. The revised figure is shown below:

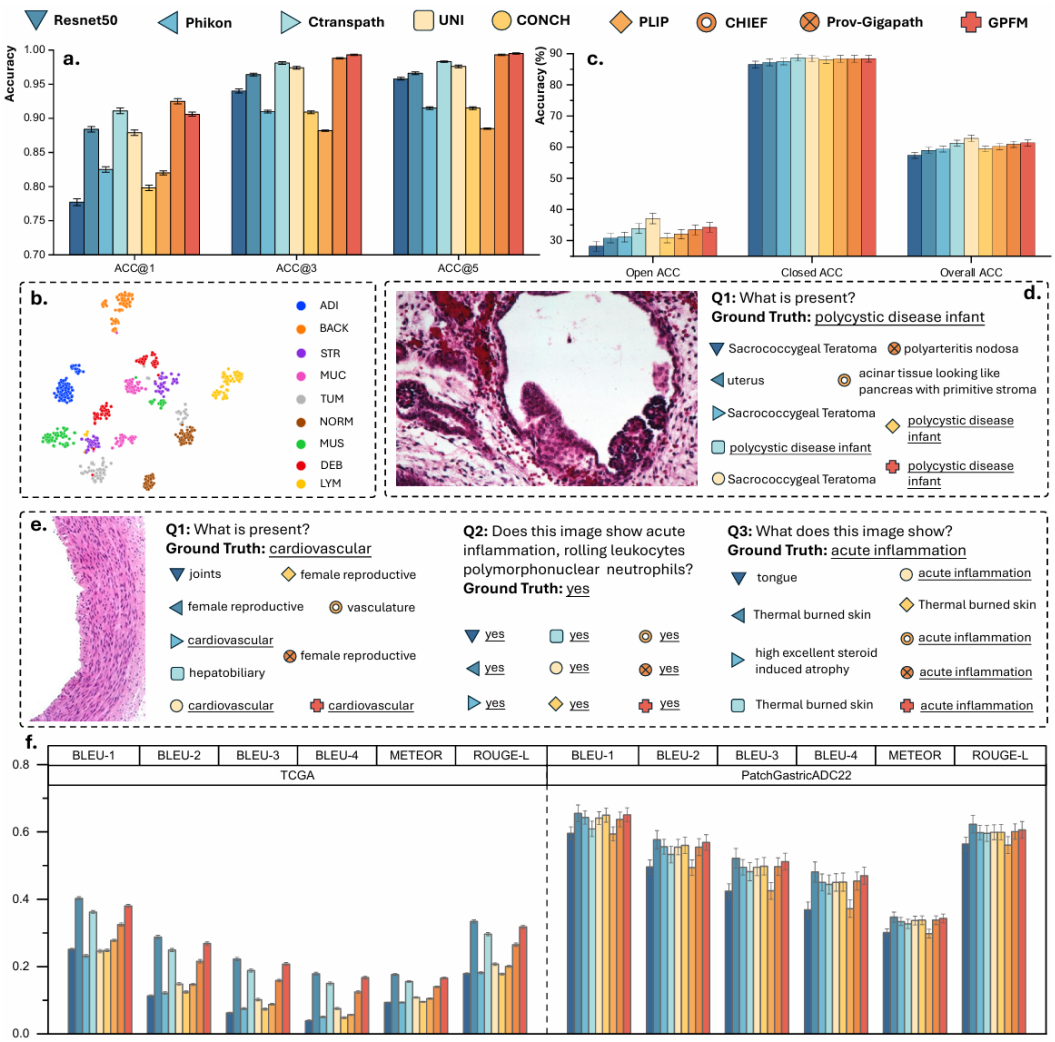


Fig. 6 Overview of Pathology Tissue Retrieval, VQA, and Report Generation. a. The top-1, top-3, top-5, and average accuracy of different FMs on pathology tissue retrieval tasks. **b.** The distribution of features extracted by GPFM. For each class, 100 samples from the test set were used, and a total of 900 samples were subjected to t-SNE dimensionality reduction to 2D. **c.** The performance of VQA on PathVQA dataset, measured by open-ended accuracy, closed-ended accuracy, and overall accuracy, for different FMs. **d.** An open-ended question along with the answers generated by various FMs. **e.** Three questions and the answers generated by FMs related to the query image. **f.** The performance of report generation on TCGA and PatchGastricAD22 data. The models are measured by six different metrics. In all subfigures, the error bars indicate standard deviation.

Code: The code provides a README file with sufficient instructions for installing and running the application.

Response:

We appreciate your acknowledgment of the sufficient instructions provided for installing and running the application. If there are any additional suggestions or areas for improvement, please let us know!