

Generating crossmodal gene expression from cancer histopathology improves multimodal AI predictions

Corresponding Author: Dr Tapabrata Chakraborty

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This paper introduces PathGen, which is a diffusion-based generative model that aims to synthesize transcriptomic data from histopathology WSIs to improve cancer diagnosis and prognosis prediction. Experiments on glioma and renal cancer datasets demonstrate that PathGen-synthesized transcriptomes have high correlation with real patient data (Spearman correlation ~ 0.71 - 0.72) and, when combined with histopathology features in downstream tasks, it can improve both cancer grading (AUC improvement from 0.823 to 0.890 for glioma) and survival risk estimation (C-Index improvement from 0.842 to 0.861) compared to using WSIs alone.

Below are my concerns:

Major concerns:

- Regarding the experiments, the study only uses two cancer types (namely glioma and renal cancer) from TCGA, where no external validation or independent cohorts outside of TCGA was used, which limits its evaluation on the generalization of the model.
- Although PathGen achieved a Spearman correlation values of 0.71-0.72, it needs careful interpretation rather than immediate acceptance as evidence of successful synthesis. Specifically, Spearman correlation basically aims to measure the monotonic relationships between variables rather than exact agreement. In these words, a correlation of 0.71 can implies that approximately 50% of the variance in real data remains unexplained by the synthetic data as the r^2 would be around 50%. Additionally, this correlation may obscure poor performance on clinically important but less abundant gene expressions. Hence, I would recommend looking into the correlation across gene categories. Also, please evaluate or discuss its performance on rare or extreme gene expression values, which be helpful.
- The authors claimed that the MAE=0.1 is low, but without context regarding the typical range and variability of gene expression values, it's difficult to understand how exactly MAE=0.1 performs. Some other options like normalized MAE can be an alternative.
- The authors claimed to provide detailed attention based heatmaps for both grading and survival risk prediction to improve the user trust (last paragraph in the intro), however no human evaluation is conducted to support this claim. The same applies to the conformal prediction parts, where no evaluations are conducted to show the effectiveness.
- For the gradation and survival prediction tasks, it lacks the comparison with other diffusion-based generative multimodal methods.
- The authors acknowledge that "for a few population groups across both cohorts, the required coverage has not been met, because of fewer samples in the calibration set." It reveals a potential flaw in the study's statistical framework, where conformal prediction's theoretical guarantees that prediction intervals contain the true value with a specified probability ($1-\alpha$) should rely on having sufficient calibration data that represents the full spectrum of the patient population. The authors' approach allocates only 104 WSIs from 82 cases for TCGA-GBMLGG and a mere 48 WSIs from 45 cases for TCGA-KIRC to the calibration set. This insufficient calibration data problem is exacerbated for demographic subgroups. To properly address this limitation, the study should alternative conformal prediction variants such as stratified conformal prediction, targeted calibration for minority groups.

Minor concerns:

- The abstract lacks quantitative descriptions (trained on how many WSIs) and results. It also lacks a high-level descriptions of their diffusion model as well as the difference of their diffusion models with existing diffusion models.
- Excessive use of subjective terminology like "significantly" throughout the introduction.

- No discussion of computational requirements or inference time for clinical deployment.
- As the experiments involve two datasets, I am wondering if stain normalization is involved in the preprocessing?

(Remarks on code availability)

Overall the codes look good and reasonable, with a clear readme file, and the dependent libraries.

Reviewer #2

(Remarks to the Author)

The paper presents PathGen, a commendable effort to address a significant challenge in deploying multimodal AI models in clinical practice: the frequent unavailability of transcriptomic data. The core idea of synthesizing transcriptomic data from routinely available histopathology images using a diffusion model is innovative for this specific application. The authors demonstrate that their approach can generate transcriptomes that, when used with WSI data in their MCAT-GR model, yield strong predictive performance for cancer grading and survival on TCGA datasets, often comparable to using real transcriptomic data and competitive with other state-of-the-art methods.

The inclusion of interpretability through attention maps and uncertainty quantification via conformal prediction are also valuable contributions towards building trust in AI models. While the study presents promising results, several critical aspects warrant further discussion and could temper some of the broader claims:

1. The study relies exclusively on TCGA datasets (GBMLGG and KIRC). While TCGA is a valuable public resource, models developed solely on it may not always generalize robustly to datasets from different institutions, with varying patient populations, image acquisition protocols (different scanners, staining variations), or pre-processing pipelines.
2. While applying diffusion models to synthesize transcriptomics from pathology is kind of novel, the core technical components themselves are largely based on existing, well-established works. The diffusion model framework is standard, UNI is a pre-existing foundation model for patch embedding, and MCAT is a prior architecture for multimodal fusion and prediction.
3. The authors introduce a dedicated "calibration set" for their conformal prediction analysis. While this is good practice for developing reliable conformal predictors, it inherently alters the data splits compared to studies that might use k-fold cross-validation or a simpler train/validation/test split without a separate calibration set. This means the training and testing sets available for PathGen's core predictive tasks might be smaller or distributed differently than in some of the SOTA methods they compare against. For a truly fair comparison with methods not employing such a calibration split, performance should ideally be evaluated using identical training and testing partitions as those used by the benchmarked studies, or the SOTA methods should be re-implemented/re-evaluated on PathGen's specific splits.
4. PathGen's approach is to explicitly generate transcriptomic data. An alternative class of methods focuses on aligning features from different modalities (e.g., learning a joint embedding space or using one modality to guide feature extraction in another) to improve prediction, without necessarily generating one modality from the other. For example, the paper cites PathoGen-X, which aims to "trans-align" genomic and WSI features. The paper doesn't deeply discuss why the explicit generation approach was chosen over these alignment strategies, which might be more direct if the ultimate goal is prediction and could potentially avoid the compounding of errors inherent in a generative step. While PathGen's generated data has the advantage of being interpretable as a "transcriptome," alignment methods can also offer interpretability through cross-modal attention or feature correlation in the learned joint space.

(Remarks on code availability)

The provided code repository appears well-structured, and the accompanying instructions are clear. However, the absence of pre-trained model checkpoints is a notable limitation, as this prevents direct evaluation of the trained models and hinders efforts to precisely replicate the reported results without undertaking full model retraining.

Reviewer #3

(Remarks to the Author)

In this manuscript, Dey et al. present PathGen, a deep-learning framework that integrates histology images with synthetic RNA-seq data to predict tumor grade and cancer patient survival. Using publicly available glioma (GBM/LGG) and kidney cancer (KIRC) datasets, they demonstrate the advantage of multimodal integration over single-modality approaches.

The strengths of this study include: (1) Development of a novel deep-learning model for predicting transcriptomic features from H&E-stained images. (2) Patch-level interpretability to capture intra-tumoral heterogeneity. (3) Use of conformal predictions to address model uncertainty.

Overall, the study tests an innovative hypothesis with instructive findings for research in the field. My primary concerns are technical and pertain to data selection, preprocessing, and presentation.

Datasets:

1. The rationale for restricting analysis to GBM/LGG and KIRC from the broader TCGA cohort (spanning over 30 cancer types) is not clearly explained.
2. Predictive performance for both tumor grade and survival was approximately 15% higher in GBM/LGG compared to KIRC,

indicating inconsistency for the predictive performance across different cancer types. Assessing the predictive performance in additional cancer types would help determine model generalizability across different cancers. While the authors claim this as a future goal in the Discussion, it would be beneficial to demonstrate the robustness and consistency of the model in studies for Nature Communications.

Data Preprocessing:

1. The manuscript does not specify which gene length normalization method (TPM, FPKM, or RSEM) was used prior to computing z-scores.
2. Figure 2(i): More technical details are needed on how predictive performance for individual genes was aggregated at the pathway/gene-group level.

Results and Data Presentation:

1. Section 2.2: The statement that "gene expression levels for gene groups show no significant differences between real and synthesized transcriptomes" ($p > 0.05$, unpaired t-test) was inconsistent with the results shown in Supplementary Table 2, where all p-values are highly significant. This discrepancy should be clarified.
2. Given the substantial survival differences between GBM and LGG, evaluating model performance on the combined TCGA-GBMLGG cohort is not surprising. The authors should report separate predictive performance (grade and C-index) for GBM and LGG individually, at least in Supplementary materials.
3. Minor issues on Figure 1c which contains a redundant text box below "Patch embeddings."
4. In Section 2.4.1, for a patient with ground truth grade I (Figure 4), the majority of patches within the tumor slide are predicted as grade IV (red). Is this expected?
5. Figures 3 & Supplementary Figure 4: Do lower-grade tumor regions consistently exhibit higher oncogene expression levels? In Supplementary Figure 4, the middle tumor portion shows lower risk scores but higher oncogene expression than the tumor edges—is this expected?
6. Supplementary Figure 4(d): Minor issues on the image annotation on the left side which is duplicated.

(Remarks on code availability)

The authors have provided pertinent code with user instructions for reproducing their analysis.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Most of the concerns have been addressed the revision. The authors mentioned that for the clinically important but less abundant gene expressions, this analysis shall involve much substantial domain expertise and input, which is totally understandable, please kindly consider put this into the discussion.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

Thank you for the revised submission. The revisions have successfully addressed my previous concerns.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The authors have carefully addressed all of this reviewer's comments.

(Remarks on code availability)

The authors have provided pertinent code with user instructions for reproducing their analysis.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

AUTHOR RESPONSE DOCUMENT

Reviewer #1 (Remarks to the Author):

This paper introduces PathGen, which is a diffusion-based generative model that aims to synthesize transcriptomic data from histopathology WSIs to improve cancer diagnosis and prognosis prediction. Experiments on glioma and renal cancer datasets demonstrate that PathGen-synthesized transcriptomes have high correlation with real patient data (Spearman correlation ~ 0.71 - 0.72) and, when combined with histopathology features in downstream tasks, it can improve both cancer grading (AUC improvement from 0.823 to 0.890 for glioma) and survival risk estimation (C-Index improvement from 0.842 to 0.861) compared to using WSIs alone.

Below are my concerns:

Major concerns:

- Regarding the experiments, the study only uses two cancer types (namely glioma and renal cancer) from TCGA, where no external validation or independent cohorts outside of TCGA was used, which limits its evaluation on the generalization of the model.

Author Response: We thank the reviewer for this valuable suggestion. In response, we have extended our experiments on two new cohorts from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) dataset and two additional cohorts from the previous TCGA dataset. Thus, in total, results for four additional data cohorts have now been added to the paper, namely, TCGA-UCEC (uterine), TCGA-BRCA (breast), CPTAC-GBM (glioblastoma) and CPTAC-UCEC (uterine). For external validation, models trained on TCGA data were directly evaluated on CPTAC cohorts without re-training. For all the considered cohorts from both TCGA and CPTAC datasets, we find that the addition of synthesized transcriptomic data significantly improves the performance of both gradation and survival risk estimation over using WSIs only (p -value < 0.05). Also, the performance of using transcriptomic data synthesized by our model, PathGen, is significantly close to that of using their real counterparts for all data cohorts (p -value < 0.05). The consistent results show that our model generalizes well over cancer cohorts and datasets. The results have been discussed in the Sections “*Similarity evaluation of synthesized and real transcriptomic data*” and “*Performance evaluation of synthesized transcriptomic data in multimodal prediction*”, and presented in Figure 2.b. and Supplementary Table 6. We have also modified Sections “*Transparency evaluation for clinical interpretability*” and “*Uncertainty quantification for clinical reliability*” to incorporate transparency evaluations and uncertainty quantification with respect to the newly added cohorts.

- Although PathGen achieved a Spearman correlation value of 0.71-0.72, it needs careful interpretation rather than immediate acceptance as evidence of successful synthesis. Specifically, Spearman correlation basically aims to measure the monotonic relationships between variables rather than exact agreement. In these words, a correlation of 0.71 can imply that approximately 50% of the variance in real data remains unexplained by the synthetic data as the r^2 would be around 50%. Additionally, this correlation may obscure poor performance on clinically important but less abundant gene expressions. Hence, I would recommend looking into the correlation across gene categories. Also, please evaluate or discuss its performance on rare or extreme gene expression values, which will be helpful. The authors claimed that the MAE=0.1 is low, but without context regarding the typical range and variability of gene expression values, it's difficult to understand how exactly MAE=0.1 performs. Some other options like normalized MAE can be an alternative.

Author Response: We thank the reviewer for the suggestion. We have now incorporated the results for the new datasets and cohorts, and presented the updated results together

(previous and new) in Figure 2.a. and Supplementary Table 5. We observe a significantly high correlation across gene categories for all data cohorts. Also, following the reviewers' suggestion, we have now evaluated the difference between real and synthesized transcriptomic features using normalized MAE (nMAE) instead of MAE.

Regarding the reviewer's concern about clinically important but less abundant gene expressions, we acknowledge that this is an important direction. In this context, we consulted with our clinician co-author (C. Banerji), who emphasized that *"the aggregation of the transcriptome in this way reduces the number of extreme values we observe compared to estimating every gene individually, and our validation set is therefore less subject to experimental noise driving extreme rare values due to the averaging process."* However, such an analysis would require substantial domain expertise and dedicated clinical input, which we plan to pursue as part of future work.

- The authors claimed to provide detailed attention based heatmaps for both grading and survival risk prediction to improve the user trust (last paragraph in the intro), however no human evaluation is conducted to support this claim. The same applies to the conformal prediction parts, where no evaluations are conducted to show the effectiveness.

Author Response: We thank the reviewer for this comment. In the introductory paragraph in question, we stated: 'For practical utility, our model must be trusted by clinicians and attune to emerging regulation'. This statement does not constitute a claim that our approaches objectively increase user trust; rather, this statement is a motivation for their introduction.

Our approaches were selected with input from a practising histopathologist (co-author C Banerji) based at University College Hospitals to reflect methods that would be most valuable and trusted in real clinical workflows. He also went through all the results presented in the paper (including visual ones) as well as the written explanations in the manuscript. Thus, though a full clinical validation would be out of scope of the current proof of concept methodological piece, the direct clinician in the loop development and feedback through C Banerji, does provide a level of clinician trust for the time being.

Our focus is instead on the technical development of these methods in the context of grading and survival prediction, with design choices guided by emerging regulatory requirements (e.g., EU AI Act Articles 13–15). These requirements: transparency, uncertainty quantification and human oversight were in fact proposed by our clinician co-author C Banerji in his *Nature Medicine*, 2023 paper as factors likely to facilitate clinician trust: <https://www.nature.com/articles/s41591-023-02562-7>. Moreover, in response to the reviewer's comment we have now added the following line in the *Discussion* section: "Thus, to add explainability to the model predictions, we have undertaken measures with design choices guided by emerging regulatory requirements (e.g., EU AI Act Articles 13–15) and in consultation with a practising histopathologist based at University College Hospitals, London."

- For the gradation and survival prediction tasks, it lacks the comparison with other diffusion-based generative multimodal methods.

Author Response: We thank the reviewer for this observation. In our work, the diffusion-based framework, PathGen, has been proposed to synthesize transcriptomic data from WSIs. However, the downstream gradation and survival prediction tasks are non-generative tasks and performed using a transformer-based multiple instance learning (MIL) model, MCAT_GR, inspired by established work (MCAT). Since these predictive tasks are not carried out by the diffusion model itself, but by a transformer-based MIL model, we

have compared our performance with other multimodal transformer-based or MIL approaches, which we believe provides the most meaningful and fair comparison.

- The authors acknowledge that "for a few population groups across both cohorts, the required coverage has not been met, because of fewer samples in the calibration set." It reveals a potential flaw in the study's statistical framework, where conformal prediction's theoretical guarantees that prediction intervals contain the true value with a specified probability $(1-\alpha)$ should rely on having sufficient calibration data that represents the full spectrum of the patient population. The authors' approach allocates only 104 WSIs from 82 cases for TCGA-GBMLGG and a mere 48 WSIs from 45 cases for TCGA-KIRC to the calibration set. This insufficient calibration data problem is exacerbated for demographic subgroups. To properly address this limitation, the study should alternative conformal prediction variants such as stratified conformal prediction, targeted calibration for minority groups.

Author Response: We thank the reviewer for highlighting this concern. We agree that conformal prediction's theoretical guarantees depend on calibration set size, and that smaller calibration groups may naturally exhibit larger deviations from the nominal coverage. However, this reflects the inherent statistical properties of finite-sample calibration. For calibration sets of size 1000, the coverage for $\alpha = 0.1$ may typically fluctuate to 0.88, rather than being exactly 0.90. (<https://arxiv.org/pdf/2107.07511>) The deviation depends on calibration set size, and the smaller the calibration set, the larger the permissible variation. In our revised manuscript, we have thus computed the theoretical coverage slack for all calibration set sizes used in our study (Supplementary Table 14) and verified that the observed coverage is within the permissible deviation. This ensures the validity of our conformal strategy and provides statistically sound calibration for both grade and survival risk predictions under marginal as well as conditional settings.

Furthermore, we perform stratified conditional conformal prediction and in response to the reviewer's suggestion, now report not only marginal coverage and uncertainty (as in the original submission) but also conditional coverage and uncertainty stratified across demographic subgroups (e.g., age, gender, and vital status). This provides a more comprehensive evaluation of subgroup performance. We note that both coverages satisfied the theoretical guarantees within the expected deviation, demonstrating that our conformal calibration approach remains robust across cohorts and specific subpopulations. We have accordingly updated Section "*Uncertainty quantification for clinical reliability*" and reported the results in the Supplementary Tables 9-13 and Figure 6. The methodology has been discussed further in *Methods* section.

Minor concerns:

- The abstract lacks quantitative descriptions (trained on how many WSIs) and results. It also lacks a high-level descriptions of their diffusion model as well as the difference of their diffusion models with existing diffusion models.

Author Response: We thank the reviewer for this constructive observation. In response, we have revised the abstract to incorporate dataset details and results of our experiments. Specifically, we added the following lines: "We experiment on two publicly available multimodal datasets, The Cancer Genomic Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC), spanning four independent cohorts: glioma-glioblastoma, renal, uterine, and breast, and observe significant performance gains in gradation and risk estimation ($p\text{-value} < 0.05$) when incorporating synthesized transcriptomic data with WSIs. Also, predictions using synthesized features were statistically close to those obtained with real transcriptomic data ($p\text{-value} > 0.05$), consistently across cohorts."

We agree with the reviewer that adding details such as the number of WSIs, quantitative results, and diffusion model description would provide further clarity. However, Nature Communications requires the abstract to remain under 150 words ([formatting instructions](#)), which restricts us from including additional details. However, we would like to note that, for TCGA-GBMLGG, 912 whole slide images (WSIs) coming from 745 cases are used, of which 183 cases belong to TCGA-GBM and 562 cases belong to TCGA-LGG. We consider 485 WSIs from 462 cases for TCGA-KIRC, 294 WSIs from 267 cases for TCGA-UCEC, 1010 WSIs from 946 cases for TCGA-BRCA, 242 WSIs from 62 cases for CPTAC-GBM, and 364 WSIs from 71 cases for CPTAC-UCEC for our experiments. Further, we have provided all quantitative dataset details, WSI counts, results, model description and model comparisons in the *Results* and *Methods* sections, ensuring transparency and accessibility for readers.

- Excessive use of subjective terminology like "significantly" throughout the introduction.

Author Response: We thank the reviewer for this comment. The word 'significantly' is used 3 times in the introduction, twice it is used in a non-subjective manner to refer objective statistical significance: first: '...significant improvements in the predictive accuracy...' and second: '...the predictive performance of both diagnosis (cancer grading) and prognosis (survival risk) significantly improves...'. We acknowledge that the word is used in a subjective context, not relevant to statistical significance, on the third mention: 'convey significant health benefits' and so have changed the word in the sentence to 'substantial'. We have now edited the manuscript to ensure significance is only used in the context of statistical assessment, and not subjectively.

- No discussion of computational requirements or inference time for clinical deployment.

Author Response: We thank the reviewer for this suggestion. We have now included the inference time in the *Discussion* section in the following line: "Also, we note the average inference time for cross-modal synthesis of transcriptomic data from WSIs using PathGen is 21.15 seconds, and the average time for estimating survival and grade is 0.71 seconds, which varies over the magnification level and the size of the WSI, as expected."

- As the experiments involve two datasets, I am wondering if stain normalization is involved in the preprocessing?

Author Response: We thank the reviewer for the question. We would like to clarify that stain normalization was not applied during the preprocessing of the datasets, being influenced by recent works highlighting that stain normalization has no significant effect on the predictive performance (<https://www.nature.com/articles/s41598-023-46619-6>). In response to the reviewer's comment, we have now included the following line in the processing step discussed in *Methods* section - "*We do not use stain normalization in preprocessing the WSIs in line with recent works that find no significant effect of using stain normalization as a preprocessing step.*"

Reviewer #1 (Remarks on code availability):

Overall the codes look good and reasonable, with a clear readme file, and the dependent libraries.

Author Response: We sincerely thank the reviewer for the positive feedback on the quality of our code, documentation, and dependencies.

Reviewer #2 (Remarks to the Author):

The paper presents PathGen, a commendable effort to address a significant challenge in deploying multimodal AI models in clinical practice: the frequent unavailability of transcriptomic data. The core idea of synthesizing transcriptomic data from routinely available histopathology images using a diffusion model is innovative for this specific application. The authors demonstrate that their approach can generate transcriptomes that, when used with WSI data in their MCAT-GR model, yield strong predictive performance for cancer grading and survival on TCGA datasets, often comparable to using real transcriptomic data and competitive with other state-of-the-art methods.

The inclusion of interpretability through attention maps and uncertainty quantification via conformal prediction are also valuable contributions towards building trust in AI models. While the study presents promising results, several critical aspects warrant further discussion and could temper some of the broader claims:

1. The study relies exclusively on TCGA datasets (GBMLGG and KIRC). While TCGA is a valuable public resource, models developed solely on it may not always generalize robustly to datasets from different institutions, with varying patient populations, image acquisition protocols (different scanners, staining variations), or pre-processing pipelines.

Author Response: We thank the reviewer for this valuable suggestion. In response, we extended our experiments by incorporating additional, independent cohorts beyond TCGA. Specifically, we included data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). In total, results for four more data cohorts have now been added to the paper, namely, TCGA-UCEC (uterine), TCGA-BRCA (breast), CPTAC-GBM (glioblastoma) and CPTAC-UCEC (uterine), along with existing TCGA-GBMLGG (glioma-glioblastoma) and TCGA-KIRC (renal). These cohorts encompass data collected from distinct institutions at distinct geographical locations, having varied acquisition protocols. For instance, the CPTAC-GBM cohort was supplemented with retrospective samples from CHOP (Children's Hospital of Philadelphia)-Upenn, Philadelphia ([CPTAC-GBM](#)), whereas TCGA-GBM contains data from Henry Ford Hospital, Detroit; University of California, San Francisco; MD Anderson Cancer Center, Houston; Fondazione IRCCS Istituto Neurologico C. Besta, Milan and various other hospitals ([TCGA-GBM](#)). The patient demographics also vary across each cohort and dataset (Supplementary Table 3).

Our findings demonstrate that models trained on TCGA datasets generalize effectively to CPTAC cohorts. Moreover, the improvement in performance on the addition of synthesized transcriptomic data over WSI, and closeness of grade and survival estimation for real and synthesized transcriptomic data, is significant and consistent over diverse cancer types and data from diverse institutional settings. The detailed results have been discussed in Sections "*Similarity evaluation of synthesized and real transcriptomic data*" and "*Performance evaluation of synthesized transcriptomic data in multimodal prediction*", and presented in Figure 2.b. and Supplementary Table 6, and corresponding explanations have been incorporated in Sections "*Transparency evaluation for clinical interpretability*" and "*Uncertainty quantification for clinical reliability*", including a discussion of the fairness of prediction over the patient demographics.

2. While applying diffusion models to synthesize transcriptomics from pathology is kind of novel, the core technical components themselves are largely based on existing, well-established works. The diffusion model framework is standard, UNI is a pre-existing foundation model for patch embedding, and MCAT is a prior architecture for multimodal fusion and prediction.

Author Response: We thank the reviewer for this thoughtful observation. We agree that the core components of our framework are built upon established methods. The diffusion model framework was indeed originally developed for image generation ([Ho et al.](#)) and has since been successfully extended to NLP and multimodal applications ([Zhu et al.](#)). However, to the best of our knowledge, our work is the first to adapt this framework for the cross-modal task of synthesizing transcriptomic data from WSIs. It is also common practice and an area of study in computational pathology to employ pretrained foundation models for obtaining robust WSI embeddings ([Yellapragada et al.](#)). In line with this, we utilize UNI as a well-established backbone for patch embedding. While MCAT has been proposed previously for survival prediction, in our work, we extend this architecture to simultaneously address both grade prediction and survival risk estimation, incorporating an ensemble loss for joint training. These modifications are described in detail in *Methods* (Section “*Multimodal cancer gradation and survival risk estimation*”).

We would like to humbly suggest that in a high stakes translational scenario such as a computational cancer it might be more prudent sometimes to employ a tried and tested approach in a novel manner (as in this work) rather than a new but unestablished method which might have less potential of being trusted and used in the clinic, which would defeat the purpose altogether.

3. The authors introduce a dedicated "calibration set" for their conformal prediction analysis. While this is good practice for developing reliable conformal predictors, it inherently alters the data splits compared to studies that might use k-fold cross-validation or a simpler train/validation/test split without a separate calibration set. This means the training and testing sets available for PathGen's core predictive tasks might be smaller or distributed differently than in some of the SOTA methods they compare against. For a truly fair comparison with methods not employing such a calibration split, performance should ideally be evaluated using identical training and testing partitions as those used by the benchmarked studies, or the SOTA methods should be re-implemented/re-evaluated on PathGen's specific splits.

Author Response: We thank the reviewer for the suggestion. To address this, we performed k-fold cross-validation on the newly added TCGA cohorts, ensuring a fairer comparison with the SOTA methods. On TCGA-UCEC, our methodology achieved a C-index of 0.723, whereas DeepSets, DeepAttnMISL, and MCAT achieved 0.593, 0.586, and 0.622, respectively. On TCGA-BRCA, our methodology achieved a C-index of 0.645, compared to 0.522, 0.577, and 0.580 for DeepSets, DeepAttnMISL, and MCAT, respectively. These results demonstrate that even under comparable partitioning strategies, our method consistently outperforms existing approaches. We note that the considered SOTA models do not perform gradation, so a direct comparison on grade prediction was not possible. Finally, due to computational constraints, we were unable to repeat experiments on the earlier TCGA-GBMLGG or TCGA-KIRC cohorts under identical splits, as this would have required considerable additional resources and time, well beyond the scope of this review period. Nevertheless, the consistent superiority of our approach on the newly evaluated cohorts provides strong evidence of its robustness and fairness in comparison.

4. PathGen's approach is to explicitly generate transcriptomic data. An alternative class of methods focuses on aligning features from different modalities (e.g., learning a joint embedding space or using one modality to guide feature extraction in another) to improve prediction, without necessarily generating one modality from the other. For example, the paper cites PathoGen-X, which aims to "trans-align" genomic and WSI features. The paper doesn't deeply discuss why the explicit generation approach was chosen over these

alignment strategies, which might be more direct if the ultimate goal is prediction and could potentially avoid the compounding of errors inherent in a generative step. While PathGen's generated data has the advantage of being interpretable as a "transcriptome," alignment methods can also offer interpretability through cross-modal attention or feature correlation in the learned joint space.

Author Response: We thank the reviewer for raising this important point. We would like to note that approaches that rely on one modality to guide feature extraction from another (such as cross-modal attention) are similar to what our predictive model MCAT_GR does. However, such approaches require the availability of both modalities at inference time. This is often impractical, as transcriptomic data is not always available in clinical settings. Our rationale for explicit cross-modal synthesis is precisely to address this limitation — by generating the missing transcriptomic modality from WSIs, we enable downstream multimodal prediction at inference while still benefiting from the strengths of both modalities. Also, while alignment-based methods that learn a shared embedding space across modalities are indeed a viable strategy, they inherently make one modality dominant in the joint representation.

Thus, explicit transcriptomic synthesis not only enables interpretability by producing an interpretable "transcriptome," but also provides a practical and effective solution for real-world clinical inference where transcriptomic data is absent. The synthesized transcriptomic data may be used to evaluate the add-on benefits of performing a transcriptomic test.

Reviewer #2 (Remarks on code availability):

The provided code repository appears well-structured, and the accompanying instructions are clear. However, the absence of pre-trained model checkpoints is a notable limitation, as this prevents direct evaluation of the trained models and hinders efforts to precisely replicate the reported results without undertaking full model retraining.

Author Response: We thank the reviewer for the positive feedback regarding the structure and clarity of our code repository. We acknowledge the concern regarding the absence of pre-trained model checkpoints. We have now added downloadable links for the checkpoints of all our models in the [GitHub repository](#).

Reviewer #3 (Remarks to the Author):

In this manuscript, Dey et al. present PathGen, a deep-learning framework that integrates histology images with synthetic RNA-seq data to predict tumor grade and cancer patient survival. Using publicly available glioma (GBM/LGG) and kidney cancer (KIRC) datasets, they demonstrate the advantage of multimodal integration over single-modality approaches. The strengths of this study include: (1) Development of a novel deep-learning model for predicting transcriptomic features from H&E-stained images. (2) Patch-level interpretability to capture intra-tumoral heterogeneity. (3) Use of conformal predictions to address model uncertainty.

Overall, the study tests an innovative hypothesis with instructive findings for research in the field. My primary concerns are technical and pertain to data selection, preprocessing, and presentation.

Datasets:

1. The rationale for restricting analysis to GBM/LGG and KIRC from the broader TCGA cohort (spanning over 30 cancer types) is not clearly explained.

Author Response: We thank the reviewer for this helpful comment. In the initial version of our study, we focused on TCGA-GBMLGG and TCGA-KIRC because TCGA does not provide tumour grade information for all cancer cohorts, and these two cohorts included both grade and survival annotations. This was important since our model, MCAT_GR, is specifically designed to jointly predict grade and survival risk. However, we have now extended our experiments to two additional TCGA cohorts, TCGA-UCEC and TCGA-BRCA.

2. Predictive performance for both tumor grade and survival was approximately 15% higher in GBM/LGG compared to KIRC, indicating inconsistency for the predictive performance across different cancer types. Assessing the predictive performance in additional cancer types would help determine model generalizability across different cancers. While the authors claim this as a future goal in the Discussion, it would be beneficial to demonstrate the robustness and consistency of the model in studies for Nature Communications.

Author Response: We thank the reviewer for this valuable suggestion. In response, we have extended our experiments to include data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) as well as two additional TCGA cohorts, uterine and breast. In total, four additional cohorts have now been incorporated: TCGA-UCEC (uterine), TCGA-BRCA (breast), CPTAC-GBM (glioblastoma), and CPTAC-UCEC (uterine). We would also like to note that variations in absolute predictive performance across different cancer types are expected and have been observed in prior established works (MCAT, PathomicFusion). Our focus, therefore, is on the generalizability of the claim that synthesized transcriptomic data provides significant added value over WSI-only predictions, and that predictions based on real and synthesized transcriptomic data remain significantly close. The new results and analyses are presented in Figure 2b and Supplementary Table 6, and discussed in Sections “*Similarity evaluation of synthesized and real transcriptomic data*” and “*Performance evaluation of synthesized transcriptomic data in multimodal prediction*”. We have also updated Sections “*Transparency evaluation for clinical interpretability*” and “*Uncertainty quantification for clinical reliability*” to incorporate transparency evaluations and uncertainty quantification with respect to the newly added cohorts.

Data Preprocessing:

1. The manuscript does not specify which gene length normalization method (TPM, FPKM, or RSEM) was used prior to computing z-scores.

Author Response: We thank the reviewer for pointing this out. The transcriptomic data for the TCGA and CPTAC datasets we used were obtained from cBioPortal (<https://www.cbioportal.org>), which already provides gene length-normalized expression values. Accordingly, in our preprocessing, we directly applied z-score normalization in cases where the values were not already standardized. For the new experiments performed on the additional datasets and cohorts (TCGA-UCEC, TCGA-BRCA, CPTAC-GBM, CPTAC-UCEC), the transcriptomic profiles were downloaded from cBioPortal, where the data are processed using RSEM. However, for the earlier experiments, we relied on the transcriptomic data provided by the authors of MCAT via their public [GitHub repository](#). The original paper does not explicitly specify the normalization method, but according to cBioPortal, “*RNASeqV2 from TCGA is processed and normalized using RSEM; specifically, the RNASeqV2 data in cBioPortal corresponds to the rsem.genes.normalized_results file from TCGA.*” Hence, we infer that RSEM normalization was also applied in those datasets.

2. Figure 2(i): More technical details are needed on how predictive performance for individual genes was aggregated at the pathway/gene-group level.

Author Response: We thank the reviewer for this comment. We would like to clarify that Figure 2(a) presents the comparison between real and synthesized transcriptomic data

across all gene groups as well as for individual gene groups, using normalized MAE and correlation metrics. It does not illustrate technical details of the aggregation of individual genes into pathway/gene-group levels. The methodology describing the pathway is discussed in brief in *Methods* (Section “*Multimodal cancer gradation and survival risk estimation*”) and illustrated in Figure 7(b, d). We learn an 1 X 1024 embedding for all genes in a gene group using the gene encoder (Figure 7.d) comprising 4 linear layers, each followed by an ELU activation. These embeddings corresponding to the gene groups are used to coattend to WSI features extracted using UNI. A dedicated pathway is used for survival risk prediction using the transcriptomic features followed by a global attention pooling which is then combined with the globally pooled feature map obtained from the co-attended feature pipeline to predict the survival risk (Figure 7(b)). Further, the co-attention maps are visualized in Figures 3,4 and Supplementary Figures 3,4,5 and explained in Section “Transparency evaluation for clinical interpretability”.

Results and Data Presentation:

1. Section 2.2: The statement that “gene expression levels for gene groups show no significant differences between real and synthesized transcriptomes” ($p > 0.05$, unpaired t-test) was inconsistent with the results shown in Supplementary Table 2, where all p-values are highly significant. This discrepancy should be clarified.

Author Response: To clarify, the p-values in Supplementary Table 2 (now Supplementary Table 5) correspond to Spearman correlation results, which indicate highly significant correlations between the real and synthesized transcriptomic features. In the earlier version, we had also performed unpaired t-tests on the existing TCGA-GBMLGG and TCGA-KIRC cohorts, but their p-values were not included in the table. We thank the reviewer for carefully pointing this out.

However, while extending our experiments to additional cohorts, we found the unpaired t-test to be redundant, since the correlation analysis already demonstrated highly significant associations (p -values < 0.05). Therefore, in the revised manuscript, we have removed the reference to the unpaired t-test and now state instead: “*Thus, showing highly significant correlation between real and synthesized transcriptomic data (Spearman correlation p -value < 0.05)*”, in Section “*Similarity evaluation of synthesized and real transcriptomic data*”. We have also removed the explicit p-values from Supplementary Table 5, since all were displayed as 0.000, and clarified in the table caption that all Spearman correlation p-values are < 0.05 .

2. Given the substantial survival differences between GBM and LGG, evaluating model performance on the combined TCGA-GBMLGG cohort is not surprising. The authors should report separate predictive performance (grade and C-index) for GBM and LGG individually, at least in Supplementary materials.

Author Response: We thank the reviewer for this suggestion. In response, we have now reported the predictive performance (grade and C-index) separately for TCGA GBM and LGG cohorts in Supplementary Table 7 of the revised manuscript.

3. Minor issues on Figure 1c which contains a redundant text box below “Patch embeddings.”

Author Response: We thank the reviewer for pointing this out. The redundant text box in Figure 1 (previously 1c) has been removed in the revised manuscript.

4. In Section 2.4.1, for a patient with ground truth grade I (Figure 4), the majority of patches within the tumor slide are predicted as grade IV (red). Is this expected?

Author Response: We thank the reviewer for flagging this. There was indeed an error in referring to the figure, Figure 4 actually corresponds to a ground truth grade IV case, and not grade I. We have corrected this in the revised manuscript. To avoid any such ambiguity in

the future, we now provide the patient's age, gender, ground truth grade, survival time, and survival status in the first line of the figure captions.

5. Figures 3 & Supplementary Figure 4: Do lower-grade tumor regions consistently exhibit higher oncogene expression levels? In Supplementary Figure 4, the middle tumor portion shows lower risk scores but higher oncogene expression than the tumor edges—is this expected?

Author Response: We thank the reviewer for this thoughtful question. The association between oncogene expression and tumour grade is complex and findings are inconsistent. Some studies have found oncogene expression to be positively correlated with tumour grade (<https://pmc.ncbi.nlm.nih.gov/articles/PMC1977917/>), while other studies report a negative correlation (<https://www.sciencedirect.com/science/article/pii/S0306987710004263>), or no correlation at all (<https://www.sciencedirect.com/science/article/pii/S0031302516352692>). Because of this it is hard to comment on whether the data we present in this regard is expected or not.

We would like to clarify that Figures 3, 4 and Supplementary Figures 3, 4 do not represent actual gene expression levels. Instead, they visualize the co-attention maps obtained from our predictive model, which we use as a tool for explainability. In these maps, higher values indicate stronger model attention rather than higher gene expression. Given this caveat and the biological complexity, we are cautious about drawing any definitive biological conclusions about the association between oncogene expression and tumour grade from our data.

However, our work provides a step towards computational methodologies for investigating the complex relationship between tumour grade and gene expression patterns in future studies, via the positing of testable hypotheses for wet lab validation. It provides an additional visual presentation of tumour heterogeneity to the clinician expert (pathologist) to choose whether or not to draw any further intuition from them. Thus, given the heterogeneity of cancer, the AI system deliberately merely presents these maps to the human user instead of making any generalised claims from it.

6. Supplementary Figure 4(d): Minor issues on the image annotation on the left side which is duplicated.

Author Response: We thank the reviewer for noticing this. The duplicated annotation in Supplementary Figure 4(d) has been removed in the revised manuscript.

Reviewer #3 (Remarks on code availability):

The authors have provided pertinent code with user instructions for reproducing their analysis.

Author Response: We thank the reviewer for acknowledging the availability and clarity of our code and user instructions. We have made every effort to ensure reproducibility and ease of use for the research community.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

Most of the concerns have been addressed the revision. The authors mentioned that for the clinically important but less abundant gene expressions, this analysis shall involve much substantial domain expertise and input, which is totally understandable, please kindly consider put this into the discussion.

Author Response: We thank the reviewer for the thoughtful suggestion. As advised, we have now accordingly revised the Discussion section to include this clarification: “Also, evaluating the clinically important but less abundant synthesized gene expressions would require substantial domain expertise.”

Reviewer #2 (Remarks to the Author):

Thank you for the revised submission. The revisions have successfully addressed my previous concerns.

Author Response: We sincerely thank the reviewer for the positive feedback and appreciation of our revisions.

Reviewer #3 (Remarks to the Author):

The authors have carefully addressed all of this reviewer's comments.

Author Response: We thank the reviewer for their time, effort, and positive assessment of our revised manuscript.

Reviewer #3 (Remarks on code availability):

The authors have provided pertinent code with user instructions for reproducing their Analysis.

Author Response: We thank the reviewer for acknowledging the reproducibility of our work and the clarity of the provided code and instructions.