

Multimodal Fusion of Radio-Pathology and Proteogenomics Identify Integrated Glioma Subtypes with Prognostic and Therapeutic Opportunities

Corresponding Author: Dr Zhenyu Zhang

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)

The system converged on three classes of multimodality-defined GBM: MOFS1, which is mainly non-enhancing, pathologically of low cellularity and atypia, and is associated with neuronal gene; MOFS2 which is associated with nodular enhancement, atypia and increased cellularity, and genes/proteins involved in the cell cycle; and MOFS3 which is associated with ring enhancement, infiltration of lymphocytes, and genes/proteins related to an inflamed tumor microenvironment. Survival analysis identified MOFS1 and MOFS2 to be prognostic and related to previously described signatures of proneural and proliferative GBM states, respectively. The authors nominate a STRAP2 as a histopathological marker of MOFS2 but provide little evidence that the test for its expression is robust. The authors then develop a deep neural network algorithm to classify GBM MOFS classes based on imaging studies in a large cohort. This represents a key advance in the field. The inclusion of a new survival cohort is significant. The tool to classify GBM MOFS classes based on radiology is available to researchers, which is another significant advance. I believe this work is of high significance to the field and will have a lasting impact, provided the following points are addressed.

A key technical issue that is not addressed is the following: What is that the rationale for integrating all the modalities? For example, it is not clear that adding histopathological data improves clustering. Could the model have generated the same classifications, which appear to align with prior predominantly RNAseq-based classifications, with fewer modalities?

The main critique I have is related to MOFS3 and mesenchymal GBM. The Mesenchymal GBM state has reproducibly been shown to exhibit a worse prognosis; how do the authors reconcile the finding with their data on prognosis showing that proliferative GBM MOFS2 is associated with a worse prognosis? One key limitation in determining the prognostic significance of MOFS3 is the following: MOFS3 is heterogeneous, and its stability as it is defined can be challenged. In S4A, the proteomic data shows that it is a mixture of samples with predominantly inflamed TME and ones that have inflamed TME and high levels of proliferation proteins. This is also seen in Figure S3A. This heterogeneity should be addressed, and sub-clustering MOFS3 may be attempted before survival analysis can be done on MOFS3 sub-classes.

Relevant to this point, the authors subclassify MOFS3 into stroma-high and stroma-L samples and find that the stroma-high samples have an equally bad prognosis compared to MOFS2. How do these categories correspond with the MOFS3 sample with high cell-cycle gene expression?

Another major point is that the authors indicate that MOFS3 showed increased lymphocyte infiltration. This is unusual in untreated GBM. Can the authors quantify this effect and indicate if it is focal or more widespread? In looking at the metadata table (Clinical), all GBMs are primary – the term secondary GBM is no longer used. Is this opposed to recurrence (i.e., primary pre-treatment vs. secondary post-treatment)? There is a column that indicates recurrence status, though that has many missing values, and it is difficult to find information on whether the samples analyzed in the MOFS modeling were a mixture of primary and recurrent GBM – the latter is known to be enriched for an immune TME, or included treatment-naïve cases (primary) only, and the column in the metadata for recurrence is from prospective analysis. The methods note that prior history of chemo/radiotherapy is an exclusion criterion. Providing this information is critical to understand the results of the paper.

A critical point in interpreting survival analysis: As a general point, I strongly encourage the authors to conduct cox-proportional hazards models where they account for age and MGMT methylation classes in their survival analysis. If the MOFS categories are still predictive of prognosis after accounting for the two main biological variables that drive the risk of death, then the significance of this new schema is fortified. Otherwise, the significance of the system should be re-examined. As a minor point - Figure S6 – the legend does not describe panels A and B. The main text indicates these panels should be non-MOFS classes. It is not clear if the categories examined are the original Wang classes or the MOFS categories. Regardless, in some cases, the PN class has essentially no cases assigned to it, making the interpretation of the survival differences difficult.

With regard to the comparison of the MOFS system to other classification systems: The authors concluded that there was little overlap between their MOFS classes and the system promoted by Wang et al. Examination of Figure 3A shows that MOFS1 and MOFS2 are stable and correspond to Proneural and mesenchymal classes, respectively, but MOFS2 is more heterogeneous. The conclusion that there is little overlap is a bit of an overstatement. In this regard, it is important to acknowledge other studies that used snRNAseq or scRNAseq to classify GBM states based on cellular composition – e.g., Al-Dalahmah et al 2023 (10.1038/s41467-023-38186-1) where the authors described an analogous GBM tissue states system that is prognostic and captures the essence of the MOFS classes.

The section on predicting IC50 in clinical samples based on results from a published GBM cell line drug-treatment database is vague and lacks validation. This section tries to predict which MOFS classes would respond to a specific drug; however, assigning MOFS classes to cell lines as a way to provide some validation is problematic because MOFS3 is defined as the TME, which is lacking in vitro. MOFS1 is defined by the presence of normal brain cells (Neurons, astrocytes, oligodendrocytes), which is also lacking in vitro. Also, the methods do not provide sufficient details on how the analysis was done. This section is confusing, and I recommend that it should be removed from the manuscript.

Finally, in the discussion, the authors should acknowledge the limitations of the current model in assigning MOFS classes to different parts of the tumor examined on imaging. For example, it is very common to have a large non-enhancing component of the tumor (MOFS1) and a small component of nodular enhancement (MOFS2). The way the data is presented, the entire tumor receives a MOFS2 class – if I understood their model correctly. It is possible that refining the system spatially may improve its predictive power. This is especially important in the post-treatment recurrence setting. The authors should also include a statement on the challenge of distinguishing treatment effect from actual tumor in the post-treatment recurrence setting and comment on the potential utility of this system in advancing the solution for this age-old question.

Minor points:

1-Are the features that were retained by the MOFS interpretable? Were all features equally important?

2-The WSI's were scanned at 5x, and then the images were segmented with cell profiler. It is unclear if the WSI's were done at 5x (low resolution) or at 20x (adequate resolution) – both are listed in the methods.

3-The MRI images and WSI with corresponding features for each WSI/MRI image are not provided. The authors should make this dataset accessible.

4-The IHC in Figure 4G and S10 are not compelling. Additional validation studies of the antibody used are needed. There is barely any signal in the MOFS2 images. The authors should show positive controls and negative controls. Based on that, the interpretation of panel 4J is difficult. The IHC panels in the S10 are low resolution and cannot be assessed.

5- same point for S100A4 IHC (Fig 5N and S11D)– the panels are difficult to interpret. The images are low resolution, and the signal-to-background ratio is low.

6- Figure 5D shows that the authors are inferring the presence of B-cells and mast cells in GBM tumors – this is surprising. Neuropathologic assessments of primary untreated GBM sections rarely reveal T-cells, let alone B-cells. Mast cells – I have never seen them in GBM section – and I have looked at more than a thousand glioma sections. What is the evidence that these cells exist? It is not clear from the text or methods how this panel was generated. These details must be provided.

7- The paragraph starting with “Specific Drug Candidates for MOFS Subtypes” should indicate that the GBM cell line data is published by another group.

8- Figure 6C – survival analysis must control for MGMT methylation status.

9- the term CAF is not defined

Reviewer #3

(Remarks to the Author)

A postdoctoral researcher in my laboratory and myself were invited to co-review this manuscript for Nature Communications.

Overall, the paper addresses a critical problem in glioblastoma translational and clinical research related to glioblastoma classification (or subtyping), which is essential to study the underlying biology of these highly heterogeneous tumours and

for stratification of patients for therapies better tailored to different tumour subtypes. A better classification of glioblastoma tumours with prognostic value is hence urgently needed.

Several attempts have been made to classify these tumours into different subtypes based on RNA expression (e.g. Philips PMID: 16530701; Wang PMID: 28697342), but the predictive prognostic value of such classification is highly debatable and not entirely consistent across different cohorts of patients, suggesting at some extent, these classifications are not robust and unstable.

Heterogeneity, however, does not occur only at the RNA level but is also reflected in gene mutations, chromosomal gain/loss, gene and protein expression profiles, tumour location, histopathological features, different transcriptional tumour cell states and non-malignant cell composition in the tumour microenvironment. Yet, most classification methods only focus on one modality (e.g., RNAseq, MRI, or histopathology). The use of single modalities prevents cross-association of different features of individual tumours, producing non-optimal classification that lacks prognosis value. This work addresses this issue and implements multimodal analysis for the classification of wild-type IDH glioblastoma based on multimodal data.

We found the work quite interesting and well-designed. Moreover, integrating multimodal data and developing a deep learning classifier for non-invasive subtype prediction is a significant step forward in glioma research. However, we think these are some concerns the Authors need to address to maximise the reproducibility, impact and scope of the results presented, which we believe can significantly contribute to the Neurooncology community in particular and broad field of cancer research and AI applications in cancer.

Main comments:

1. Multimodal fusion model (MOFS model). The methods section mentions that to create this model, "top-n" variables were selected from each layer (radiology, WSI, transcriptomics, and proteomics). 2640 variable combinations were then used for unsupervised cluster analysis. Can the authors include/mention/describe the variables used for this and include this as Supplemental material? I found this is critical as the model can be biased towards one of the specific layers (e.g. transcriptomics or proteomics).
2. Multimodal fusion model (MOFS model). The imbalance between subtypes, particularly MOFS3, could bias model performance. Did the authors take this into account and could describe whether techniques such as oversampling or class weighting were also used?
3. Data Augmentation: No mention of data augmentation (in this multimodal model and also in the MRI model) was made, which could help mitigate overfitting.
4. Multimodal fusion model (MOFS model). Can the authors expand on the logic and the algorithms for CPI and GAP calculations?
5. Multimodal fusion model (MOFS model). The authors should show (quantitatively) the differences in performance when using multimodal data vs. single-modal data to classify the different MOFS subtypes. For example, after defining the model final model (MOFS model), use this to label patients' groups as ground truth. Then, assess the performance of radiology variables alone, WSI variables alone, transcriptomics variables alone and proteomics variables alone and compute AUC or confusion matrices to determine/show how the multimodal approach overcomes obvious issues in single modal data.
6. Multimodal fusion model (MOFS model). I am puzzled by the authors' failure to consider obvious demographics (sex and age) and MGMT methylation status that are critically relevant to a patient's prognosis (i.e., survival data to the standard of care). In this regard, would it be helpful if the authors could highlight in the PCA plot (or add additional PCA plots) where dots in the graphs are also labelled according to the patient's sex, patient age (higher or lower than the median for GBM), and MGMT methylation status?
7. The authors should consider including interpretability tools (e.g., SHAP or Grad-CAM) to increase transparency in clinical predictions associated with MOFS classes, which is crucial for trust in medical AI.
8. Related to Figure 2A, "Radio-pathological and Biological Peculiarities of MOFS Subtypes", section, Page 8). The authors comment on differences observed at the histopathological and radiological levels across the MOFS groups. Authors should add stats on these observations, either formal (through image analysis) or qualitative (e.g., by stating that they were observed in 10/12 MOFS1 patients) quantitation. If qualitative results are provided, please include a multipanel with corresponding images. In the images, please also include the patient's demographic information (sex, age, methylation status, if possible).
9. Table S4. The legend does not provide sufficient information to determine whether this table includes RNA or protein data.
10. Figure 2C is not cited in the paper.
11. MOFS ensemble classifier (MOFS classification model based on RNA transcriptomics data only). Related to the methods section.
 - (i) The authors should provide training results for the NN method employed.
 - (ii) The authors should provide a confusion matrix for validation tests (point 6 in the methods). Values reported as

percentages are not very informative if the sample size is low. Authors should also split the validation results based on sex, age, and MGMT methylation status to identify bias in the model's performance about patients' demographics. Together, this will determine the consistency and stability of the proposed new classification.

12. Validation of MOFS Subtypes and Comparison with Traditional Classifications. I found it quite informative the authors aim to benchmark this classification approach with previous classifications. However, a new report also provides a new classification based on TME composition with prognostic value (White K et al., *Annals of Oncology*, 2023 PMID: 36494005) that can overlap with the MOFS classification proposed in this study. It is critical that the authors also benchmark the MOFS model against the TMElow/medium/high classification, repeating the same analysis as in Figure 3A and Figure S6.

13. Similarly to comment 12, recently, patient-derived cell lines were subgrouped into three groups based on pharmacological response and RNA expression (PMID: 39112531) that can somewhat overlap with the MOFS classification. Again, the authors are encouraged to benchmark the MOFS classification against the model in this work (using RNAseq available from primary tumours).

14. Figure title and Legends for Figures S6 and S7 are swapped.

15. It was mentioned that the MOFS ensemble classifier has an accuracy of ~90% (methods section, see also comment 8). Figure S7 shows pathway enrichment analysis for MOFS subtypes using this MOFS ensemble classifier. Although the general trend is accepted (matches the results in Figure 2B), many patients do not replicate such a trend. So, it would be suitable for the authors to report what number of patients agree with the results in 2B and how many do not. This also relates to the use of single-modal vs multimodal models.

16. Figure 5. Stromal component. The authors did not include blood vessels in the stromal component analysis (i.e. endothelial cells and pericytes). Figure 5M suggests that this could be quite an essential non-malignant cell component in MOFS3 (e.g. FAP, ACTA2, PDRGFA expression).

17. Figure 5L_labels in the inset. Should be "high/low-Str" not "high/low-CAFs".

18. Including MGMT methylation status is critical in the context of TMZ sensitivity and Figure 6B.

19. MRI Classifier. This is an excellent practical implementation of the results of this work. However, we believe it needs more detailed validations as follows

i) Why did the authors use DNN rather than CNN, which is more robust and accurate? The DNN model used for classification could benefit from a more sophisticated CNN architecture that better suits image data.

ii) Can the authors describe the features chosen (and the rationale for selection) to build the MRI classifier?

iii) Overall, more details on what was used as ground truth for model training, testing and validation of the MRI model are needed in the methods section and include discussions on the variations in AUC results if different benchmarking is used. See also the comment below.

iv) Page 21 mentions data on the MRI classifier performance against the FAHZZU2 cohort. This cohort is not a whole multimodal. We believe that this validation should focus on multimodal data (FAHZZU1). However, if FAHZZU2 is included, we wonder whether the labels on this cohort were based on transcript expression using the MOFS ensemble classifier, and, if this is the case, discuss whether this could be a cause of the drop of values in AUC results. Overall, more details on what was used as ground truth for model training, testing and validation of the MRI model are needed in the methods section and include discussions on the variations in AUC results if different benchmarking is used.

v) There is a risk that the MRI classifier loses accuracy regarding the new multimodal classification proposed in this study. So, in this regard, it would be helpful to know how "transcriptomic classes predictions" made by the MRI classifier by itself correspond (or not) to (i) traditional classifications (Philips, Wang and White) and (ii) the MOFS ensemble classifier applied to FAHZZU1 and FAHZZU2 and for (i) and (ii) display data as in Figure 3A.

vi) Amend/correct the labels in Figure S12 "training, testing, validation".

vii) An essential limitation of this study is the lack of external validation of the connection between the MRI model and drug-sensitivity predictions. The authors should look into performing this analysis (maybe establishing new collaborations) to enhance their findings' validity further.

viii) Recent analysis of patient-derived cell lines reported drug sensitivity assays for 3 different pharmacological subgroups in glioma (PMID: 39112531). (see also comment 13). Is there any correlation between drug sensitivities reported for groups in this study vs drugs (and groups) in PMID: 39112531.

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career

Reviewer #5

(Remarks to the Author)

The manuscript introduces an innovative multimodal fusion subtyping (MOFS) framework that integrates diverse data modalities—including radiology, pathology, whole-exome sequencing, RNA sequencing, and mass spectrometry—from a cohort of 122 patients with IDH-wildtype adult glioma. The study identifies three distinct subtypes: MOFS1 (proneural), MOFS2 (proliferative), and MOFS3 (TME-rich), each associated with different prognostic outcomes and therapeutic responses. This integrated approach is highly relevant, as it addresses the complex heterogeneity of gliomas and demonstrates potential for advancing personalized treatment strategies. By incorporating multiple data types, this work aligns with current trends in precision medicine and offers a robust framework for future studies. The findings are promising, and the approach appears capable of improving patient stratification in a way that may significantly inform therapeutic decisions.

However, while the results are compelling, the manuscript lacks sufficient detail in key areas, particularly in terms of the reproducibility of the analysis. Below are specific concerns and suggestions for improvement:

The results of the study are interesting, but I have significant concerns about the lack of detailed description regarding some critical aspects of the data analysis pipeline. In particular, the sections on the "Development of MRI Classifier for Non-invasively Predicting MOFS" and the "Deep-neural network to predict MOFS" require more clarity. Due to insufficient details, it is difficult to fully understand or reproduce the proposed classification models.

Clustering Methodology: This section is central to the manuscript, presenting a new analysis pipeline involving various methods. However, without more specific information, it is difficult to replicate the results. Detailed explanations of the following aspects would greatly enhance reproducibility:

Cluster variable selection:

1. **Selection Criteria:** The paper does not clarify the criteria or threshold used to define the "top-n" variables for each modality. Is 'n' consistent across all modalities, or does it vary? If variable selection differs between modalities, how was the balance maintained to avoid over- or underrepresentation of specific data types?
2. **Preprocessing:** Additional information on preprocessing steps such as normalization, scaling, or transformation is needed, especially prior to calculating the median absolute deviation (MAD). Were these steps applied uniformly across all modalities?
3. **CPI/GAP Calculation:** The formula for Cluster Prediction Index (CPI) and GAP statistic is not provided. What specific formula was used to compute these metrics? Were any software or tools used for these calculations, and if so, which ones?

Parameter Settings for the 11 Algorithms:

4. **Missing Parameters:** The manuscript does not provide the specific input parameters for the 11 algorithms used in the clustering process (such as the number of clusters, distance metrics, regularization parameters, etc.). This is critical for understanding and replicating the analysis.
5. **Algorithm Choice:** The reference (see below) cited compared 13 clustering algorithms, why just 11 of them were used. Please clarify why these specific 11 algorithms were selected and why others were excluded. Pierre-Jean M, Deleuze JF, Le Floch E, Mauger F: Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Brief Bioinform* 2020, 21(6):2011-2030.

Jaccard Matix and COCA:

6. **Binary Transformation:** The method used to transform clustering results into binary data is unclear. How were clustering results encoded as binary vectors?
7. **Jaccard Distance Calculation:** The manuscript lacks details on how the Jaccard distance matrix was calculated. Were any thresholds or adjustments applied during this process?
8. **COCA:** The choice of number of consensus clusters is not detailed. It's not specified if any validation or bootstrapping techniques were applied to assess the stability of the consensus clusters.

Silhouette coefficient:

9. **Distance Metric:** How was the silhouette coefficient computed? Was Euclidean distance or another measure used to assess clustering quality?
10. **Clustering Algorithm:** Which clustering algorithm was used to compute the silhouette coefficient? This information is necessary for assessing the reliability of the clustering analysis.

MRI processing and Radiomics Concerns:

11. The manuscript mentions the creation of ADC maps on a voxel-by-voxel basis using a single-exponential model. Was any external software used for this process, and if so, which program?
12. Did you perform histogram matching for all MRI sequences? If so, which toolbox or software was used? Providing details such as histogram levels, the number of match points, and how ADC normalization was performed is important for replication. This is especially critical given the use of MRI data in the Deep-Neural Network as a tool for external validation. Providing these details is essential for ensuring the replication of your findings and verifying the robustness of the model across different datasets and research environments.
13. The authors resampled all input modalities from 5 mm slice thickness to 1 mm³ isotropic voxels. This process could introduce artifacts or biases. How was this resampling validated, and what impact might it have on the analysis?
14. The manuscript mentions the use of PyRadiomics for feature extraction, but there are no details about the specific

parameters or version used. Providing the PyRadiomics parameters file would significantly enhance reproducibility.

Unclassified samples:

15. Some patients were not assigned to a MOFS subtype, and a core set of 116 cases was used. What are the reasons for the failure to classify some patients? Please elaborate on how these cases differed from the classified ones.

Deep-neural network to predict MOFS: while I think the strength of the paper lies in the tool to predict MOFS types only based on imaging parameters:

16. Accuracy Drop: The Deep-Neural Network trained on the FAHZZU1 dataset exhibited a drop in accuracy (0.825) when applied to FAHZZU2. Was this drop due to reduced generalizability, overrepresentation of radiological features, or another factor? Please clarify.

17. Independent Dataset: Since the dataset used to develop the Deep-Neural Network (DNN) comes from a single institution (FAHZZU), I recommend testing the model on an independent dataset to validate its generalizability. Utilizing an open-source dataset, such as the UCSF glioma dataset, would provide a more robust evaluation and ensure that the model performs well across different populations. Has external validation been considered in your analysis?

Code and Reproducibility:

18. Software & Code: I strongly encourage the authors to share the code used for the analysis, including specific versions of R, Python, or other software packages. This level of transparency will enhance reproducibility and allow other researchers to validate the findings.

Conclusion

In summary, while the manuscript presents valuable and relevant findings regarding MOFS subtypes in IDH-wildtype gliomas, it lacks sufficient methodological detail to ensure reproducibility. Addressing these specific concerns will greatly enhance the quality of this work and ensure it meets the standards of transparency and rigor expected in the field. By providing additional details on data preprocessing, algorithm parameters, and software versions, the authors can significantly strengthen the manuscript and facilitate future research based on these promising findings.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

(Remarks on code availability)

I thank the authors for addressing my comments. There is still an issue that is still not clear and is rather confusing. Regarding the alignment of the MOFS clusters with established such as Wang et al. classes – the proliferative MOFS2 appears to align with the proneural classifications in Figure 3C. However, this data conflicts with that shown in figure S14B, where the RNA or MRI MOFS classifiers results demonstrate that MOFS2 aligns with Wang's mesenchymal classification, and MOFS3 align with Wang's proneural classification, as opposed to Wang's mesenchymal class as shown in figure 3C. This issue is partially addressed in the text, as the authors highlight the importance of having multiple modalities for accurate classification and prognostication, and the comments on loss of accuracy with the DNN algorithm. The issue remains that the two figures are quite different, and explanation is needed. It is also noteworthy that in their new data provided in figure 5I, MOFS2 and MOFS3 H-stroma appear to behave similarly. However, the statistics to support that should be included. This point is significant and should be emphasized.

Reviewer #3

(Remarks to the Author)

Thanks for the authors for taking into consideration all the comments I initially raised.

I appreciate their understanding of transparency and benchmarking with previous work as a way to further test the capacity of the proposed new classification method for primarily diagnosed glioblastomas.

I consider the authors has satisfactorily addressed all my comments.

Just I noted that in response 6-8 to Reviewer 5 (not this reviewer) there is reference to "(e.g. PMID:123 and PMID:456)", so correct references should be added.

I also completely understand the limitations of the work and recognise that publication will increase interest and contribute to external validation of the model.

Overall, I found the article very interesting and recommend it for publication.

(Remarks on code availability)

I reviewed the content of the link describing the code with the readme documentation and the corresponding manual. The codes are well organised and structured with examples lines. I have not installed the code myself and I have not run it.

Reviewer #5

(Remarks to the Author)

Dear Authors,

Thank you for your thorough responses to my comments and for the significant efforts made to address the concerns I raised. After reviewing your responses and the updates to the manuscript, I believe that you have adequately addressed the points of concern.

External Validation:

Your explanation regarding the challenges of using an independent dataset, such as the UCSF glioma dataset, is clear and reasonable. The detailed discussion of the absence of ADC and RNA-seq data in the UCSF dataset effectively explains why external validation using this dataset was not feasible. I appreciate your transparency in acknowledging this limitation and your proactive approach to exploring collaborative opportunities for developing multimodal datasets in future research. The inclusion of this discussion in the manuscript provides an honest assessment of the current limitations while underscoring the importance of cross-cohort validations in future studies.

Code and Reproducibility:

I commend you for making the complete analysis pipeline, along with detailed documentation, available through a public GitHub repository. This initiative is crucial for transparency and reproducibility, as it allows other researchers to replicate your findings and apply your methods to their datasets. The inclusion of software versions, parameter settings, and user guides further enhances the usability of the MOFS framework, making it accessible to a broader audience in the research community.

Methodological Transparency:

I appreciate the additional details you provided regarding data preprocessing, feature selection, clustering methodologies, and algorithm parameters. The expanded descriptions of the computational steps and parameter configurations significantly enhance the clarity and reproducibility of the analysis pipeline. These updates ensure that the methodological aspects of the study are robust and well-documented.

Overall, while the limitation of not being able to validate using an external open-source dataset is acknowledged, the steps you have taken—particularly the release of the code—provide other researchers the opportunity to test your methods on their own datasets. This contribution strengthens the reproducibility and applicability of your framework in glioblastoma research.

Thank you for your responsiveness and for the improvements made to the manuscript. I believe the work now represents a valuable contribution to the field.

(Remarks on code availability)

I reviewed the code provided in the GitHub repository, and it appears well-structured and understandable. The repository includes a README file with detailed instructions for installing and running the application. Furthermore, the inclusion of a comprehensive user manual adds substantial value by providing clear guidance on the various functions and workflows implemented in the MOFS framework.

The code is meticulously documented, with specific sections dedicated to the clustering, classification, and feature selection processes. These details are highly beneficial for understanding how the methods were implemented and applied in the analysis. The repository also specifies the versions of R and libraries used, ensuring that users have the necessary information to replicate the computational environment.

However, I did not attempt to install or run the code myself, primarily due to the inclusion of numerous highly specific libraries and dependencies, some of which may have compatibility issues depending on the user's computational setup. This reliance on specialized libraries, while essential for the advanced analyses performed in the study, could limit the usability of the code for some researchers without extensive computational expertise or the exact required environment.

In summary, while the code seems to be a well-prepared resource for the research community, the complexity of its dependencies may present challenges for some users. That said, the availability of clear documentation and a thorough manual should help mitigate these challenges for those motivated to replicate the study.

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/>

Dear Reviewers,

Thanks very much for taking your time to review this manuscript. We really appreciate all your comments and suggestions! Please find our itemized responses below and our revisions in the re-submitted files.

Thanks again!

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The system converged on three classes of multimodality-defined GBM: MOFS1, which is mainly non-enhancing, pathologically of low cellularity and atypia, and is associated with neuronal gene; MOFS2 which is associated with nodular enhancement, atypia and increased cellularity, and genes/proteins involved in the cell cycle; and MOFS3 which is associated with ring enhancement, infiltration of lymphocytes, and genes/proteins related to an inflamed tumor microenvironment. Survival analysis identified MOFS1 and MOFS2 to be prognostic and related to previously described signatures of proneural and proliferative GBM states, respectively. The authors nominate a STRAP2 as a histopathological marker of MOFS2 but provide little evidence that the test for its expression is robust. The authors then develop a deep neural network algorithm to classify GBM MOFS classes based on imaging studies in a large cohort. This represents a key advance in the field. The inclusion of a new survival cohort is significant. The tool to classify GBM MOFS classes based on radiology is available to researchers, which is another significant advance. I believe this work is of high significance to the field and will have a lasting impact, provided the following points are addressed.

A key technical issue that is not addressed is the following: What is that the rationale for integrating all the modalities? For example, it is not clear that adding histopathological data improves clustering. Could the model have generated the same classifications, which appear to align with prior predominantly RNAseq-based classifications, with fewer modalities?

Response: We sincerely thank you for your insightful comments and for highlighting this important point. We would like to clarify the rationale behind integrating multiple modalities into our analysis. This study aims to integrate multiple data modalities—DNA (genetic level), RNA (transcriptional level), protein (functional level), pathology (cellular level), and radiology (anatomical/tissue level). Each of these modalities represents a distinct dimension of disease biology, with special characteristics and inherent differences. However, these dimensions are not isolated; they are causally interconnected, reflecting different levels of cellular and tissue processes that influence and drive one another (ref 1-5). For example, genetic mutations (DNA) lead to transcriptional alterations (RNA), which subsequently affect protein expression and function, ultimately influencing cellular structures and the

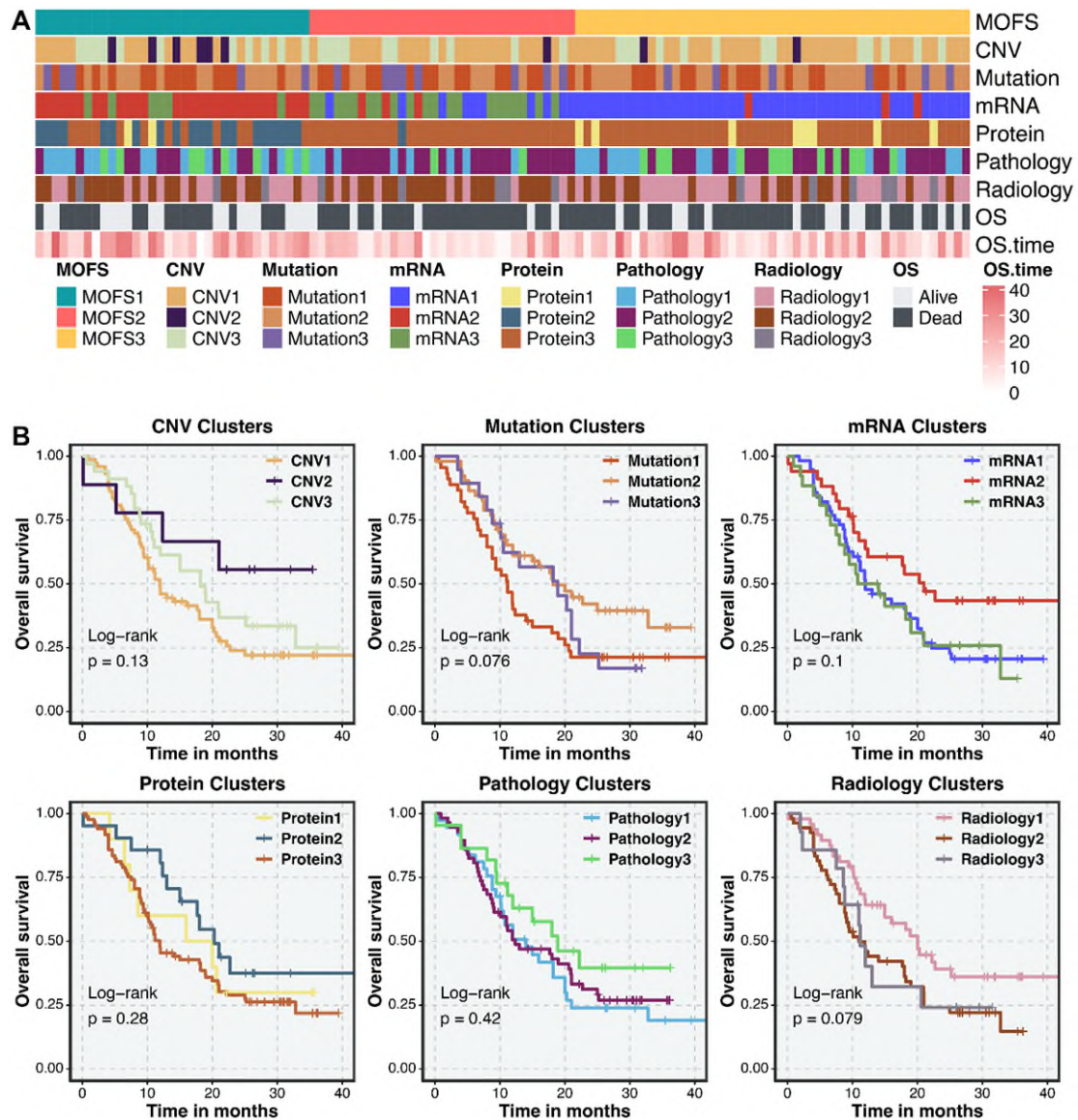
tumor microenvironment, as visualized in pathology images. Similarly, radiological imaging captures macro-level features that reflect the underlying molecular and cellular changes. Specifically, cellular-level microscopic phenotypes, resulting from gene mutations, aberrant signaling, and abnormal expression, profoundly affect key cellular processes such as cell proliferation, inflammation, and angiogenesis (ref 6). These alterations are captured by histopathological images. Macroscopic tumor features, such as shape, texture, edema, and necrosis, can be visualized using advanced imaging techniques like multiparametric MRIs. Previous studies have shown that radiological and histopathological images can predict mutational status or gene expression levels (ref 7-9). This not only establishes a link between imaging data and molecular omics but also provides the theoretical foundation for multimodal fusion analysis.

By integrating these modalities, we aim to construct a more comprehensive, multi-dimensional profile of the disease. This allows us to leverage the complementary strengths of each modality, creating a more holistic view of disease mechanisms that can reveal novel insights that would be obscured by single-modality analyses. This fusion strategy enables us to explore the relationships between modalities and how they contribute to discrepancies in patient outcomes, such as survival and drug response (ref 1 and 10). Fusion strategies can be categorized into early, intermediate, and late fusion (ref 1-2). Early fusion concatenates all modalities into a single matrix, but this can lead to the "curse of dimensionality" and variable shifts, which may negatively affect downstream analyses. Late fusion involves analyzing each omics layer separately and then integrating the results, but it may miss out on the complementary interactions between modalities. Intermediate fusion integrates data during clustering, allowing for the identification of multimodal joint clusters, connecting dependencies between different omics layers, and revealing underlying biological mechanisms (ref 1, 2, 5). Given the complementarity of information in each omics layer, intermediate fusion is particularly effective for uncovering complex biological processes (ref 1-2).

We incorporated histopathological features into our multimodal model to enhance both its interpretability and clinical applicability. Histopathological data provide critical microstructural information—such as cell morphology, density, and immune cell infiltration—which are partially captured by RNA and protein data. Additionally, the inclusion of histopathological features improves the clinical interpretability of our model, allowing clinicians to better understand the distinctions between subtypes. This potentially enables clinicians to preliminarily assess a patient's subtype by examining pathology slides or imaging, thus supporting more informed clinical decision-making. Furthermore, integrating histopathological data creates the foundation for potential perioperative diagnostic models, where imaging-pathology combinations could improve accuracy and assist clinicians in making reliable subtype predictions.

To further demonstrate the added value of histopathological data, we conducted a series of comparative clustering experiments to assess the specific contribution of each modality layer, including histopathological features, to the prognostic

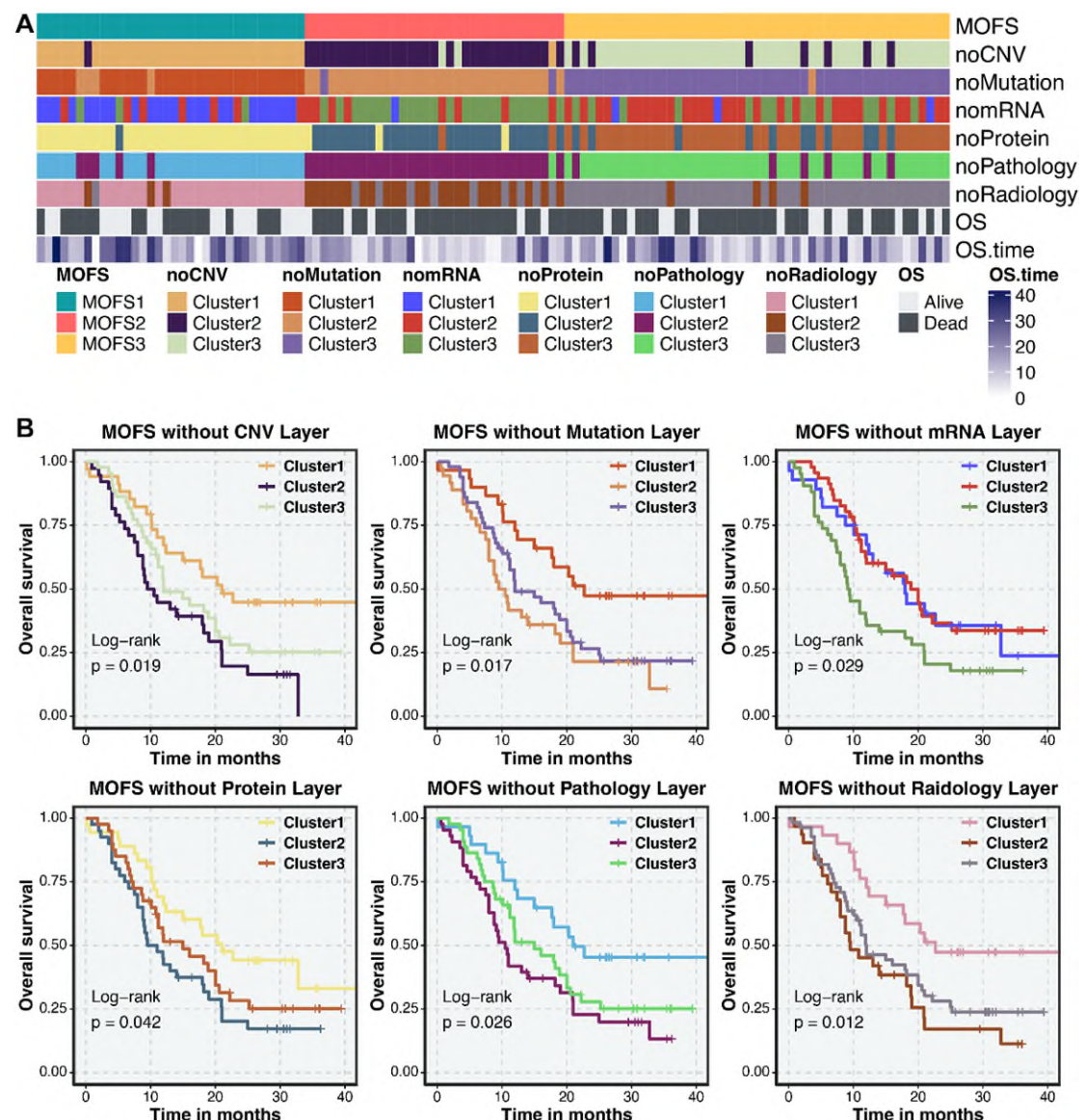
performance of patient clustering. In these experiments, we performed consensus clustering analysis with individual modality layers and also evaluated the effect of excluding individual modality layers. For consistency, we maintained the same set of selected features for each omics layer and a clustering number of three, as in the MOFS framework. For single-modality clustering, we observed that the agreement between single-modality clusters and the MOFS subtypes was generally low (**Ref Figure 1a**). Moreover, survival analysis demonstrated limited prognostic significance for single-modality clusters, with none achieving statistical significance in distinguishing survival outcomes (**Ref Figure 1B**). When excluding individual modalities, the remaining multimodal clusters showed a higher concordance with MOFS subtypes (**Ref Figure 2A**). These clusters were also prognostic, achieving statistically significant separation of survival curves (**Ref Figure 2B**). However, their survival discrimination was consistently inferior to the full multimodal MOFS framework (**Ref Figure 3**). This suggests that each modality contributes unique and complementary information to the clustering framework. In particular, the inclusion of histopathological features enhanced both the prognostic power and the clinical interpretability of the MOFS framework. Histopathological data provide critical microstructural insights, such as immune infiltration and cell atypia/density. By integrating histopathology with other omics layers, the MOFS framework achieves a more holistic and clinically meaningful classification of IDH wild-type glioma subtypes. These findings confirm the pivotal role of histopathological data in improving patient stratification and support the robustness of our multimodal clustering approach. We have incorporated these results into the revised manuscript to emphasize the specific value of each modality.



Ref Figure 1. Single-modality clustering and its prognostic significance

(A) Heatmap showing the clustering results derived from single modalities, including copy number variation (CNV), mutation, mRNA expression, protein expression, pathology, and radiology. Each row represents the clustering results for a specific modality, and each column represents a patient. Colors indicate the clusters assigned by each modality.

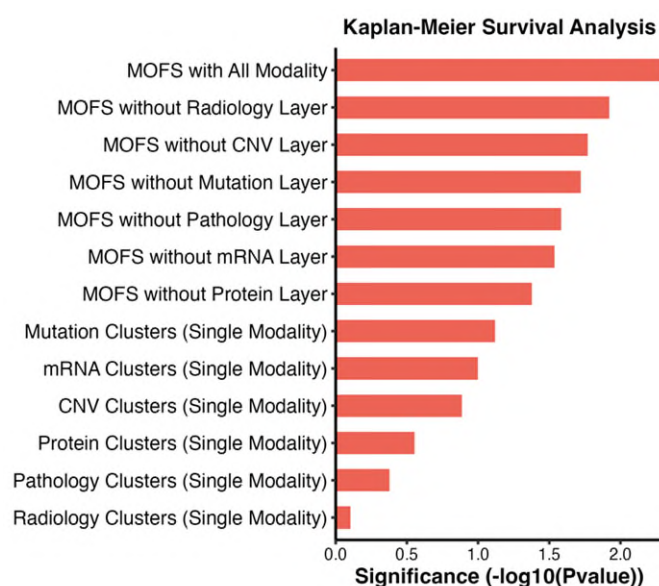
(B) Kaplan-Meier survival curves for clusters derived from individual modalities. The curves represent overall survival outcomes for patients in each cluster. Log-rank p-values indicate the prognostic significance of each single-modality clustering. None of the single-modality clusters achieved statistical significance, highlighting their limited ability to distinguish survival outcomes.



Ref Figure 2. Effect of excluding individual modalities on clustering and survival analysis

(A) Heatmap illustrating the clustering results when each modality was excluded from the multimodal MOFS framework. Rows represent clustering results after excluding specific modalities (e.g., noCNV, noMutation, etc.), while the top row represents the original MOFS subtypes for reference. Colors denote patient cluster assignments.

(B) Kaplan-Meier survival curves for clusters obtained by excluding specific modalities.



Ref Figure 3. Survival significance of multimodal and single-modality clustering.

Bar plot showing the significance of Kaplan-Meier survival analysis for different clustering strategies, represented as $-\log_{10}(\text{p-value})$ from log-rank tests. The top bar represents the MOFS framework incorporating all modalities ("MOFS with All Modality"), achieving the highest prognostic significance. Subsequent bars represent clustering results when individual modalities are excluded ("MOFS without X Layer") and single-modality clusters ("X Clusters"). Single-modality clustering shows limited prognostic power, while multimodal approaches demonstrate significantly enhanced survival discrimination, especially when all modalities are included.

Regarding your point about previous classifications, we acknowledge that earlier subtyping systems, which were based primarily on RNA-seq data, defined subtypes such as Proneural, Classic (proliferative), and Mesenchymal, as shown in new Figure 3c. Our subtypes show only moderate correlation with these traditional classifications. Moreover, the 2016 and 2021 WHO classifications of CNS tumors have significantly advanced the classification of adult diffuse gliomas based on IDH mutations (ref 11-12), with the latest edition prioritizing IDH mutations over histological features. Given the poor prognosis and high prevalence of IDH-wildtype GBM, this subpopulation remains particularly relevant (ref 13). The earlier subtyping systems did not account for IDH mutations and were derived solely from transcriptomic data, which limits their prognostic value (ref 14-16). Although our subtypes are similar to previously reported ones, we believe that the subtypes we have defined are more biologically consistent and show better prognostic significance. For instance, we observed that the MOFS3 subtype, which was identified through multimodal integration, does not uniformly retain the tumor microenvironment (TME)-rich characteristics across all patients when analyzed through protein or RNA-seq data. While most of patients exhibit strong TME-rich features, some also display a mixture of proliferative traits. This heterogeneity is particularly challenging if the analysis is based solely on either RNA-seq or protein

data, as such unidimensional analyses might lead to higher clustering homogeneity, akin to traditional subtyping systems. In these systems, molecular features such as RNA or protein expression might obscure the diverse phenotypic behaviors seen in our MOFS3 subtype, thus failing to capture its full biological complexity. Additionally, our proliferative subtype, which exhibits the worst prognosis among the identified subtypes, deviates from previous reports. However, its stability and prognostic significance were consistently validated across external datasets, unlike other subtyping systems (new Figure S8-9). This finding suggests that our classification system offers more detailed stratification, particularly regarding the clinical significance of the proliferative subtype. Unlike earlier studies, which often did not account for the broader biological context or integrated multiple data layers, our approach ensures that these subtypes have better biological conservation and prognostic relevance. Therefore, it is clear that the subtypes we identified are distinct from earlier classifications, both in terms of their biological features and their ability to predict patient outcomes. This highlights the value of multimodal data integration in uncovering subtypes with more precise clinical implications.

We hope that these clarifications and additional findings will help address your concerns and will strengthen our manuscript. Thank you once again for your valuable feedback.

References:

1. Lipkova J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. 2022;40(10):1095-1110. doi:10.1016/j.ccell.2022.09.012
2. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer*. 2022;22(2):114-126. doi:10.1038/s41568-021-00408-3
3. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. 2015;16(2):85-97. doi:10.1038/nrg3868
4. Tini G, Marchetti L, Priami C, Scott-Boyer MP. Multi-omics integration-a comparison of unsupervised clustering methodologies. *Brief Bioinform*. 2019;20(4):1269-1279. doi:10.1093/bib/bbx167
5. Lobato-Delgado B, Priego-Torres B, Sanchez-Morillo D. Combining Molecular, Imaging, and Clinical Data Analysis for Predicting Cancer Prognosis. *Cancers (Basel)*. 2022;14(13):3215. Published 2022 Jun 30. doi:10.3390/cancers14133215
6. Weller M, Wick W, Aldape K, et al. Glioma. *Nat Rev Dis Primers*. 2015;1:15017. Published 2015 Jul 16. doi:10.1038/nrdp.2015.17
7. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J Clin*. 2019;69(2):127-157. doi:10.3322/caac.21552
8. Aftab K, Aamir FB, Mallick S, et al. Radiomics for precision medicine in glioblastoma. *J Neurooncol*. 2022;156(2):217-231.

doi:10.1007/s11060-021-03933-1

9. Schmauch B, Romagnoni A, Pronier E, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat Commun.* 2020;11(1):3877. Published 2020 Aug 3. doi:10.1038/s41467-020-17678-4
10. Şenbabaoğlu Y, Michailidis G, Li JZ. Critical limitations of consensus clustering in class discovery. *Sci Rep.* 2014;4:6207. Published 2014 Aug 27. doi:10.1038/srep06207
11. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol.* 2021;23(8):1231-1251. doi:10.1093/neuonc/noab106
12. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.* 2016;131(6):803-820. doi:10.1007/s00401-016-1545-1
13. McKinnon C, Nandhabalan M, Murray SA, Plaha P. Glioblastoma: clinical presentation, diagnosis, and management. *BMJ.* 2021;374:n1560. Published 2021 Jul 14. doi:10.1136/bmj.n1560
14. Phillips HS, Kharbanda S, Chen R, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell.* 2006;9(3):157-173. doi:10.1016/j.ccr.2006.02.019
15. Wang Q, Hu B, Hu X, et al. Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell.* 2018;33(1):152. doi:10.1016/j.ccell.2017.12.012
16. Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell.* 2010;17(1):98-110. doi:10.1016/j.ccr.2009.12.020

The main critique I have is related to MOFS3 and mesenchymal GBM. The Mesenchymal GBM state has reproducibly been shown to exhibit a worse prognosis; how do the authors reconcile the finding with their data on prognosis showing that proliferative GBM MOFS2 is associated with a worse prognosis? One key limitation in determining the prognostic significance of MOFS3 is the following: MOFS3 is heterogeneous, and its stability as it is defined can be challenged. In S4A, the proteomic data shows that it is a mixture of samples with predominantly inflamed TME and ones that have inflamed TME and high levels of proliferation proteins. This is also seen in Figure S3A. This heterogeneity should be addressed, and sub-clustering MOFS3 may be attempted before survival analysis can be done on MOFS3 sub-classes.

Relevant to this point, the authors subclassify MOFS3 into stroma-high and stroma-L samples and find that the stroma-high samples have an equally bad prognosis compared to MOFS2. How do these categories correspond with the MOFS3 sample with high cell-cycle gene expression?

Response: We sincerely thank you for your insightful comments regarding the

heterogeneity of MOFS3 and its potential impact on prognosis. We have carefully examined the heterogeneity of MOFS3 through further subgroup analyses and have gained valuable insights, which we address below.

First, we acknowledge that approximately one-third of MOFS3 samples exhibit high cell cycle activity, consistent with the heatmap shown in **Ref Figure 4A-B**. However, our analysis indicates that proliferation status does not correspond to stromal score within MOFS3. Specifically, there is no significant difference in stromal scores between the high and low proliferation groups within MOFS3, and the median stromal score is even slightly higher in the low proliferation group (**Ref Figure 4C**). Additionally, there is no correlation between cell cycle score and stromal score within MOFS3 ($P = 0.961$, $Cor = 0.007$).

Second, differential gene expression and functional enrichment analyses comparing MOFS3 high proliferation with MOFS2 samples (**Ref Figure 4D**) reveal that MOFS3 high proliferation samples remain enriched in TME-related pathways, consistent with the defining characteristics of the MOFS3 subtype. This result underscores that even in MOFS3 samples with high proliferation scores, their primary biological identity is still TME-rich, rather than proliferation-driven.

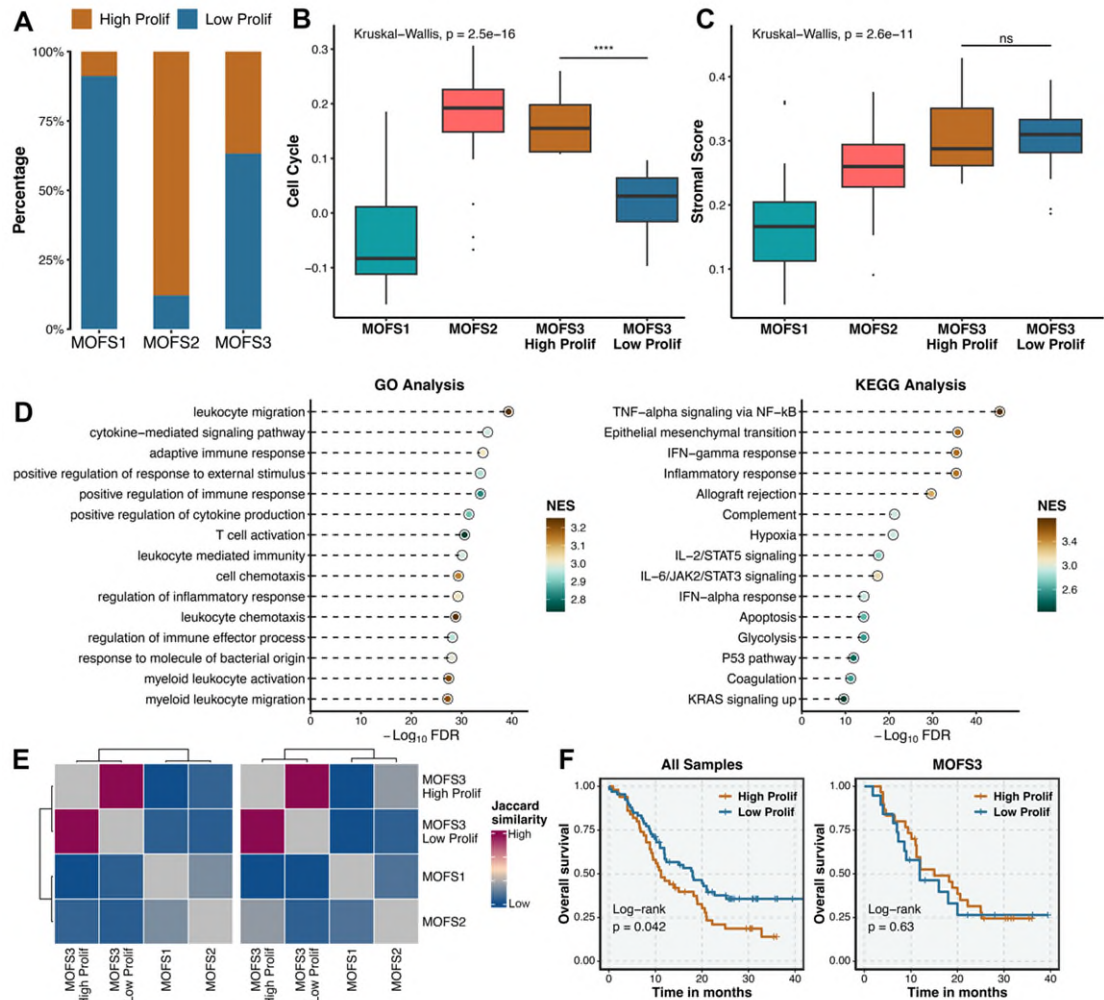
Third, our Jaccard similarity analysis (**Ref Figure 4E**) further supports this conclusion. Among all subgroups analyzed, the high and low proliferation groups within MOFS3 are the most similar, indicating that the primary attribute of MOFS3 remains its TME-rich nature, regardless of proliferation status.

Finally, survival analyses reveal distinct patterns. While cell cycle score is associated with worse prognosis across all samples ($p = 0.042$), it has no prognostic significance within MOFS3 samples ($p = 0.63$) (**Ref Figure 4F**). In contrast, stromal score is significantly prognostic within MOFS3 but not in MOFS1 or MOFS2 (Supplementary Figure S15D). These findings suggest that proliferation score is not a major determinant of prognosis within MOFS3, further supporting the stability and biological relevance of MOFS3 as defined in our study.

One important reason why our MOFS framework identifies subtypes differently from traditional RNA-seq-based classifications lies in the use of multimodal integration. Traditional clustering methods based solely on RNA-seq can achieve highly homogeneous clusters in terms of transcriptional profiles. However, they often oversimplify tumor heterogeneity by neglecting other biologically relevant dimensions, such as protein activity, histopathological features, and radiological characteristics. In contrast, our multimodal integration approach captures complementary and distinct aspects of tumor biology, providing a more holistic understanding of GBM subtypes. For instance, while MOFS3 contains a mixture of proliferative and non-proliferative samples, its core TME-rich identity is preserved across all analyses, a complexity that would likely be overlooked in single-modality clustering methods.

In summary, while MOFS3 exhibits heterogeneity in cell cycle activity, its primary biological identity as a TME-rich subtype remains stable. The prognostic value of MOFS3 is more closely tied to its stromal characteristics, further underscoring the stability and biological relevance of this subtype as defined in our study. The use of

multimodal integration in the MOFS framework ensures that we capture the full biological complexity of IDH wild-type glioma subtypes, offering insights that go beyond traditional RNA-seq-based classifications. We have included these additional analyses and interpretations in the revised manuscript to provide a more comprehensive understanding of MOFS3 heterogeneity and to address your concerns.



Ref Figure 4. Proliferation and stromal heterogeneity within MOFS subtypes.

(A) Proportion of high proliferation (High Prolif) and low proliferation (Low Prolif) samples in each MOFS subtype.

(B) Cell cycle scores across MOFS1, MOFS2, and MOFS3 proliferation subgroups.

(C) Stromal scores across MOFS1, MOFS2, and MOFS3 proliferation subgroups. MOFS3 has significantly higher stromal scores than MOFS1 and MOFS2, with no difference between the high and low proliferation subgroups.

(D) GO and KEGG pathway enrichment of upregulated genes in MOFS3 High Prolif versus MOFS2. MOFS3 High Prolif samples remain enriched in TME-related pathways, retaining the TME-rich identity of MOFS3.

(E) Jaccard similarity heatmaps for upregulated (left) and downregulated (right) gene sets. MOFS3 High Prolif and Low Prolif groups show the highest similarity, indicating that MOFS3's primary TME-rich identity is preserved.

(F) Kaplan-Meier survival curves for high and low proliferation groups. High proliferation predicts worse prognosis in the overall cohort ($P = 0.042$) but not within MOFS3 ($P = 0.63$).

Another major point is that the authors indicate that MOFS3 showed increased lymphocyte infiltration. This is unusual in untreated GBM. Can the authors quantify this effect and indicate if it is focal or more widespread? In looking at the metadata table (Clinical), all GBMs are primary – the term secondary GBM is no longer used. Is this opposed to recurrence (i.e., primary pre-treatment vs. secondary post-treatment)? There is a column that indicates recurrence status, though that has many missing values, and it is difficult to find information on whether the samples analyzed in the MOFS modeling were a mixture of primary and recurrent GBM – the latter is known to be enriched for an immune TME, or included treatment-naïve cases (primary) only, and the column in the metadata for recurrence is from prospective analysis. The methods note that prior history of chemo/radiotherapy is an exclusion criterion. Providing this information is critical to understand the results of the paper.

Response: We sincerely thank the reviewer for the thoughtful comments and the opportunity to clarify these important points. All samples included in our study were from treatment-naïve primary IDH wild-type GBMs and IDH wild-type astrocytic gliomas, as confirmed through metadata review and direct consultation with our clinical collaborators. This point has been explicitly clarified in the revised Methods section. The metadata column labeled “recurrence” reflects follow-up information indicating whether the patient experienced a recurrence after the primary tumor was sampled. However, only primary samples collected prior to any treatment (e.g., no prior history of chemo/radiotherapy) were included in our MOFS analysis. This ensures that the study focuses exclusively on treatment-naïve primary IDH wild-type gliomas. To avoid confusion, we have updated the methodological description in the revised manuscript.

Regarding the immune cell infiltration observed in MOFS3, we agree that immune cell infiltration is relatively limited for untreated GBM compared to other solid tumors, which is typically considered an immune “cold” tumor. The term “lymphocyte infiltration” may have been imprecise. Upon further review, we agree that the immune cell populations within GBM are diverse and include significant contributions from macrophages and microglia (Figure 5d). As such, we have revised the description to more accurately refer to “immune cell infiltration,” which encompasses all immune cell types, including both lymphocytes and myeloid cells, in order to avoid any misinterpretation. Notably, it is also important to emphasize that the immune scoring (based on omics data) in our study is relative (MOFS3 vs. MOFS2+MOFS1). While GBM as a whole remains predominantly immune “cold,” MOFS3 demonstrates relatively, averagely enriched immune cell infiltration compared to MOFS2 and MOFS1, as shown in Figure 5 of the original manuscript. This suggests that MOFS3 may exhibit a greater degree of immune activation within the context of GBM. Besides, histopathological validation further supports this

observation. Pathological slides revealed a relatively higher presence of immune cells in MOFS3 samples compared to MOFS1 and MOFS2. Importantly, this infiltration in MOFS3 is focal rather than widespread (we have added this description in the revised manuscript), as evident from the histological examination. This focal immune enrichment may represent an immune reserve within MOFS3 that could render this subtype more responsive to immunotherapy. While further investigation is required to substantiate this hypothesis.

A critical point in interpreting survival analysis: As a general point, I strongly encourage the authors to conduct cox-proportional hazards models where they account for age and MGMT methylation classes in their survival analysis. If the MOFS categories are still predictive of prognosis after accounting for the two main biological variables that drive the risk of death, then the significance of this new schema is fortified. Otherwise, the significance of the system should be re-examined. As a minor point - Figure S6 – the legend does not describe panels A and B. The main text indicates these panels should be non-MOFS classes. It is not clear if the categories examined are the original Wang classes or the MOFS categories. Regardless, in some cases, the PN class has essentially no cases assigned to it, making the interpretation of the survival differences difficult.

Response: We sincerely thank the reviewer for these insightful comments and suggestions, which have contributed to the rigor and clarity of our study. For the survival analysis, we conducted additional multivariable Cox proportional hazards regression analyses using external validation cohorts, which included information on key prognostic factors such as age, MGMT methylation status, and additional clinical variables, including radiotherapy and chemotherapy (**Ref Table 1**). Even after adjusting for these variables, MOFS1 consistently demonstrated an association with better overall survival, MOFS2 remained predictive of poor prognosis, and MOFS3 showed no significant association with survival. These findings align with the results presented in Figure 3 of the original manuscript, further confirming the robustness and predictive value of the MOFS classification system. These new results have been incorporated into **Supplementary Table S5A**, and the corresponding text in the manuscript has been updated. We appreciate the reviewer's suggestion, as these additional analyses reinforce the clinical relevance of the MOFS schema.

Regarding **Figure S6 (new Figure S8)**, we have revised the figure legend to provide descriptions of Panels A and B.

Indeed, we observed that in some datasets, the PN class was underrepresented or even absent. This issue likely arises from limitations in the external datasets themselves, as these classifications were generated using algorithms originally trained on different cohorts and subsequently applied to normalized data from independent datasets. The observed challenges with subtype distribution balance and survival separation further suggest that these algorithms may not generalize well to other datasets. In contrast, the MOFS framework demonstrates more consistent and clinically meaningful subtype classification across cohorts, emphasizing its

robustness and translational potential.

We hope these clarifications and additional analyses address the reviewer's concerns and further validate the significance and clinical utility of our study.

Ref Table 1. Multivariable Cox proportional hazards regression analysis for MOFS subtypes after adjustment for clinical covariates.

This table presents the results of multivariable Cox proportional hazards regression analyses performed in both the in-house cohort and external validation cohorts. The analysis accounts for key clinical variables, including age, MGMT methylation status, radiotherapy, and chemotherapy. Hazard ratios (HR), along with their 95% confidence intervals (HR.95L and HR.95H), are provided for each MOFS subtype (MOFS1, MOFS2, and MOFS3).

Cohort (Multivariate analysis included clinical traits)	Subtype	HR	HR.95H	HR.95L	Pvalue
In-house (Adjust by Age+KPS+Radiotherapy+Chemotherapy)	MOFS1	0.4484	0.8068	0.2492	0.0074
In-house (Adjust by Age+KPS+Radiotherapy+Chemotherapy)	MOFS2	2.1177	3.7057	1.2102	0.0086
In-house (Adjust by Age+KPS+Radiotherapy+Chemotherapy)	MOFS3	1.1694	1.9382	0.7055	0.5439
GSE72951 (Adjust by MGMT methylation)	MOFS1	0.3289	0.8562	0.1263	0.0227
GSE72951 (Adjust by MGMT methylation)	MOFS2	2.7194	6.8778	1.0752	0.0346
GSE72951 (Adjust by MGMT methylation)	MOFS3	1.1565	2.3235	0.5756	0.6829
CGGA_RNAseq (Adjust by Age and MGMT methylation)	MOFS1	0.4725	0.7155	0.3120	0.0004
CGGA_RNAseq (Adjust by Age and MGMT methylation)	MOFS2	2.0095	3.2195	1.2543	0.0037
CGGA_RNAseq (Adjust by Age and MGMT methylation)	MOFS3	1.2945	1.8883	0.8874	0.1803
CGGA_RNAseq (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS1	0.4468	0.6891	0.2897	0.0003
CGGA_RNAseq (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS2	2.2507	3.6453	1.3896	0.0010
CGGA_RNAseq (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS3	1.2551	1.8639	0.8451	0.2602
CGGA_Array (Adjust by Age and MGMT methylation)	MOFS1	0.3581	0.6976	0.1839	0.0025
CGGA_Array (Adjust by Age and MGMT methylation)	MOFS2	2.2076	4.2178	1.1555	0.0165
CGGA_Array (Adjust by Age and MGMT methylation)	MOFS3	1.2986	2.2271	0.7572	0.3425
CGGA_Array (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS1	0.3993	0.8174	0.1951	0.0120
CGGA_Array (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS2	2.3784	4.6650	1.2127	0.0117
CGGA_Array (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS3	1.0844	1.9561	0.6012	0.7878

With regard to the comparison of the MOFS system to other classification systems: The authors concluded that there was little overlap between their MOFS classes and the system promoted by Wang et al. Examination of Figure 3a shows that MOFS1 and MOFS2 are stable and correspond to Proneural and mesenchymal classes, respectively, but MOFS2 is more heterogeneous. The conclusion that there is little overlap is a bit of an overstatement. In this regard, it is important to acknowledge other studies that used snRNAseq or scRNAseq to classify GBM states based on cellular composition – e.g., Al-Dalahmah et al 2023 (10.1038/s41467-023-38186-1) where the authors described an analogous GBM tissue states system that is prognostic and captures the essence of the MOFS classes.

Response: We sincerely thank you for your thoughtful feedback and for pointing out the need for clarification regarding the overlap between the MOFS system and other classification systems. Upon careful review, we recognize that our initial phrasing of "little overlap" was imprecise and did not fully capture the nuanced relationship between MOFS subtypes and the Wang et al. classification. Specifically, MOFS1 aligns

closely with the Proneural class, and MOFS2 partially overlaps with the Mesenchymal class, though MOFS2 exhibits greater heterogeneity (as we elaborated in our response to your first comment). We have revised the manuscript to explicitly acknowledge these partial overlaps.

We also sincerely appreciate you bringing attention to the study by Al-Dalahmah et al. (2023), which provides an important perspective on GBM heterogeneity through snRNAseq and scRNAseq analyses. Your reference to this study is highly relevant, as their description of prognostic GBM tissue states based on cellular composition captures the essence of the MOFS framework and highlights the potential convergence between independent methodologies. We agree that this study underscores the importance of cellular composition in defining tumor states, which may complement our multimodal approach. In the revised manuscript, we have incorporated a brief mention of their work in the Discussion (Limitations) section, noting that while MOFS integrates multi-omics data, other studies using single-cell RNA sequencing have also provided valuable insights into GBM states and their prognostic relevance. A deeper integration of single-cell and multimodal frameworks may be an avenue for future research.

The section on predicting IC50 in clinical samples based on results from a published GBM cell line drug-treatment database is vague and lacks validation. This section tries to predict which MOFS classes would respond to a specific drug; however, assigning MOFS classes to cell lines as a way to provide some validation is problematic because MOFS3 is defined as the TME, which is lacking in vitro. MOFS1 is defined by the presence of normal brain cells (Neurons, astrocytes, oligodendrocytes), which is also lacking in vitro. Also, the methods do not provide sufficient details on how the analysis was done. This section is confusing, and I recommend that it should be removed from the manuscript.

Response: We sincerely thank you for your thoughtful comments and for highlighting the limitations in the section on predicting IC50 in clinical samples based on the GBM cell line drug-treatment database. Upon careful consideration of your feedback, we agree that this section has significant limitations, particularly due to the lack of tumor microenvironment (TME) components and normal brain cells (neurons, astrocytes, oligodendrocytes) in vitro, which are key features of MOFS3 and MOFS1, respectively. Additionally, we recognize that the methods provided for this analysis lacked sufficient detail, which may have contributed to the confusion.

In light of these issues, we have decided to remove this section from the manuscript. We would like to emphasize that the removal of this content does not affect the primary conclusions of our study. The focus of our work remains on the robust integration of multimodal data to define MOFS subtypes and their prognostic significance, which are well-supported by the analyses presented in other sections of the manuscript.

We appreciate your insightful feedback, which has helped us refine the manuscript and ensure that it maintains clarity and scientific rigor.

Finally, in the discussion, the authors should acknowledge the limitations of the current model in assigning MOFS classes to different parts of the tumor examined on imaging. For example, it is very common to have a large non-enhancing component of the tumor (MOFS1) and a small component of nodular enhancement (MOFS2). The way the data is presented, the entire tumor receives a MOFS2 class – if I understood their model correctly. It is possible that refining the system spatially may improve its predictive power. This is especially important in the post-treatment recurrence setting. The authors should also include a statement on the challenge of distinguishing treatment effect from actual tumor in the post-treatment recurrence setting and comment on the potential utility of this system in advancing the solution for this age-old question.

Response: We sincerely thank you for your insightful comments and for pointing out an important limitation of our current model in assigning MOFS classes to different tumor regions observed on imaging. You are correct that, as presented, our model assigns a single MOFS class to the entire tumor, which does not account for the spatial heterogeneity often observed in GBM. For instance, a tumor with a large non-enhancing component (MOFS1) and a smaller nodular enhancement (MOFS2) would currently be classified as MOFS2, based on the dominant imaging features. We agree that this approach oversimplifies the complexity of GBM and may limit the predictive power of the model. In the revised manuscript, we have acknowledged this limitation in the Discussion section.

Furthermore, we recognize the potential for improving the MOFS system by incorporating spatially resolved analyses to capture intra-tumoral heterogeneity. Such refinements could enhance the accuracy of subtype assignments, particularly in the post-treatment recurrence setting, where tumor biology and therapeutic responses are often heterogeneous. Incorporating spatial information could also provide valuable insights into how different regions of the tumor respond to therapy and how this impacts patient outcomes.

Additionally, we have added a statement addressing the challenge of distinguishing treatment effects from true tumor recurrence in the post-treatment setting. We agree that this is a long-standing issue in the field and acknowledge that the current MOFS system does not yet address this challenge directly. However, we believe that future iterations of the MOFS framework, particularly those incorporating multimodal spatial data, have the potential to contribute to solving this critical question by providing more precise characterization of tumor biology and treatment effects.

We appreciate your suggestions, which have helped us refine the manuscript and address key limitations of our approach.

Minor points:

1-Are the features that were retained by the MOFS interpretable? Were all features equally important?

Response: Thank you for your thoughtful questions. We appreciate the opportunity

to clarify the interpretability of the retained features and their relative importance. For the gene-based modalities, including transcriptomics and proteomics, the retained features are directly interpretable as they correspond to specific genes or proteins. Many of these features are linked to well-defined biological functions and pathways, as evidenced by our functional enrichment analyses, which further underscored their relevance in defining MOFS subtypes. These analyses provide insights into the biological mechanisms driving the molecular heterogeneity within GBM and the distinct characteristics of each MOFS subtype.

In contrast, for histopathological and radiological modalities, the interpretability of features is less straightforward due to the abstract nature of the extracted parameters. These features, derived from image texture, shape, intensity, and higher-order statistical properties, encapsulate microstructural or radiological patterns that might correlate with tumor heterogeneity, immune infiltration, or vascular characteristics. For example, radiomics features such as texture entropy or wavelet-based measures have been linked to variations in tumor cellularity, necrosis, and angiogenesis. Similarly, in pathology, extracted features like cell density or morphological complexity reflect tissue architecture, which aligns with the underlying tumor biology. While the specific biological interpretation of these image-derived features may require further investigation, their inclusion enhances the multimodal framework's capacity to capture complex tumor phenotypes.

Regarding the relative importance of features, the MOFS framework utilized the median absolute deviation (MAD) to identify highly variable features within each modality, ensuring that the most informative features were retained. However, as the primary focus of clustering is to group samples based on similarity rather than to assess feature importance, the clustering process does not explicitly rank features by their contribution. The feature selection step, driven by MAD, prioritized highly variable features likely to differentiate subtypes, ensuring that each modality contributed relevant and complementary information to the final MOFS framework.

2-The WSI's were scanned at 5x, and then the images were segmented with cell profiler. It is unclear if the WSI's were done at 5x (low resolution) or at 20x (adequate resolution) – both are listed in the methods.

Response: Below is a detailed workflow for processing the WSI. Pathology slides were scanned at 20x magnification using a digital pathology scanner (KF-PRO-120-HI) to obtain the original whole slide images (WSI). The WSI at the 5x resolution was converted from RGB to Lab color space, and Otsu's algorithm was then applied to calculate a segmentation threshold, which was used to segment the tissue from the WSI. The obtained tissue image was tiled into many 1024×1024 patches at 20x magnifications, where these patches were adjacent to one another covering the WSI. A Python package Yottixel (<https://doi.org/10.1016/j.media.2020.101757>) was used to select the optimal patches for further analysis. Finally, CellProfiler software was used to extract features from each selected patch. We have described this workflow in the Methods section.

3-The MRI images and WSI with corresponding features for each WSI/MRI image are not provided. The authors should make this dataset accessible.

Response: We appreciate your interest in accessing the dataset. However, due to patient privacy concerns according to our institution's regulation, we are temporarily unable to make the original WSI/MRI images publicly available without controlled access. Nevertheless, we can provide all the extracted feature data derived from the original WSI/MRI images. The relevant feature information is available in Table S1, within the WSI/Radiomics sheet.

4-The IHC in Figure 4g and S10 are not compelling. Additional validation studies of the antibody used are needed. There is barely any signal in the MOFS2 images. The authors should show positive controls and negative controls. Based on that, the interpretation of panel 4J is difficult. The IHC panels in the S10 are low resolution and cannot be assessed.

Response: We sincerely thank you for pointing out the issues with the IHC panels in Figure 4g and Supplementary Figures S10. Upon review, we realized that the selection of certain IHC images may not have been optimal, which could have affected the clarity of the results. To address this, we have replaced the IHC panels in Figure 4g with higher-quality images that better represent the staining patterns observed in our experiments.

For Supplementary Figures S10A, we acknowledge that the IHC images included in the initial submission were low resolution, making them difficult to assess. Given the large number of samples, we have now included the original quantitative data corresponding to these images in Supplementary Table S8, providing a clearer representation of the staining results for these experiments. This additional information ensures that the data are accessible for evaluation and interpretation.

5- same point for S100A4 IHC (Fig 5N and S11D)– the panels are difficult to interpret. The images are low resolution, and the signal-to-background ratio is low.

Response: We sincerely thank you for your feedback regarding the S100A4 IHC panels in Figure 5n and Supplementary Figure S11D. We acknowledge that the resolution of the images in the initial submission may have limited their interpretability. To address this, we have now provided the original quantitative matrix corresponding to the S100A4 IHC data in the revised manuscript. This information has been included in Supplementary Table S8 to ensure clarity and facilitate a more detailed evaluation of the results.

6- Figure 5d shows that the authors are inferring the presence of B-cells and mast cells in GBM tumors – this is surprising. Neuropathologic assessments of primary untreated GBM sections rarely reveal T-cells, let alone B-cells. Mast cells – I have never seen them in GBM section – and I have looked at more than a thousand glioma sections. What is the evidence that these cells exist? It is not clear from the text or methods how this panel was generated. These details must be provided.

Response: Thank you for your valuable comments regarding the cell infiltration in

GBM tumors. We appreciate the opportunity to clarify the underlying methodology and address your concerns directly.

The immune cell profiles presented in Figure 5d were generated using the ssGSEA (single-sample Gene Set Enrichment Analysis) algorithm, as described in Barbie et al. (Nature, 2009, DOI: 10.1038/nature08460). The ssGSEA measures the relative enrichment of predefined gene sets in individual samples. Specifically, we utilized immune-related gene signatures to compute enrichment scores that reflect the relative activity of various cell types, such as T cells, B cells, and mast cells, across samples. Importantly, these scores represent relative abundance and are normalized across all samples within the dataset, rather than reflecting absolute cell counts or histological observations.

This approach aligns with methodologies used in other studies, such as the work by White et al. (2023, Annals of Oncology, PMID: 36494005), which applied similar transcriptomic profiling techniques to infer immune cell populations (including T, B, NK, Monocyte, Myeloid DC, Neutrophils, Endothelial, Microglia) and define tumor microenvironment (TME)-based GBM subtypes (TME-low, TME-medium, and TME-high). In the revised manuscript, we have expanded on these comparisons, illustrating how the MOFS subtypes correspond to these inferred TME-based subtypes (e.g., new **Figures 3c and Figure S9A**).

It is well-documented that primary GBM tissue can express transcriptional signatures of immune cell populations, including T cells and B cells, as shown in recent tumor immune landscape studies (PMID: 38127790, 38242087, 39085419, 39576855). While B cells and T cells are rarely observed histologically, their transcriptional signatures may still be detectable through algorithms like ssGSEA, which rely on the expression of specific marker genes. This is especially relevant in MOFS3, where immune infiltration is relatively higher compared to other subtypes. However, these results should be interpreted as computational inferences rather than direct evidence of cell presence in histological sections.

Regarding mast cells, we acknowledge their rarity in GBM tissue. To address your concern, we have removed mast cell-related analyses from the manuscript and replaced them with endothelial cells and pericytes, as recommended by other reviewers (**Figure S13B and Figure 5d**).

Finally, we have clarified in the revised manuscript that ssGSEA-based enrichment scores are relative indicators of gene signature activity across samples. We believe this clarification will help readers better understand the strengths and limitations of the method employed in our study. Thank you again for raising these important points and helping us improve the clarity and rigor of our work.

7- The paragraph starting with “Specific Drug Candidates for MOFS Subtypes” should indicate that the GBM cell line data is published by another group.

Response: Thank you for your suggestion. As per your request, we have removed this section from the revised manuscript.

8- Figure 6c – survival analysis must control for MGMT methylation status.

Response: Thank you for your valuable reminder. We have now incorporated survival analyses that control for MGMT methylation status in the revised manuscript (Supplementary Table S5A).

9- the term CAF is not defined

Response: Thank you for pointing this out. We have revised in the new manuscript.

Reviewer #3 (Remarks to the Author):

A postdoctoral researcher in my laboratory and myself were invited to co-review this manuscript for Nature Communications.

Overall, the paper addresses a critical problem in glioblastoma translational and clinical research related to glioblastoma classification (or subtyping), which is essential to study the underlying biology of these highly heterogeneous tumours and for stratification of patients for therapies better tailored to different tumour subtypes. A better classification of glioblastoma tumours with prognostic value is hence urgently needed.

Several attempts have been made to classify these tumours into different subtypes based on RNA expression (e.g. Philips PMID: 16530701; Wang PMID: 28697342), but the predictive prognostic value of such classification is highly debatable and not entirely consistent across different cohorts of patients, suggesting at some extent, these classifications are not robust and unstable.

Heterogeneity, however, does not occur only at the RNA level but is also reflected in gene mutations, chromosomal gain/loss, gene and protein expression profiles, tumour location, histopathological features, different transcriptional tumour cell states and non-malignant cell composition in the tumour microenvironment. Yet, most classification methods only focus on one modality (e.g., RNAseq, MRI, or histopathology). The use of single modalities prevents cross-association of different features of individual tumours, producing non-optimal classification that lacks prognosis value. This work addresses this issue and implements multimodal analysis for the classification of wild-type IDH glioblastoma based on multimodal data.

We found the work quite interesting and well-designed. Moreover, integrating multimodal data and developing a deep learning classifier for non-invasive subtype prediction is a significant step forward in glioma research. However, we think these are some concerns the Authors need to address to maximize the reproducibility, impact and scope of the results presented, which we believe can significantly contribute to the Neurooncology community in particular and broad field of cancer research and AI applications in cancer.

Response: We are grateful for the reviewer's positive evaluation of our work and the

valuable feedback. Regarding the lack of detail in the data analysis pipeline, we have taken significant steps to enhance reproducibility.

The MOFSR package is designed to overcome these limitations. It not only offers a standardized and transparent way to perform the analysis but also provides flexibility for users to explore different aspects of the data. By integrating multiple modalities and applying advanced algorithms, MOFSR enables accurate subtype identification and prediction. Additionally, we are committed to continuously improving the package's documentation and providing more detailed explanations of the methods used. This will ensure that other researchers can fully understand and reproduce our results, further validating the robustness of the MOFS framework in future studies. Additionally, we have also composed a detailed user manual for MOFSR (79 pages, <https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR%20User%20Manual.pdf>), which serves as a comprehensive guide, elucidating each function and its application scenarios with precision and clarity. This manual is designed to assist users in navigating the package with ease and confidence, ensuring that they can make the most of MOFSR's capabilities.

1. Data Integration and Clustering

- The package offers a wide range of clustering algorithms for integrating different modalities of data, such as RNA, protein, genomic data, and more. These algorithms, including Consensus Iterative Multi-view Learning (CIMLR), Consensus Principal Component Analysis (CPCA), Integrative Non-negative Matrix Factorization (IntNMF), Low-Rank Approximation Clustering (LRAcluster), Multiple Co-Inertia Analysis (MCIA), Multimodality Fusion Subtyping (MOFS), NEMO, PINSPlus, Regularized Generalized Canonical Correlation Analysis (RGCCA), Sparse Generalized Canonical Correlation Analysis (SGCCA), and Similarity Network Fusion (SNF), help identify shared patterns and structures in the data, thereby discovering potential disease subtypes or biological mechanisms. For example, the `RunCIMLR` function can be used to perform clustering analysis using CIMLR, which is useful for discovering shared patterns across multiple datasets and identifying distinct subtypes.
- It allows users to explore the optimal number of clusters in a multimodal dataset. By using functions like `Find.OptClusterFeatures`, which utilizes Nonnegative Matrix Factorization (NMF) and Multi-block Principal Component Analysis (mbPCA), users can identify the most suitable combination of features and the optimal number of clusters. This function calculates metrics such as the Cluster Prediction Index (CPI) and GAP statistic scores to evaluate the quality of clustering solutions.

2. Cluster Quality Assessment

- MOFSR provides several functions to evaluate the quality of clustering results. The `CalCHI` function calculates the Calinski-Harabasz index for hierarchical clustering, which assesses cluster quality by comparing between-cluster and within-cluster dispersion. Higher values of this index suggest better-defined clusters.

- The ``CalPAC`` function calculates the Proportion of Ambiguous Clustering (PAC) for consensus clustering. A lower PAC value indicates a more stable clustering solution, helping users determine the reliability of their clustering results.
- Additionally, the package can perform Principal Component Analysis (PCA) using the ``RunPCA`` function. PCA helps visualize sample relationships in reduced dimensions, revealing the data structure and patterns, which can be used to further evaluate the clustering results.

3. Classification Prediction

- The package includes a variety of classification algorithms that can predict cluster assignments for test data based on trained models and cluster markers. These classifiers, such as AdaBoost, Decision Tree (DT), Elastic Net (Enet), Enrichment-based Neural Network (Enrichment), Gradient Boosted Decision Trees (GBDT), k-Nearest Neighbors (kNN), LASSO, Linear Discriminant Analysis (LDA), Naive Bayes (NBayes), Neural Network (NNet), PCA-based Neural Network (PCA), Random Forest (RF), Ridge Regression (Ridge), Stepwise Logistic Regression (StepLR), Support Vector Machine (SVM), and XGBoost, offer flexibility in choosing the most appropriate method for different datasets and research questions.
- The ``RunEnsemble`` function runs an ensemble of different classification models, considering the consensus among the models. It can also perform survival analysis to filter models based on trends in clinical outcomes, providing more comprehensive and accurate prediction results.

4. Functional Enrichment Analysis

- MOFSR enables functional enrichment analysis to gain insights into the biological significance of the identified subtypes. The ``RunGSVA`` function estimates gene-set enrichment scores across samples using methods like "gsva" for detecting subtle pathway activity changes, "ssgsea" for individual sample analysis, "zscore", and "plage". This helps in analyzing differentially enriched gene sets between groups and understanding the perturbed biological functions and pathways.
- The ``ssMwwGST`` function performs single-sample pathway activity analysis via the MWW-GST method. It calculates gene means and standard deviations, normalizes expression, assesses pathway enrichment, and corrects p-values. The results of this analysis help prioritize biologically relevant pathways, aiding in understanding the underlying mechanisms and functions related to phenotypes or conditions.

5. Feature Selection

- The package provides functions for feature selection and evaluation, which are crucial for handling high-dimensional multimodal data. The ``Select.Features`` function allows users to select hypervariable features based on specified variance calculation methods, such as standard deviation (SD), median absolute

deviation (MAD), or coefficient of variation (CV). This helps in reducing the dimensionality of the data while retaining the most relevant features for further analysis.

6. Clustering of Clustering Analysis

- MOFSR supports clustering of clustering analysis, which combines the results of multiple clustering algorithms to obtain a more robust and reliable clustering solution. The `RunCOCA` function performs clustering of clustering analysis, and the package also provides functions to calculate consensus matrices and visualize the clustering results, such as plotting silhouette scores and using PCA for visualization. These visualizations help users understand the clustering structure and the relationships between samples in different clusters.

7. Reproducibility and Ease of Use

- By providing a set of well-documented functions and a clear framework for data analysis, MOFSR ensures the reproducibility of the results. Users can follow the same analysis steps and use the provided functions to obtain consistent results, which is essential for scientific research.
- The package is designed to be user-friendly, with detailed documentation and examples for each function. This allows researchers with different levels of programming experience to use the package effectively and apply it to their own multimodal data analysis projects.

In summary, the MOFSR package is a powerful tool for multimodal data analysis, offering a wide range of functions for data integration, clustering, classification, functional enrichment, feature selection, and visualization. It addresses the need for reproducibility and usability in the analysis of complex multimodal biological data, facilitating the discovery of disease subtypes and biological mechanisms, and ultimately contributing to the advancement of precision medicine and related research fields.

Main comments:

1. Multimodal fusion model (MOFS model). The methods section mentions that to create this model, "top-n" variables were selected from each layer (radiology, WSI, transcriptomics, and proteomics). 2640 variable combinations were then used for unsupervised cluster analysis. Can the authors include/mention/describe the variables used for this and include this as Supplemental material? I found this is critical as the model can be biased towards one of the specific layers (e.g. transcriptomics or proteomics).

Response: Thank you for your insightful comment. We appreciate your concern regarding the potential bias in the model towards specific layers such as transcriptomics or proteomics. To address this, we employed a feature selection process that involved calculating the median absolute deviation (MAD) for each modality layer. We then selected the top-n variables from each layer, and combined

them into 2640 variable combinations. The cluster prediction index (CPI) and GAP statistic were used to determine the optimal number of clusters and the combination of features that provided the best clustering performance. We recognize the importance of transparency in this process and will include the list of selected variables in Supplementary Material (Table S3), allowing readers to fully assess the choices made for each modality and the resulting variable combinations used in the unsupervised clustering analysis. Additionally, to further support reproducibility, we will upload the code for the feature combination selection process to our GitHub repository (<https://github.com/Zaoqu-Liu/MOFS>). Moreover, we have developed an R package, *MOFSR*, which implements the MOFS framework. The *Find.OptClusterFeatures* within the package performs this process (<https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/Find.OptClusterFeatures.R>).

Regarding your concern about potential bias, we would like to emphasize that our intermediate fusion approach mitigates this risk (ref 1-5). Unlike early fusion, which combines all modalities into a single matrix and may lead to imbalance due to differences in data scale or dimensionality, intermediate fusion integrates the data while preserving the unique characteristics of each layer. This approach ensures that no single modality, such as transcriptomics or proteomics, dominates the clustering process, and it allows the model to capture the interdependencies between modalities in a balanced manner. Early fusion can suffer from the "curse of dimensionality" and be influenced by the imbalance between data layers, which may adversely affect the downstream analyses. In contrast, the intermediate fusion approach provides a more robust and biologically meaningful integration of multimodal data (ref 1-5). We hope this explanation clarifies the methodology and addresses your concern. Thank you again for your helpful feedback.

References:

1. Lipkova J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. 2022;40(10):1095-1110. doi:10.1016/j.ccell.2022.09.012
2. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer*. 2022;22(2):114-126. doi:10.1038/s41568-021-00408-3
3. Ritchie MD, Holinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. 2015;16(2):85-97. doi:10.1038/nrg3868
4. Tini G, Marchetti L, Priami C, Scott-Boyer MP. Multi-omics integration-a comparison of unsupervised clustering methodologies. *Brief Bioinform*. 2019;20(4):1269-1279. doi:10.1093/bib/bbx167
5. Lobato-Delgado B, Priego-Torres B, Sanchez-Morillo D. Combining Molecular, Imaging, and Clinical Data Analysis for Predicting Cancer Prognosis. *Cancers (Basel)*. 2022;14(13):3215. Published 2022 Jun 30. doi:10.3390/cancers14133215

2. *Multimodal fusion model (MOFS model). The imbalance between subtypes, particularly MOFS3, could bias model performance. Did the authors take this into account and could describe whether techniques such as oversampling or class weighting were also used?*

Response: We sincerely thank the reviewer for raising this important concern regarding the potential impact of subtype imbalance. To address this, we have incorporated measures to ensure the robustness and reliability of our clustering and classification approaches.

First, during the unsupervised clustering process, we employed a resampling strategy as part of the consensus clustering framework, which has now been explicitly described in the revised manuscript. Specifically, 70% of the samples were randomly selected in each iteration, and this process was repeated 10,000 times to generate a consensus matrix. To further support reproducibility, we will upload the code for clustering of clustering analysis to our GitHub repository (<https://github.com/Zaoqu-Liu/MOFS>). Moreover, we have developed an R package, *MOFSR*, which implements the MOFS framework. The *RunCOCA* function within the *MOFSR* package could perform this process (<https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/RunCOCA.R>). This clustering of clustering algorithm is designed to enhance clustering stability and robustness by aggregating results across multiple subsamples of the data, rather than relying on a single clustering outcome. The consensus matrix captures the pairwise similarity of sample clustering assignments across iterations, which is then used to identify stable clusters. This approach ensures that the derived clusters are robust to variability in sample selection. Metrics such as the proportion of ambiguous clustering (PAC) and Calinski-Harabasz index were used to optimize the number of clusters and assess cluster robustness. Then, validation of the MOFS model included a silhouette coefficient analysis to remove samples with low clustering confidence (silhouette coefficient < 0.4). This step ensured that only core samples with high cluster assignment stability were retained, further reducing the impact of potential imbalances and noise in the data. Additionally, to further support reproducibility, we will upload the code for these calculations to our GitHub repository (<https://github.com/Zaoqu-Liu/MOFS>). Moreover, we have developed an R package, *MOFSR*, which implements the MOFS framework. These functions are described below:

- CalCHI: This function calculates the Calinski-Harabasz index for evaluating the quality of clustering solutions. It is used to determine the optimal number of clusters (<https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/CalCHI.R>).
- CalPAC: This function calculates the Proportion of Ambiguous Clustering (PAC) to help evaluate the optimal number of clusters (<https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/CalPAC.R>).
- Silhouette coefficient also is encapsulated in <https://github.com/Zaoqu-Liu/MOFS/> and at <https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/RunMOFS.R>

3. Data Augmentation: No mention of data augmentation (in this multimodal model and also in the MRI model) was made, which could help mitigate overfitting.

Response: We sincerely thank the reviewer for raising the important point regarding data augmentation and its potential to mitigate overfitting. Data augmentation is widely utilized in image-based analyses to diversify training datasets through transformations like rotation, flipping, and scaling. However, in our study, the MRI data were preprocessed into quantitative feature matrices using PyRadiomics (as detailed in the Methods section and Github repository: <https://github.com/Zaoqu-Liu/MOFS/blob/master/PyRadiomics%20parameters.yaml>) rather than being analyzed as raw images. Consequently, traditional data augmentation techniques were not directly applicable to our workflow. Instead, we employed alternative strategies tailored to our multimodal framework to ensure robustness and mitigate overfitting.

For the unsupervised clustering process, we adopted an intermediate fusion strategy, integrating data across multiple modalities simultaneously, followed by a late fusion step to derive the final MOFS subtypes. Specifically, we used a resampling strategy within the clustering of clustering analysis framework, where 70% of the samples were randomly selected during each iteration, repeated 10,000 times, to construct a consensus matrix ([RunCOCA.R](#)). This robust approach minimized the influence of outliers and ensured the stability of clustering results. Additionally, our framework incorporated 11 clustering algorithms based on diverse principles (<https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/RunMOFS.R>). To enhance feature selection, we evaluated combinations of top-n variables based on their median absolute deviation (MAD) across modalities and used the cluster prediction index (CPI) and GAP statistic to identify the optimal variable set for clustering. Metrics such as the proportion of ambiguous clustering (PAC) and Calinski-Harabasz index were used to optimize the number of clusters and assess cluster robustness. Then, validation of the MOFS model included a silhouette coefficient analysis to remove samples with low clustering confidence (silhouette coefficient < 0.4). This step ensured that only core samples with high cluster assignment stability were retained, further reducing the impact of potential imbalances and noise in the data. These steps were instrumental in reducing dimensionality, addressing potential bias from dominant modalities, and improving the generalizability of the clustering framework.

For the supervised neural network model based on MRI-derived features, we employed rigorous feature selection methods, including univariate logistic regression and Lasso regression, to retain only the most informative features. Cross-validation on the FAHZZU1 training set and independent testing on the FAHZZU2 and FAHZZU3 cohorts were conducted to validate model performance.

Furthermore, external validation of the MOFS subtypes was performed using seven independent cohorts, which predominantly contained RNA-seq data due to the limited availability of public multimodal datasets. While the MOFS labels in these cohorts were derived using the RNA-based ensemble classifier, we consistently

observed significant prognostic trends in Kaplan-Meier survival analyses across these external datasets. The reproducibility of the MOFS subtypes' prognostic value, despite using transcriptomic data for classification, underscores the robustness of our framework. These findings validate the clinical relevance of the MOFS classification system and its ability to provide stable prognostic insights in diverse, independent datasets.

By combining intermediate fusion, resampling strategies, and rigorous feature selection with external validation, our methods address overfitting and ensure reliability across multiple cohorts. Thank you for your insightful suggestions, which have allowed us to improve the rigor and clarity of our study.

4. Multimodal fusion model (MOFS model). Can the authors expand on the logic and the algorithms for CPI and GAP calculations?

Response: We thank the reviewer for this thoughtful question regarding the logic and algorithms behind the Cluster Prediction Index (CPI) and GAP statistic in our multimodal fusion model (MOFS). These metrics were crucial in selecting the optimal number of clustering variables and ensuring robust cluster assignments.

The CPI calculation was implemented using the IntNMF R package (PMID: 28459819). CPI quantifies the consistency of clustering results across multiple iterations by assessing how well the features and samples are grouped together in the final clustering solution.

The GAP statistic was calculated using the MOGSA R package (PMID: 31243065). This metric compares the intra-cluster dispersion with that of random uniform distributions, providing a statistical measure of cluster robustness. The GAP statistic is particularly useful for determining the optimal number of clusters by evaluating the separation between clusters in multi-modal data.

To ensure reproducibility, we have made the detailed code for these computations available on our <https://github.com/Zaoqu-Liu/MOFS>.

The CPI, or Cluster Prediction Index, was used to evaluate the predictive stability of clustering features. The logic behind CPI is to quantify how well the selected variables can consistently reproduce cluster membership across multiple resampled subsets of the data. For each variable combination, clustering was performed on random subsamples. The clustering assignments from each subsample were then compared to those from the full dataset using similarity metrics. The average similarity across all resampled subsets was calculated as the CPI value. A higher CPI indicates that the clustering results are robust and that the selected features reliably define the clusters. This approach ensured that the variables chosen for clustering had high predictive power and contributed to the biological stability of the identified subtypes.

The GAP statistic was used to estimate the optimal number of clusters by comparing the within-cluster dispersion of the actual data to that of a randomly generated reference dataset. Specifically, the within-cluster dispersion was measured for different numbers of clusters (k) on both the real dataset and the reference dataset. The GAP statistic was calculated as the difference between the

log-transformed within-cluster dispersion of the reference and the actual datasets. The optimal number of clusters corresponds to the value of k where the GAP statistic reaches its maximum, indicating that the clusters formed in the actual data are significantly more compact than those formed in the random reference data. This ensures that the clustering structure reflects real biological patterns rather than random noise.

In summary, the CPI and GAP statistic collectively provided a robust framework for identifying both the optimal clustering features and the number of clusters in our MOFS model. In response to the reviewer's comment, we have added references to the original descriptions of these algorithms (CPI: 28459819, GAP statistic: 31243065) in the revised manuscript to provide readers with additional context and details.

5. Multimodal fusion model (MOFS model). The authors should show (quantitatively) the differences in performance when using multimodal data vs. single-modal data to classify the different MOFS subtypes. For example, after defining the model final model (MOFS model), use this to label patients' groups as ground truth. Then, assess the performance of radiology variables alone, WSI variables alone, transcriptomics variables alone and proteomics variables alone and compute AUC or confusion matrices to determine/show how the multimodal approach overcomes obvious issues in single modal data.

Response: We appreciate the reviewer's insightful suggestion to quantify the performance differences between multimodal and single-modality approaches. As per the request, we evaluated the classification performance of each individual modality layer and compared it to the integrated multimodal MOFS framework. Additionally, we assessed the impact of excluding individual modality layers from the multimodal model. Instead of using AUC, which is typically applied to binary classifications, we utilized confusion matrices and accuracy metrics for these evaluations.

The left panel of **Ref Table 1** displays the clustering results for single-modality layers. Each modality was independently clustered, and the clusters were matched to MOFS subtypes based on maximum overlap. Among the individual modalities, transcriptomics showed the highest accuracy (67.2%), followed by proteomics (51.7%), mutation (41%), CNV (43.1%), radiology (43.1%), and pathology (43.1%). While transcriptomics outperformed other modalities, the accuracy of single-modality clustering remained lower than that of the multimodal framework. This demonstrates the limited capacity of single-modality approaches to fully capture the complexity and diversity of MOFS subtypes. The relatively higher accuracy of transcriptomics suggests that this modality provides more direct biological insights into subtype distinctions compared to other layers.

In the right panel of **Ref Table 1**, we present the clustering results after excluding individual modality layers from the multimodal model. Even with one modality excluded, the remaining multimodal clusters maintained significantly higher accuracy compared to single-modality clustering. Notably, excluding transcriptomics resulted in the largest drop in accuracy (to 68.1%), indicating the critical role of transcriptomic

data in the MOFS framework. In contrast, excluding mutation data resulted in 90.5% accuracy, suggesting a relatively smaller impact on overall performance. These findings underscore that each modality contributes unique and complementary information to the clustering framework. The exclusion of any single modality weakens the model's performance, but the multimodal approach remains robust and significantly outperforms single-modality models.

Additionally, we observed that the agreement between single-modality clusters and the MOFS subtypes was generally low (**Ref Figure 1A**). Moreover, survival analysis demonstrated limited prognostic significance for single-modality clusters, with none achieving statistical significance in distinguishing survival outcomes (**Ref Figure 1b**). When excluding individual modalities, the remaining multimodal clusters showed a higher concordance with MOFS subtypes (**Ref Figure 2A**). These clusters were also prognostic, achieving statistically significant separation of survival curves (**Ref Figure 2B**). However, their survival discrimination was consistently inferior to the full multimodal MOFS framework (**Ref Figure 3**). We have incorporated these results into the revised manuscript to emphasize the specific value of each modality.

Ref Table 1. Comparison of clustering accuracy for single-modality layers and multimodal frameworks with excluded modalities.

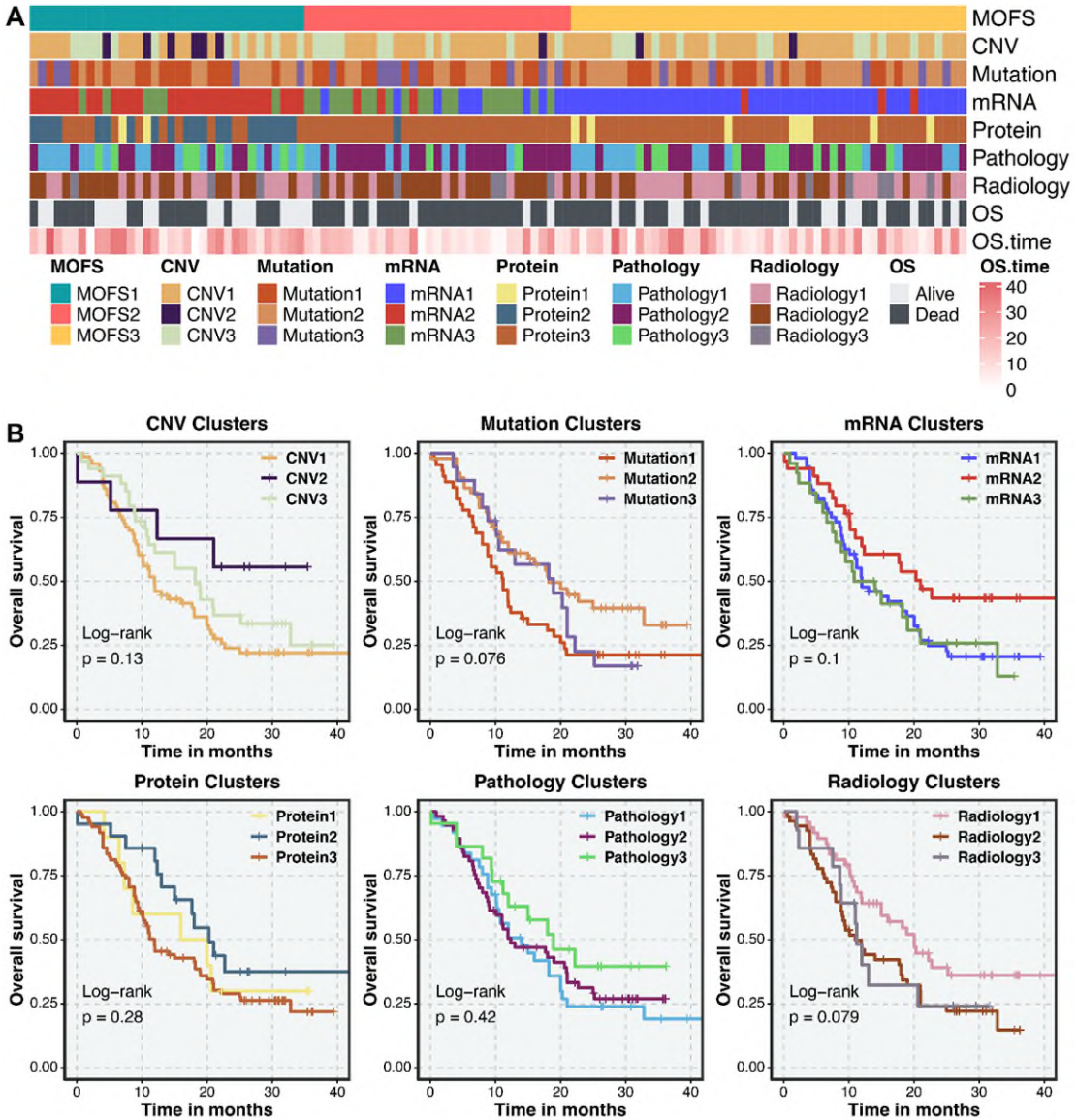
The left table represents clustering results using individual modality layers independently. The right table shows clustering results after systematically excluding individual modalities from the multimodal MOFS framework.

Cluster	Subtypes	MOFS1	MOFS2	MOFS3	Accuracy
mRNA Clusters	mRNA2	20	5	9	0.672414
	mRNA3	8	18	0	
	mRNA1	6	10	40	
Protein Clusters	Protein2	20	1	0	0.517241
	Protein3	12	32	41	
	Protein1	2	0	8	
Mutation Clusters	Mutation3	5	8	6	0.405172
	Mutation1	14	15	16	
	Mutation2	15	10	27	
CNV Clusters	CNV2	6	1	2	0.431034
	CNV3	11	10	13	
	CNV1	17	22	34	
Pathology Clusters	Pathology1	15	8	14	0.431034
	Pathology2	11	23	23	
	Pathology3	8	2	12	
Radiology Clusters	Radiology3	3	5	6	0.431034
	Radiology2	18	20	16	
	Radiology1	13	8	27	

Cluster	Subtypes	MOFS1	MOFS2	MOFS3	Accuracy
MOFS without mRNA	Cluster1	24	1	3	0.681034
	Cluster3	3	24	15	
	Cluster2	7	8	31	
MOFS without Protein	Cluster1	33	3	0	0.853448
	Cluster2	1	28	11	
	Cluster3	0	2	38	
MOFS without Mutation	Cluster1	30	0	0	0.939655
	Cluster2	4	31	1	
	Cluster3	0	2	48	
MOFS without CNV	Cluster1	33	1	0	0.905172
	Cluster2	1	30	7	
	Cluster3	0	2	42	
MOFS without Pathology	Cluster1	29	0	0	0.896552
	Cluster2	5	32	6	
	Cluster3	0	1	43	
MOFS without Radiology	Cluster1	30	0	0	0.853448
	Cluster2	3	24	4	
	Cluster3	1	9	45	

For single-modality clustering, we observed limited prognostic significance for single-modality clusters, with none achieving statistical significance in distinguishing survival outcomes (**Ref Figure 1**). When excluding individual modalities, the remaining multimodal clusters showed a higher concordance with MOFS subtypes (**Ref Figure 2A**). These clusters were also prognostic, achieving statistically significant separation of survival curves (**Ref Figure 2B**). However, their survival discrimination was consistently inferior to the full multimodal MOFS framework (**Ref Figure 3**). This

suggests that each modality contributes unique and complementary information to the clustering framework. In particular, the inclusion of histopathological features enhanced both the prognostic power and the clinical interpretability of the MOFS framework. We have incorporated these results into the revised manuscript to emphasize the specific value of each modality.

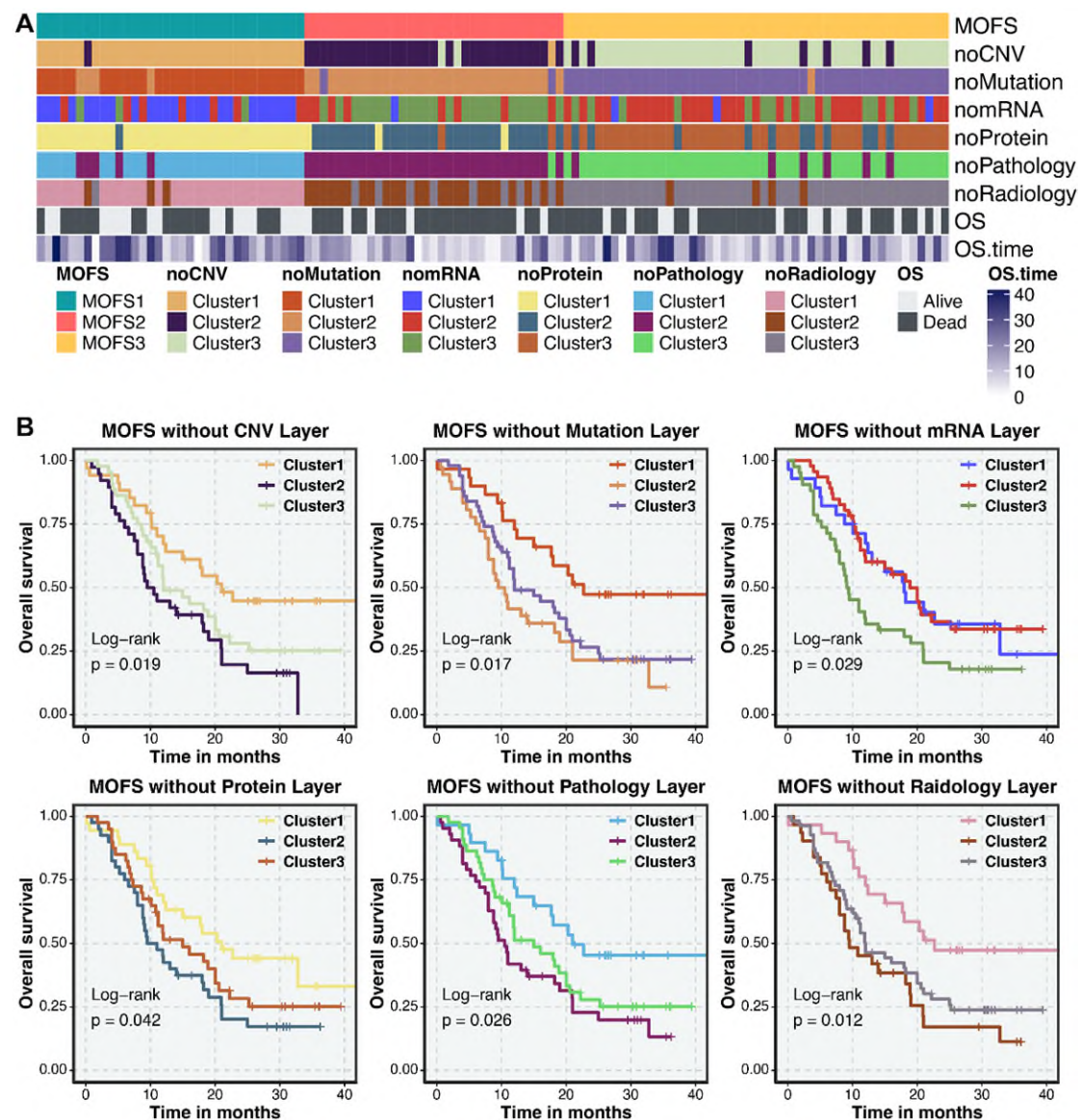


Ref Figure 1. Single-modality clustering and its prognostic significance

(A) Heatmap illustrating the clustering results derived from single modalities, including copy number variation (CNV), mutation, mRNA expression, protein expression, pathology, and radiology. Each row represents the clustering results for a specific modality, and each column represents a patient. Colors indicate the clusters assigned by each modality.

(B) Kaplan-Meier survival curves for clusters derived from individual modalities. The curves represent overall survival outcomes for patients in each cluster. Log-rank p-values indicate the prognostic significance of each single-modality clustering. None of the single-modality clusters achieved statistical significance, highlighting their

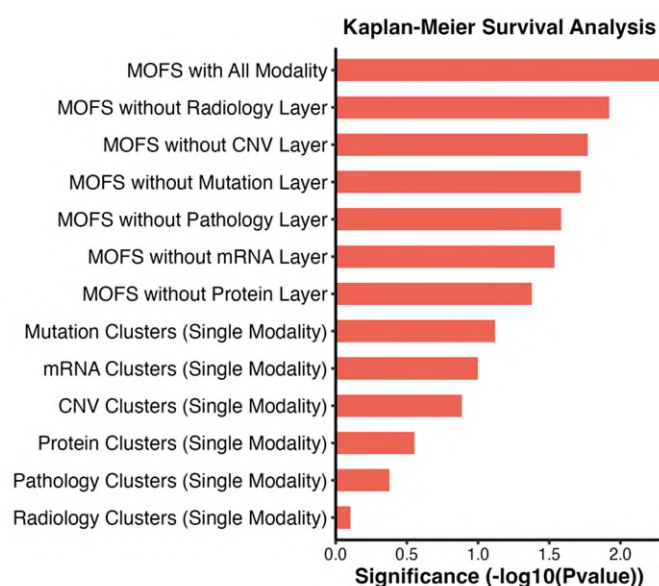
limited ability to distinguish survival outcomes.



Ref Figure 2. Effect of excluding individual modalities on clustering and survival analysis

(A) Heatmap illustrating the clustering results when each modality was excluded from the multimodal MOFS framework. Rows represent clustering results after excluding specific modalities (e.g., noCNV, noMutation, etc.), while the top row represents the original MOFS subtypes for reference. Colors denote patient cluster assignments.

(B) Kaplan-Meier survival curves for clusters obtained by excluding specific modalities. Each panel corresponds to the exclusion of a particular modality (e.g., CNV, mutation, mRNA, etc.), with three clusters shown in each case. Log-rank p -values indicate that the exclusion of any single modality still results in statistically significant survival separation, although the performance is inferior compared to the full multimodal clustering.



Ref Figure 3. Survival significance of multimodal and single-modality clustering.

Bar plot showing the significance of Kaplan-Meier survival analysis for different clustering strategies, represented as $-\log_{10}(\text{p-value})$ from log-rank tests. The top bar represents the MOFS framework incorporating all modalities ("MOFS with All Modality"), achieving the highest prognostic significance. Subsequent bars represent clustering results when individual modalities are excluded ("MOFS without X Layer") and single-modality clusters ("X Clusters"). Single-modality clustering shows limited prognostic power, while multimodal approaches demonstrate significantly enhanced survival discrimination, especially when all modalities are included.

6. Multimodal fusion model (MOFS model). I am puzzled by the authors' failure to consider obvious demographics (sex and age) and MGMT methylation status that are critically relevant to a patient's prognosis (i.e., survival data to the standard of care). In this regard, would it be helpful if the authors could highlight in the PCA plot (or add additional PCA plots) where dots in the graphs are also labelled according to the patient's sex, patient age (higher or lower than the median for GBM), and MGMT methylation status?

Response: We sincerely thank the reviewer for this excellent suggestion. We agree that considering key demographic and molecular factors such as age, sex, and MGMT methylation status is crucial for understanding prognosis and survival outcomes. Rather than focusing solely on PCA visualization, we believe that multivariable Cox proportional hazards regression is a more robust and appropriate method to account for these prognostic variables.

To address this, we performed additional multivariable Cox regression analyses using both our in-house cohort and external validation cohorts. These analyses incorporated essential clinical factors, including age, sex, MGMT methylation status, radiotherapy, and chemotherapy, to assess the independent prognostic value of the MOFS subtypes. The results, summarized in Supplementary Table S5A (**Ref Table 2**), reveal that after adjusting for these variables, MOFS1 remains associated with better

overall survival, MOFS2 continues to predict worse prognosis, and MOFS3 shows no significant association with survival outcomes. These findings are consistent with those presented in Figure 3d of the original manuscript, further validating the robustness and predictive power of the MOFS classification system. The manuscript has been updated accordingly to include these additional results, which strengthen the clinical relevance of our findings.

We greatly appreciate the reviewer's insightful comments, as they allowed us to further refine and substantiate the clinical relevance of our work. These additional analyses and clarifications reinforce the value of the MOFS schema in glioblastoma research and its potential application in clinical practice.

Ref Table 2. Multivariable Cox proportional hazards regression analysis for MOFS subtypes after adjustment for clinical covariates.

This table presents the results of multivariable Cox proportional hazards regression analyses performed in both the in-house cohort and external validation cohorts. The analysis accounts for key clinical variables, including age, MGMT methylation status, radiotherapy, and chemotherapy. Hazard ratios (HR), along with their 95% confidence intervals (HR.95L and HR.95H), are provided for each MOFS subtype (MOFS1, MOFS2, and MOFS3).

Cohort (Multivariate analysis included clinical traits)	Subtype	HR	HR.95H	HR.95L	Pvalue
In-house (Adjust by Age+KPS+Radiotherapy+Chemotherapy)	MOFS1	0.4484	0.8068	0.2492	0.0074
In-house (Adjust by Age+KPS+Radiotherapy+Chemotherapy)	MOFS2	2.1177	3.7057	1.2102	0.0086
In-house (Adjust by Age+KPS+Radiotherapy+Chemotherapy)	MOFS3	1.1694	1.9382	0.7055	0.5439
GSE72951 (Adjust by MGMT methylation)	MOFS1	0.3289	0.8562	0.1263	0.0227
GSE72951 (Adjust by MGMT methylation)	MOFS2	2.7194	6.8778	1.0752	0.0346
GSE72951 (Adjust by MGMT methylation)	MOFS3	1.1565	2.3235	0.5756	0.6829
CGGA_RNAseq (Adjust by Age and MGMT methylation)	MOFS1	0.4725	0.7155	0.3120	0.0004
CGGA_RNAseq (Adjust by Age and MGMT methylation)	MOFS2	2.0095	3.2195	1.2543	0.0037
CGGA_RNAseq (Adjust by Age and MGMT methylation)	MOFS3	1.2945	1.8883	0.8874	0.1803
CGGA_RNAseq (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS1	0.4468	0.6891	0.2897	0.0003
CGGA_RNAseq (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS2	2.2507	3.6453	1.3896	0.0010
CGGA_RNAseq (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS3	1.2551	1.8639	0.8451	0.2602
CGGA_Array (Adjust by Age and MGMT methylation)	MOFS1	0.3581	0.6976	0.1839	0.0025
CGGA_Array (Adjust by Age and MGMT methylation)	MOFS2	2.2076	4.2178	1.1555	0.0165
CGGA_Array (Adjust by Age and MGMT methylation)	MOFS3	1.2986	2.2271	0.7572	0.3425
CGGA_Array (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS1	0.3993	0.8174	0.1951	0.0120
CGGA_Array (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS2	2.3784	4.6650	1.2127	0.0117
CGGA_Array (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS3	1.0844	1.9561	0.6012	0.7878

7. The authors should consider including interpretability tools (e.g., SHAP or Grad-CAM) to increase transparency in clinical predictions associated with MOFS classes, which is crucial for trust in medical AI.

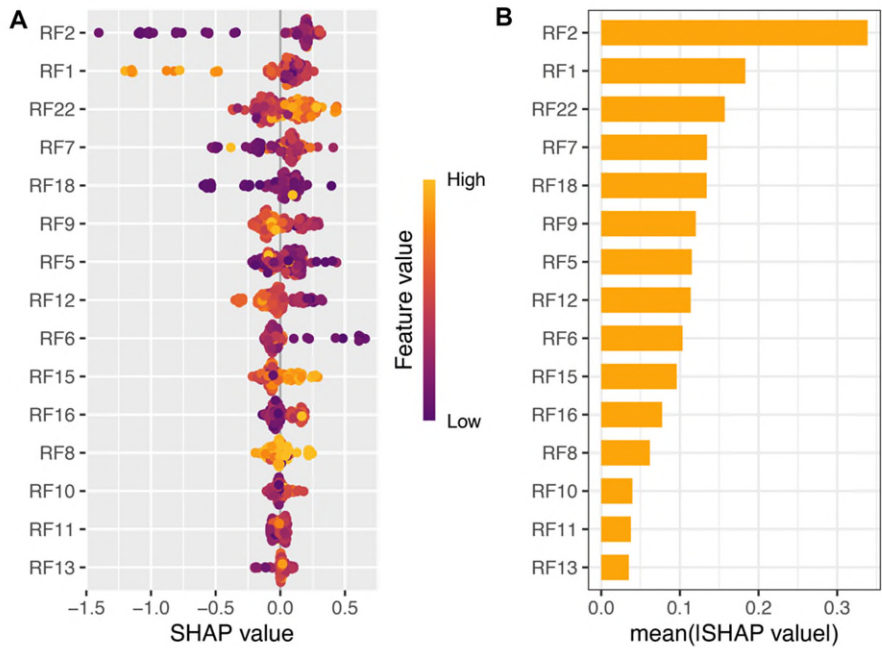
Response: Thank you for your insightful suggestion regarding the use of interpretability tools. We fully agree that enhancing transparency in clinical predictions is essential for building trust in medical AI applications. To address this, we incorporated SHAP (SHapley Additive exPlanations) analyses into our framework to evaluate feature importance and their contributions to MOFS subtype

classification. The results are visualized in the **Ref. Figure 4**.

The SHAP summary plot (left panel) highlights the variable contributions of individual features to the classification, with RF2 (T1C_original_firstorder_InterquartileRange) and RF1 (ADC_wavelet-HLH_glcml_SumEntropy) identified as the most influential features. The color gradient represents the feature values, illustrating how specific imaging-derived metrics positively or negatively impact subtype predictions. The bar plot (right panel) ranks features based on their mean absolute SHAP values, emphasizing the relative importance of each feature across the dataset. Details of the radiomics features (RF1-RF22) are provided in Supplementary Table S10 for further clarity.

These analyses significantly enhance the interpretability of our classification model by offering a detailed understanding of the critical imaging-derived features influencing MOFS subtype predictions. This transparency not only improves the trustworthiness of our model but also aids clinicians and researchers in better understanding the biological and clinical implications of the identified subtypes. We have updated the revised manuscript to include these findings and additional methodological details to ensure clarity.

We hope these additions address your concerns and provide a more transparent foundation for the clinical application of our work.



Ref Figure 4: SHAP Analysis of Feature Contributions in MOFS Subtype Classification.

(A) SHAP summary plot illustrating the contributions of individual imaging-derived features to MOFS subtype classification. Each dot represents a feature's SHAP value for a specific sample, indicating its influence on the prediction. Features are ranked by importance, with RF2 (T1C_original_firstorder_InterquartileRange) showing the highest impact. The color gradient represents feature values, where yellow indicates

high values and purple indicates low values.

(B) Mean absolute SHAP values for each feature, ranked by their overall contribution to the classification model. RF2, RF1, and RF22 are the top three features, emphasizing their critical roles in distinguishing MOFS subtypes.

8. Related to Figure 2a, "Radio-pathological and Biological Peculiarities of MOFS Subtypes", section, Page 8). The authors comment on differences observed at the histopathological and radiological levels across the MOFS groups. Authors should add stats on these observations, either formal (through image analysis) or qualitative (e.g., by stating that they were observed in 10 /12 MOFS1 patients) quantitation. If qualitative results are provided, please include a multipanel with corresponding images. In the images, please also include the patient's demographic information (sex, age, methylation status, if possible).

Response: Thanks very much for your suggestion. We have conducted a manual review and statistical analysis of tumor enhancement patterns across all patients' MRIs. This evaluation was performed collaboratively by a deputy chief physician in neuroradiology (JY) and a deputy chief physician in neurosurgery (ZYZ) both with over 10 years of working experience. If the evaluation results by the two experts differ, they will engage in mutual consultation and discussion to reach a consensus. These findings are summarized in **Supplementary Table S1 and Figure S2E**. Specifically: MOFS1 is characterized by a higher prevalence of no enhancement and limited enhancement types. MOFS2 predominantly exhibits mass-like enhancement. MOFS3 is primarily defined by ring enhancement as the dominant pattern. Representative MRI images are presented in Figure 2a. These results provide a comprehensive overview of the radiological peculiarities within each MOFS subtype, supporting the observed differences in tumor characteristics.

Additionally, two deputy chief physician in neuropathology (WWW and LC) randomly select five patches in each patient's WSI to jointly assess the number of immune cells in each patch. The average immune cell count across these patches was used to represent the patient's immune cell infiltration level. If the evaluations by the two experts differ, they will engage in mutual consultation and discussion to reach a consensus. Statistical analysis of the data was performed, and the results are presented in **Supplementary Table S1 and Figure S2E**. Representative images of pathological patches are presented in Figure 2a. These findings highlight distinct immunological characteristics of the MOFS3 subtype, providing further insights into its tumor microenvironment.

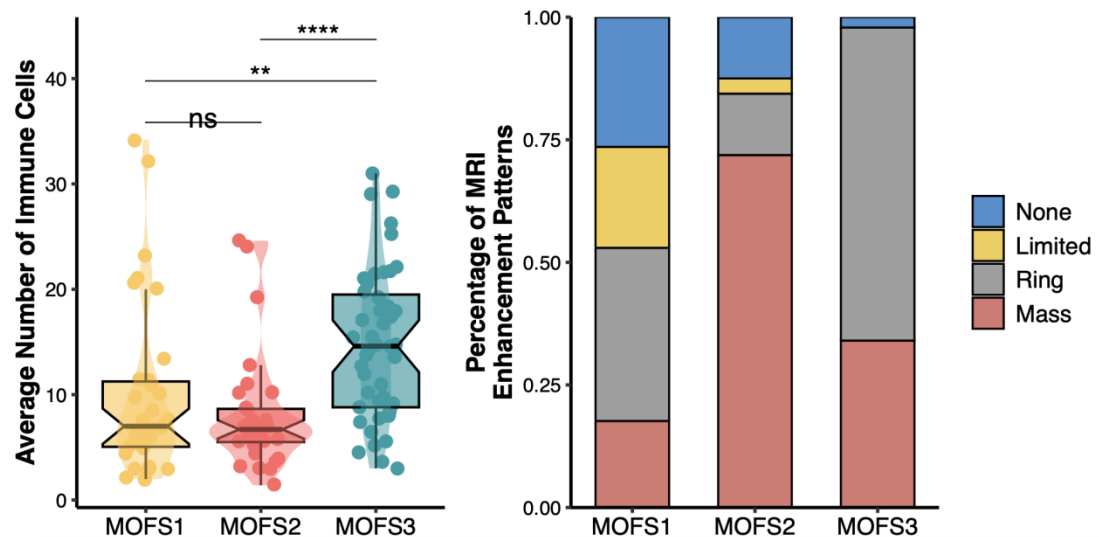


Figure S2E: Box plot showing the average number of immune cells per patch among MOFS subtypes; bar plot illustrating the percentage distribution of different MRI enhancement patterns (none, limited, ring, and mass) among MOFS1, MOFS2, and MOFS3 (**P < 0.01, ****P < 0.0001, ns = not significant).

9. Table S4. The legend does not provide sufficient information to determine whether this table includes RNA or protein data.

Response: Thank you for pointing out the need for greater clarity in the legend of Table S4. In response, we have updated the legend and annotations in the revised manuscript. Specifically, the six sheets in Table S4 are labeled as ssGST-RNA, ORA-RNA, GSEA-RNA, ssGST-Protein, ORA-Protein, and GSEA-Protein, making it clear which results are derived from RNA data and which are derived from protein data.

We hope this revision resolves the issue and enhances the comprehensibility of the supplementary materials. Thank you for highlighting this important detail.

10. Figure 2c is not cited in the paper.

Response: Thank you for pointing this out. We have corrected this oversight, and Figure 2c is now properly cited in the revised manuscript.

11. MOFS ensemble classifier (MOFS classification model based on RNA transcriptomics data only). Related to the methods section.

(i) The authors should provide training results for the method employed.

(ii) The authors should provide a confusion matrix for validation tests (point 6 in the methods). Values reported as percentages are not very informative if the sample size is low. Authors should also split the validation results based on sex, age, and MGMT methylation status to identify bias in the model's performance about patients' demographics. Together, this will determine the consistency and stability of the proposed new classification.

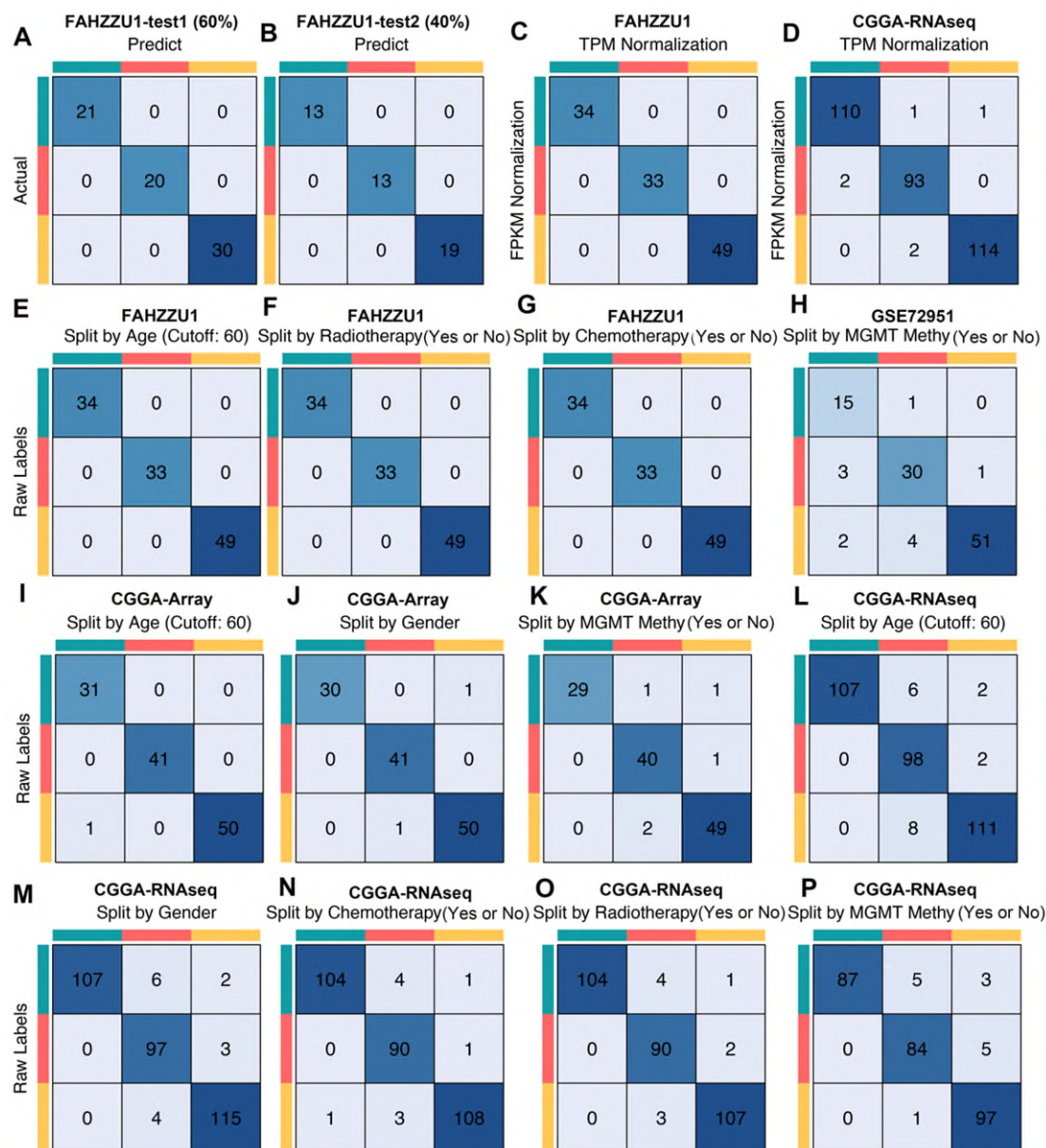
Response: Thank you for your thoughtful suggestion. To address the concerns

regarding validation, transparency, and demographic fairness, we have conducted extensive additional analyses for the MOFS RNA-seq-based ensemble classifier. The results are presented in **Ref Figure 5** and described in detail below. These results aim to provide a more comprehensive evaluation of the classifier's performance and robustness across datasets, normalization schemes, and patient demographics. Additionally, we have uploaded the full code for the classifier and ensemble classifier to our GitHub repository (<https://github.com/Zaoqu-Liu/MOFS>). In the repository, we have organized the code into separate functions for each classifier, making it easy for users to apply these models to their own classification systems or to validate our results. The code is included in the R package *MOFSR*, which can be found at [MOFSR GitHub repository](#).

First, within the FAHZZU1 cohort, we performed a 60:40 random split to create two independent subsets: FAHZZU1-test1 (**Ref Figure 5A**) and FAHZZU1-test2 (**Ref Figure 5B**). The confusion matrices for both subsets demonstrate consistent classification accuracy across the three subtypes (MOFS1, MOFS2, and MOFS3), showing strong agreement between predicted and actual subtype labels. These results highlight the robustness of the classifier to internal variability and confirm the integrity of subtype assignments within our training cohort.

Second, the classifier's compatibility with different RNA-seq quantification methods was assessed. We tested the model using TPM (Transcripts Per Million) and FPKM (Fragments Per Kilobase of transcript per Million mapped reads) normalization schemes across multiple datasets. The confusion matrices for FAHZZU1 (**Ref Figure 5C**) and CGGA-RNAseq (**Ref Figure 5D**) demonstrate that the classifier performs equally well across these normalization methods. This suggests that the model is robust to variability introduced by RNA-seq data processing pipelines, an essential consideration when dealing with public datasets originating from diverse sequencing platforms and protocols.

To evaluate demographic fairness and consistency, we conducted stratified analyses based on clinical factors and treatment-related variables. For the FAHZZU1 cohort, we split the samples by age (cutoff: 60 years, **Ref Figure 5E**), radiotherapy (**Ref Figure 5F**), and chemotherapy (**Ref Figure 5G**). Similarly, for the CGGA cohorts, we stratified patients by MGMT methylation status (**Ref Figure 5K, P**), gender (**Ref Figure 5J, M**), and age (**Ref Figure 5I, L**). Across all stratifications, the confusion matrices reveal balanced performance, with no evidence of bias in classification outcomes.



Ref Figure 5. Confusion matrices for MOFS ensemble classifier performance across multiple cohorts and demographic splits.

(A-B) Performance of the classifier on FAHZZU1-test1 (60% random split) and FAHZZU1-test2 (40% random split) subsets.

(C-D) Confusion matrices for FAHZZU1 and CGGA-RNAseq cohorts, tested with TPM-normalized RNA-seq data, showing the classifier's robustness to different normalization methods.

(E-G) Stratified analysis of FAHZZU1 cohort by age (cutoff: 60 years), radiotherapy (Yes/No), and chemotherapy (Yes/No).

(H) Stratified analysis of the GSE72951 data by MGMT methylation status.

(I-K) Stratified analysis of the CGGA-Array cohort by age (cutoff: 60 years), gender, and MGMT methylation status.

(L-P) Stratified analysis of the CGGA-RNAseq cohort by age (cutoff: 60 years), gender, chemotherapy (Yes/No), radiotherapy (Yes/No), and MGMT methylation status.

Each confusion matrix includes the actual labels (rows) and predicted labels (columns) for MOFS1, MOFS2, and MOFS3 subtypes. Color intensity reflects the proportion of samples in each category, with darker shades indicating higher values.

The primary goal of developing the RNA-seq-based MOFS ensemble classifier was to enable its application in publicly available datasets that lack the comprehensive multimodal data required for our original MOFS framework. Previous subtyping studies, such as those by Wang et al., Philips et al., White K et al. (PMID: 36494005, as you mentioned), and Wu et al. (PMID: 39112531, as you mentioned), which relied on predefined gene sets for enrichment-based scoring, our classifier directly predicts MOFS subtypes using ensemble machine learning. It is important to note that the classifier's primary purpose is not clinical application but rather biological exploration. Glioblastoma datasets with RNA-seq data represent valuable opportunities to explore the molecular underpinnings of MOFS subtypes, even when full multimodal data are unavailable. The classifier allows us to apply the MOFS framework to such datasets, enabling rigorous analyses of subtype-specific features and associations with survival outcomes. This approach builds on the foundations of prior subtyping studies while offering a more systematic and robust method for extending our findings.

In response to the reviewer's comment, we have included these additional analyses and validations in the revised manuscript and supplementary materials. These results address concerns about transparency, consistency, and demographic fairness while emphasizing the classifier's potential for advancing glioblastoma research. Thank you for your valuable insights, which have strengthened the rigor and clarity of our work.

12. Validation of MOFS Subtypes and Comparison with Traditional Classifications. I found it quite informative the authors aim to benchmark this classification approach with previous classifications. However, a new report also provides a new classification based on TME composition with prognostic value (White K et al., Annals of Oncology, 2023 PMID: 36494005) that can overlap with the MOFS classification proposed in this study. It is critical that the authors also benchmark the MOFS model against the TMElow/medium/high classification, repeating the same analysis as in Figure 3a and Figure S6.

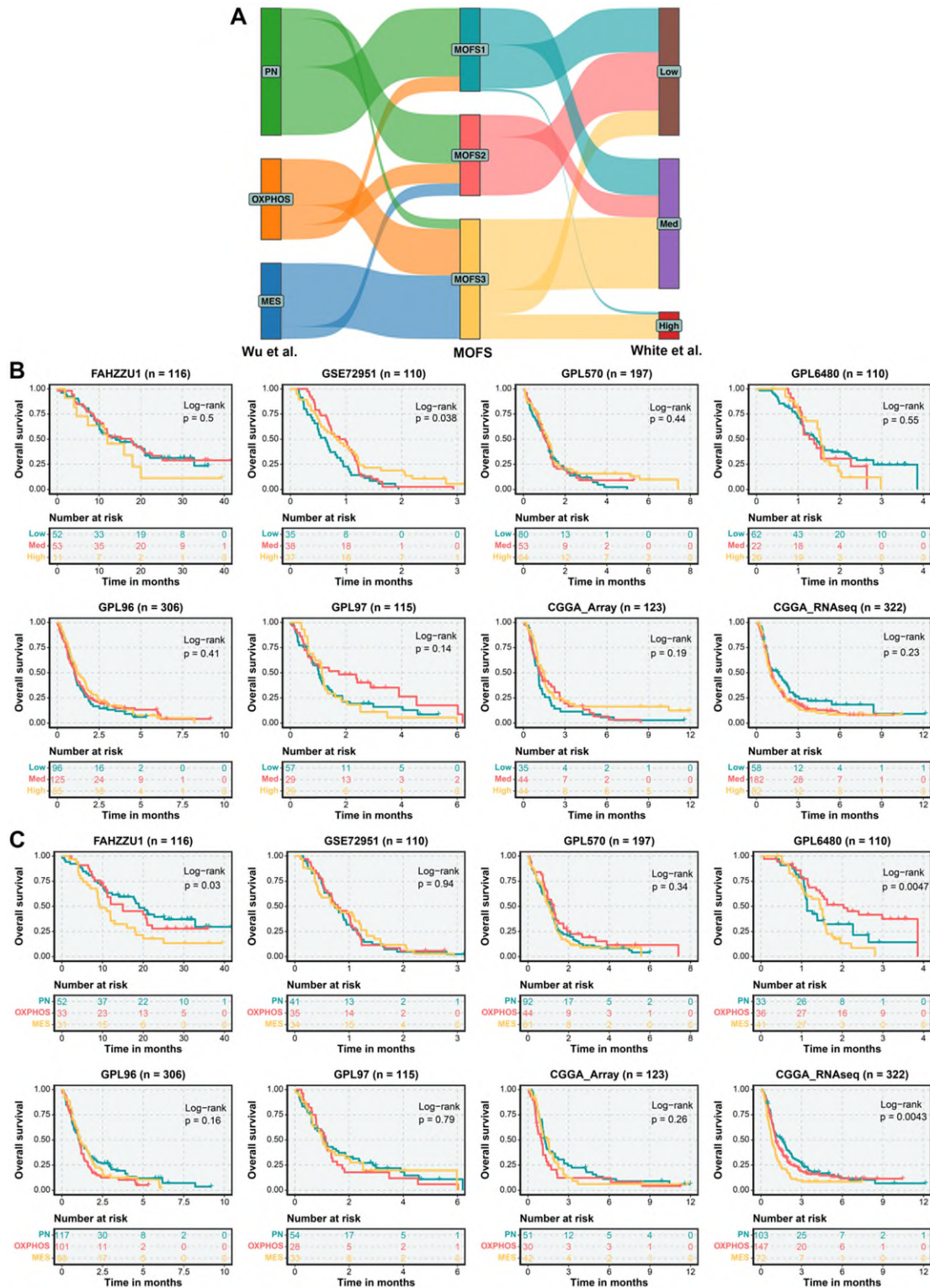
13. Similarly to comment 12, recently, patient-derived cell lines were subgrouped into three groups based on pharmacological response and RNA expression (PMID: 39112531) that can somewhat overlap with the MOFS classification. Again, the authors are encouraged to benchmark the MOFS classification against the model in this work (using RNAseq available from primary tumours).

Response to 12&13: Thank you for your insightful suggestion to benchmark the MOFS classification against the recently proposed TME-based classification (White K et al., Annals of Oncology, 2023 PMID: 36494005) and *patient-derived cell line (PDCL)-based* classification (Wu et al., Nature Communications, 2023 PMID: 39112531). We have conducted the suggested analyses to assess the overlap and

prognostic value of these classification systems in comparison with our MOFS framework.

To benchmark MOFS against the TME-based subtypes (Low, Med, High) and the PDCL-based subtypes (PN, OXPHOS, MES), we followed the methods outlined in the respective studies to identify these subtypes within our datasets. The overlap analysis between MOFS subtypes and these classifications is presented in **Ref Figure 6A**, which shows moderate concordance between the systems.

Additionally, we assessed the prognostic value of the TME and PDCL-based subtypes across external validation cohorts using the same Kaplan-Meier survival analysis framework as in Figures Figure S9 of our revised manuscript. The survival outcomes associated with these alternative classifications are shown in **Ref Figure 6B** and **Ref Figure 6C**, respectively. While the TME and PDCL-based subtypes showed some prognostic relevance in specific datasets, their stratification was less consistent and robust across external cohorts compared to MOFS. These results underscore the superior clinical utility and generalizability of the MOFS framework. Thank you again for your suggestion, which has enhanced the rigor and clarity of our work.



Ref Figure 6. Benchmarking MOFS subtypes against TME-based and PDCL-based classifications.

(A) Sankey diagram comparing MOFS subtypes (MOFS1, MOFS2, and MOFS3) with TME-based subtypes (Low, Med, High) as defined in White K et al. and PDCL-based subtypes (PN, OXPHOS, MES) as described in Wu et al. The flow indicates the overlap in sample classification between the different systems.

(B) Kaplan-Meier survival analyses of TME-based subtypes (Low, Med, High) across

multiple datasets, including FAHZZU1, GSE72951, GPL570, GPL6480, GPL96, GPL97, CGGA-Array, and CGGA-RNAseq. The log-rank p-values indicate the significance of survival differences within each cohort.

(C) Kaplan-Meier survival analyses of PDCL-based subtypes (PN, OXPHOS, MES) in the same datasets as in (B). Survival curves demonstrate the prognostic value of these subtypes within each cohort, with log-rank p-values assessing statistical significance.

14. Figure title and Legends for Figures S6 and S7 are swapped.

Response: Thank you for pointing this out. We have corrected this oversight in the revised manuscript.

15. It was mentioned that the MOFS ensemble classifier has an accuracy of ~90% (methods section, see also comment 8). Figure S7 shows pathway enrichment analysis for MOFS subtypes using this MOFS ensemble classifier. Although the general trend is accepted (matches the results in Figure 2b), many patients do not replicate such a trend. So, it would be suitable for the authors to report what number of patients agree with the results in 2B and how many do not. This also relates to the use of single-modal vs multimodal models.

Response: Thank you for pointing out this important aspect regarding Figure S10 (previously Figure S7). It is essential to clarify that the results in Figure S10 are derived from seven external cohorts where MOFS subtypes were predicted using the MOFS ensemble classifier. These predictions were based on transcriptomic data, and therefore, they cannot be directly compared to the ground truth subtypes obtained through multimodal clustering in the primary cohort (Figure 2b). The results in Figure 2b represent the clustering results from the FAHZZU1 cohort using multimodal data.

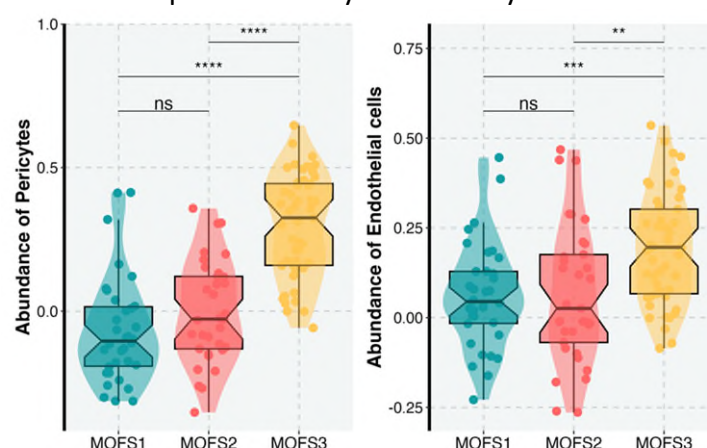
However, the MOFS ensemble classifier has been validated to achieve high accuracy, underscoring its reliability for transcriptomic-based subtype prediction. In the external cohorts, the classifier successfully preserved the biological and prognostic characteristics of the MOFS subtypes. Specifically, the functional enrichment patterns for MOFS1, MOFS2, and MOFS3 (e.g., neurodevelopmental, proliferation, and TME-related pathways, respectively) remain consistent with those identified in Figure 2b. Additionally, the classifier maintained robust prognostic stratification across all external cohorts, further validating its applicability.

We acknowledge that some discrepancies in individual patient classifications may occur due to the reliance on single-modality data in external cohorts, as transcriptomic-only data inherently lack the depth of multimodal integration. Nonetheless, the overarching trends reinforce the robustness of the MOFS framework. Thank you again for your constructive suggestion, which has enhanced the depth and clarity of our analysis.

16. Figure 5. Stromal component. The authors did not include blood vessels in the stromal component analysis (i.e. endothelial cells and pericytes). Figure 5m suggests that this could be quite an essential non-malignant cell component in MOFS3 (e.g. FAP, ACTA2, PDRGFA expression).

Response: Thank you for bringing this to our attention. We appreciate the reviewer's insightful comment regarding the importance of stromal components such as endothelial cells and pericytes. To address this, we conducted additional analyses to evaluate the abundance of endothelial cells and pericytes across the MOFS subtypes. As shown in **Ref Figure 7**, both endothelial cells and pericytes were found to be significantly more abundant in MOFS3 compared to MOFS1 and MOFS2. This finding aligns with the observed stromal-rich nature of MOFS3.

We have incorporated these results into the revised manuscript (Figure 5d and Supplementary Figure S15B) and expanded the results to highlight the significance of vascular components in MOFS3. Thank you again for your constructive suggestion, which has enhanced the depth and clarity of our analysis.



Ref Figure 7. Abundance of pericytes and endothelial cells across MOFS subtypes. Violin plots illustrating the relative abundance of (left) pericytes and (right) endothelial cells in MOFS1 (blue), MOFS2 (red), and MOFS3 (yellow). The median and interquartile range are shown within each plot. Statistical significance was calculated using the Kruskal-Wallis test followed by Dunn's multiple comparisons test. MOFS3 exhibits significantly higher abundance of both pericytes and endothelial cells compared to MOFS1 and MOFS2 ($p < 0.01$, $p < 0.001$, $**p < 0.0001$, ns = not significant). These findings underscore the enriched vascular stromal components in the MOFS3 subtype.

17. Figure 5I_labels in the inset. Should be "high/low-Str" not "high/low-CAFs".

Response: Thank you for pointing this out. We have corrected this oversight in the revised manuscript.

18. Including MGMT methylation status is critical in the context of TMZ sensitivity and Figure 6b.

Response: Thank you for highlighting the importance of including MGMT methylation status in the context of TMZ sensitivity. As requested, we revised the manuscript and incorporated MGMT methylation status into our analysis to ensure a more comprehensive evaluation of TMZ response across MOFS subtypes. While drug response data was removed as per Reviewer 1's suggestion, we retained the analysis of TMZ sensitivity within the clinical outcomes section, now presented in

Supplementary Table S5B.

Our results indicate that, after adjusting for key confounders such as MGMT methylation status, age, and gender, MOFS1 and MOFS3 patients who received TMZ therapy demonstrated significantly better overall survival ($HR < 1$, $P < 0.05$). Conversely, in MOFS2 patients, no survival benefit was observed between those treated with TMZ versus untreated cases ($P = 0.179$) (**Ref Table 3 and new Table S5B**). This finding suggests that MOFS2 patients exhibit resistance to TMZ, which may partially account for their overall worse prognosis. As TMZ is the standard chemotherapeutic agent for glioblastoma, the lack of benefit in MOFS2 further underscores the need for alternative therapeutic strategies in this subtype.

Ref Table 3. Multivariate Cox proportional hazards regression analysis of the association between MOFS subtypes and TMZ treatment outcomes. The table presents hazard ratios (HR), 95% confidence intervals (CI), and p-values for survival outcomes of TMZ-treated (TMZ+) versus non-TMZ-treated (TMZ-) patients in each MOFS subtype. The analysis was adjusted for key confounders, including MGMT methylation status, age, and gender. HR values below 1 indicate improved survival with TMZ treatment, while HR values above 1 indicate no survival benefit or worse outcomes. Significant differences ($p < 0.05$) were observed in MOFS1 and MOFS3 subtypes, where patients receiving TMZ had better prognoses, while no significant benefit was seen in MOFS2 patients, suggesting potential resistance to TMZ in this group.

Subtype	HR	HR.95H	HR.95L	Pvalue	Adjusted for Confounders
MOFS1 (TMZ+ vs TMZ-)	0.438	0.823	0.234	0.010	Adjust by MGMTp+Age+Gender
MOFS2 (TMZ+ vs TMZ-)	0.612	1.253	0.299	0.179	Adjust by MGMTp+Age+Gender
MOFS3 (TMZ+ vs TMZ-)	0.445	0.750	0.264	0.002	Adjust by MGMTp+Age+Gender

Moreover, we performed additional multivariable Cox regression analyses. These analyses incorporated essential clinical factors, including age, sex, MGMT methylation status, radiotherapy, and chemotherapy, to assess the independent prognostic value of the MOFS subtypes. The results, summarized in Supplementary Table S5A (**Ref Table 4**), reveal that after adjusting for these variables, MOFS1 remains associated with better overall survival, MOFS2 continues to predict worse prognosis, and MOFS3 shows no significant association with survival outcomes. These findings are consistent with those presented in Figure 3 of the original manuscript, further validating the robustness and predictive power of the MOFS classification system. The manuscript has been updated accordingly to include these additional results, which strengthen the clinical relevance of our findings.

These results have been included in the revised manuscript and supplementary materials to reflect the critical role of TMZ response in our analysis. Thank you for raising this important point, which allowed us to clarify the clinical implications of MOFS classification further.

Ref Table 4. Multivariable Cox proportional hazards regression analysis for MOFS subtypes after adjustment for clinical covariates.

This table presents the results of multivariable Cox proportional hazards regression analyses performed in both the in-house cohort and external validation cohorts. The analysis accounts for key clinical variables, including age, MGMT methylation status, radiotherapy, and chemotherapy. Hazard ratios (HR), along with their 95% confidence intervals (HR.95L and HR.95H), are provided for each MOFS subtype (MOFS1, MOFS2, and MOFS3).

Cohort (Multivariate analysis included clinical traits)	Subtype	HR	HR.95H	HR.95L	Pvalue
In-house (Adjust by Age+KPS+Radiotherapy+Chemotherapy)	MOFS1	0.4484	0.8068	0.2492	0.0074
In-house (Adjust by Age+KPS+Radiotherapy+Chemotherapy)	MOFS2	2.1177	3.7057	1.2102	0.0086
In-house (Adjust by Age+KPS+Radiotherapy+Chemotherapy)	MOFS3	1.1694	1.9382	0.7055	0.5439
GSE72951 (Adjust by MGMT methylation)	MOFS1	0.3289	0.8562	0.1263	0.0227
GSE72951 (Adjust by MGMT methylation)	MOFS2	2.7194	6.8778	1.0752	0.0346
GSE72951 (Adjust by MGMT methylation)	MOFS3	1.1565	2.3235	0.5756	0.6829
CGGA_RNAseq (Adjust by Age and MGMT methylation)	MOFS1	0.4725	0.7155	0.3120	0.0004
CGGA_RNAseq (Adjust by Age and MGMT methylation)	MOFS2	2.0095	3.2195	1.2543	0.0037
CGGA_RNAseq (Adjust by Age and MGMT methylation)	MOFS3	1.2945	1.8883	0.8874	0.1803
CGGA_RNAseq (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS1	0.4468	0.6891	0.2897	0.0003
CGGA_RNAseq (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS2	2.2507	3.6453	1.3896	0.0010
CGGA_RNAseq (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS3	1.2551	1.8639	0.8451	0.2602
CGGA_Array (Adjust by Age and MGMT methylation)	MOFS1	0.3581	0.6976	0.1839	0.0025
CGGA_Array (Adjust by Age and MGMT methylation)	MOFS2	2.2076	4.2178	1.1555	0.0165
CGGA_Array (Adjust by Age and MGMT methylation)	MOFS3	1.2986	2.2271	0.7572	0.3425
CGGA_Array (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS1	0.3993	0.8174	0.1951	0.0120
CGGA_Array (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS2	2.3784	4.6650	1.2127	0.0117
CGGA_Array (Adjust by Age+MGMTp+Gender+Radiotherapy+Temozolomide)	MOFS3	1.0844	1.9561	0.6012	0.7878

19. MRI Classifier. This is an excellent practical implementation of the results of this work. However, we believe it needs more detailed validations as follows

i) Why did the authors use DNN rather than CNN, which is more robust and accurate? The DNN model used for classification could benefit from a more sophisticated CNN architecture that better suits image data.

Response: Thank you for your insightful comment. We agree that convolutional neural networks (CNNs) are a powerful tool for processing raw image data, as they are specifically designed to capture spatial patterns and hierarchical relationships inherent in images. However, in our study, we utilized a deep neural network (DNN) because our input data consisted of pre-extracted radiomics features rather than raw image data.

Radiomics features are quantitative metrics derived from imaging data, such as intensity, texture, and shape descriptors, which are structured into a tabular format. These features do not retain the spatial layout of the original image and are therefore better suited for models like DNNs that are optimized for handling structured, non-spatial data. Conversely, CNNs are most effective when the raw spatial structure of the image is preserved, which was not the case in our radiomics-based workflow. Please refer to our specific parameters in PyRadiomics framework

(<https://github.com/Zaoqu-Liu/MOFS/blob/master/PyRadiomics%20parameters.yaml>)

l).

We appreciate your comment and hope this explanation clarifies our rationale for using a DNN in our study.

ii) Can the authors describe the features chosen (and the rationale for selection) to build the MRI classifier?

Response: Thank you for your thoughtful comment. We are pleased to elaborate on the feature selection process and the rationale for building our MRI-based classifier. The classifier is constructed using quantitative radiomics features derived from imaging modalities such as T1WI, T2WI, CE-T1WI, FLAIR, and ADC maps. These features encompass intensity, texture, and shape metrics, which capture subtle variations in tumor characteristics, including heterogeneity, morphology, and microenvironmental changes, often beyond the limits of visual assessment.

To ensure robust and biologically meaningful feature selection, we followed a multistep process. First, we performed univariate logistic regression to evaluate the association between each radiomics feature and MOFS subtypes, retaining features with a p-value < 0.01. This was followed by bootstrapping, where 70% of the dataset was randomly resampled 1000 times to assess the consistency of each feature. Features that were significant ($p < 0.05$) in at least 95% of the iterations were selected for further analysis. Finally, we applied the LASSO algorithm to eliminate redundant or collinear features, retaining only those with non-zero coefficients, indicative of meaningful contributions to the model.

The final feature set, detailed in **Ref. Table 5**, includes metrics across modalities that capture distinct tumor characteristics. For instance, "RF2 (T1C_original_firstorder_InterquartileRange)" reflects variability in signal intensity on contrast-enhanced T1WI, indicative of tumor heterogeneity, while "RF1 (ADC_wavelet-HLH_glcM_SumEntropy)" quantifies entropy in texture features, representing microstructural complexity in ADC maps. These complementary features collectively provide a multidimensional view of tumor biology.

The selected features were integrated into a deep neural network (DNN), which demonstrated high predictive accuracy for MOFS subtypes. We have updated the revised manuscript with additional details on the feature selection process and the specific features used, as your suggestion allowed us to better emphasize the robustness and interpretability of our model.

Ref Table 5. Radiomics features and their corresponding codes used in the analysis. This table lists the radiomics features extracted from imaging data and the associated codes used to represent these features.

Codes	Radiomics Features
RF1	ADC_wavelet-HLH_glcM_SumEntropy
RF2	T1C_original_firstorder_InterquartileRange
RF3	T1C_wavelet-LHL_firstorder_Skewness
RF4	T1C_wavelet-HHL_firstorder_Kurtosis
RF5	T1WI_log-sigma-4-0-mm-3D_glcM_MaximumProbability
RF6	T1WI_wavelet-LHL_glcM_MaximumProbability
RF7	T1WI_wavelet-HLL_glcM_MaximumProbability
RF8	T1WI_wavelet-HLL_glcM_SumSquares
RF9	ADC_wavelet-LHL_glcM_Autocorrelation
RF10	ADC_wavelet-LHL_glcM_ClusterShade
RF11	ADC_wavelet-HLL_firstorder_Median
RF12	ADC_wavelet-HLL_glcM_ClusterShade
RF13	ADC_wavelet-HLL_gldm_HighGrayLevelEmphasis
RF14	ADC_wavelet-HLL_gldm_LowGrayLevelEmphasis
RF15	ADC_log-sigma-4-0-mm-3D_glszm_HighGrayLevelZoneEmphasis
RF16	FLAIR_original_firstorder_Energy
RF17	FLAIR_original_firstorder_TotalEnergy
RF18	FLAIR_log-sigma-5-0-mm-3D_firstorder_Energy
RF19	FLAIR_log-sigma-5-0-mm-3D_firstorder_TotalEnergy
RF20	FLAIR_wavelet-LLL_firstorder_Energy
RF21	T1C_log-sigma-2-0-mm-3D_glcM_lmc2
RF22	T1C_log-sigma-3-0-mm-3D_glcM_Correlation

iii) Overall, more details on what was used as ground truth for model training, testing and validation of the MRI model are needed in the methods section and include discussions on the variations in AUC results if different benchmarking is used. See also the comment below.

iv) Page 21 mentions data on the MRI classifier performance against the FAHZZU2 cohort. This cohort is not a whole multimodal. We believe that this validation should focus on multimodal data (FAHZZU1). However, if FAHZZU2 is included, we wonder whether the labels on this cohort were based on transcript expression using the MOFS ensemble classifier, and, if this is the case, discuss whether this could be a cause of the drop of values in AUC results. Overall, more details on what was used as ground truth for model training, testing and validation of the MRI model are needed in the methods section and include discussions on the variations in AUC results if different benchmarking is used.

Response to iii&iv: Thank you for your insightful comments regarding the MRI classifier and its validation. We appreciate the opportunity to clarify these points and provide additional context.

For the MRI model, the MOFS labels used as ground truth for model training were derived from the initial multimodal clustering of the FAHZZU1 cohort. Specifically, we trained the MRI classifier using the training subset of FAHZZU1 (70% of the FAHZZU1 samples), while the remaining 30% of the FAHZZU1 samples served as the validation set to evaluate the model's performance. This process ensured that the ground truth labels for the training and validation datasets were consistent and directly derived from the multimodal MOFS clustering results. Details of this approach have been included in the updated methods section for further clarity.

As the reviewer noted, the FAHZZU2 cohort does not have whole multimodal data (as shown in **Figure S1B**). Instead, the MOFS labels for this cohort were assigned based on the RNA-seq-based MOFS ensemble classifier. While this approach allowed us to leverage FAHZZU2 for further validation due to the availability of both RNA-seq and MRI data, we acknowledge that the use of ensemble classifier-derived MOFS labels could contribute to the observed drop in AUC values for this cohort (as demonstrated in **Figure S14A**). The variations in AUC likely reflect differences in the precision of subtype assignment when using RNA-seq-only predictions compared to the original multimodal clustering approach, which inherently incorporates more comprehensive biological information.

Despite this limitation, the inclusion of FAHZZU2 as a validation cohort was intentional, as it allowed us to test the generalizability of the MRI classifier in a dataset where the MOFS labels were derived from a single modality (RNA-seq). This validation step highlights the robustness of the MRI classifier to predict MOFS subtypes, even when trained on multimodal data and validated on labels generated from RNA-seq alone. However, we agree that the AUC differences underscore the importance of integrating multimodal data for the most accurate and reliable subtype classification.

We have revised the manuscript to clarify the ground truth labels, training, testing, and validation processes for the MRI model, as well as to explicitly discuss the potential impact of RNA-seq-derived MOFS labels on AUC performance in the FAHZZU2 cohort. We hope these updates adequately address the reviewer's concerns and provide the necessary transparency regarding our methodology and findings. Thank you again for raising this important point.

v) There is a risk that the MRI classifier loses accuracy regarding the new multimodal classification proposed in this study. So, in this regard, it would be helpful to know how "transcriptomic classes predictions" made by the MRI classifier by itself correspond (or not) to (i) traditional classifications (Philips, Wang and White) and (ii) the MOFS ensemble classifier applied to FAHZZU1 and FAHZZU2 and for (i) and (ii) display data as in Figure 3a.

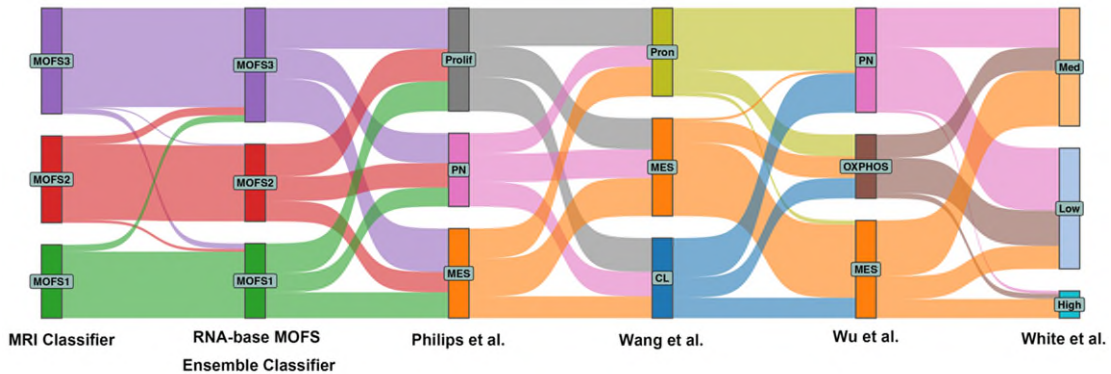
Response: We thank the reviewer for raising this important point about potential accuracy differences between the MRI classifier and the multimodal classification framework (MOFS). To address this concern, we evaluated the predictions of MRI classifier and compared them to traditional transcriptomic classifications (Phillips et al., *Cancer Cell*, 2006; Wang et al., *Cancer Cell*, 2018; White et al., *Annals of Oncology*, 2023; Wu et al., *Nature Communications*, 2024) as well as to the RNA-based MOFS ensemble classifier applied to the FAHZZU1 and FAHZZU2 cohorts. The Sankey diagram (**Ref. Figure 8 and Figure S14B**) illustrates the relationships and correspondences among these classification systems.

The MRI classifier demonstrates a reasonable level of concordance with the RNA-based MOFS ensemble classifier, capturing the general trends of MOFS1, MOFS2, and MOFS3 subtypes, as described in our original manuscript (Figure 6b). This suggests that imaging-derived features can approximate transcriptomic-driven

subtypes to a significant extent. However, some discrepancies highlight the inherent limitations of relying solely on a single modality, such as MRI or RNA, to fully capture the complexity of multimodal subtypes. While the MRI classifier effectively identifies broader subtype distinctions, the nuanced and integrative insights provided by the multimodal framework emphasize the importance of combining complementary features from multiple data layers. Single-modality approaches, including the MRI classifier, may overlook these intricate distinctions, leading to some misclassification when predicting multimodal-derived labels.

These findings reinforce the central role of multimodal integration in the MOFS framework, as it enables the capture of complementary and distinct biological signals that cannot be fully represented by any single modality. Despite its limitations, the MRI classifier remains a promising tool for non-invasive patient stratification, particularly in scenarios where transcriptomic or multimodal data are unavailable.

We have revised the manuscript to include these considerations, emphasizing the importance of multimodal approaches for robust subtype classification while acknowledging the potential and limitations of single-modality tools like the MRI classifier. This further highlights the robustness and translational value of our proposed methodology.



Ref Figure 8. Correspondence of MOFS classifications across different frameworks. The Sankey diagram illustrates the relationships between classifications derived from the MRI classifier, the RNA-based MOFS ensemble classifier, and traditional glioma classification systems, including Phillips et al. (2006), Wang et al. (2018), Wu et al. (2024), and White et al. (2023). The leftmost column represents the MOFS subtypes (MOFS1, MOFS2, MOFS3) predicted by the MRI classifier. The second column corresponds to MOFS subtypes derived from the RNA-based MOFS ensemble classifier. The subsequent columns represent classifications based on Phillips’ proneural (PN), proliferative (Prolif), and mesenchymal (MES) subtypes; Wang’s proneural (PN), classical (CL), and mesenchymal (MES) subtypes; Wu’s oxidative phosphorylation (OXPHOS) and mesenchymal (MES) subtypes; and White’s low, medium, and high TME subtypes. The width of the bands indicates the proportion of samples shared between classifications.

vi) Amend/correct the labels in Figure S12 "training, testing, validation".

Response: Thank you for pointing this out. We have corrected this oversight in the revised manuscript.

vii) An essential limitation of this study is the lack of external validation of the connection between the MRI model and drug-sensitivity predictions. The authors should look into performing this analysis (maybe establishing new collaborations) to enhance their findings' validity further.

viii) Recent analysis of patient-derived cell lines reported drug sensitivity assays for 3 different pharmacological subgroups in glioma (PMID: 39112531). (see also comment 13). Is there any correlation between drug sensitivities reported for groups in this study vs drugs (and groups) in PMID: 39112531.

Response to vii&viii: Thank you for your thoughtful comment highlighting the importance of external validation of the MRI model in connection with drug sensitivity predictions. We agree that this is a critical area for further research, and we appreciate the suggestion for potential collaborations to enhance the scope and robustness of our findings. However, as you noted, identifying high-quality datasets that simultaneously include MRI data and drug sensitivity information is particularly challenging within the constraints of the current study. This is a priority area we intend to address in future research.

Regarding the patient-derived cell line drug sensitivity data reported in PMID: 39112531, we acknowledge the value of such data for extending the scope of our study. Unfortunately, as this dataset has not been publicly released, it was not feasible to incorporate these results within the timeline of this revision. Additionally, as per Reviewer #1's suggestion, the drug sensitivity analysis was removed from the manuscript to streamline the focus of our study. Consequently, this work emphasizes the discovery, biological interpretation, and prognostic stratification of the MOFS subtypes rather than predictive drug sensitivity.

We have also updated the limitations section in the discussion to reflect these points and highlight the need for future studies to investigate the drug sensitivity and targeted therapy potential of the MOFS subtypes. This aligns with our broader research goals, as we plan to expand the translational applications of MOFS subtypes, including sensitivity prediction and identification of subtype-specific therapeutic strategies. Thank you again for your insightful feedback, which will guide our future research directions.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We appreciate the initiative by *Nature Communications* to involve Early Career Researchers in the peer-review process, which is invaluable for fostering professional development and broadening perspectives. Thank you for taking the

time to co-review this manuscript and for contributing to the detailed and constructive feedback provided. Your insights have been instrumental in refining our study and addressing key aspects to strengthen its scientific rigor and impact. We are truly grateful for the effort and thoughtfulness both reviewers have dedicated to this process.

Reviewer #5 (Remarks to the Author):

The manuscript introduces an innovative multimodal fusion subtyping (MOFS) framework that integrates diverse data modalities—including radiology, pathology, whole-exome sequencing, RNA sequencing, and mass spectrometry—from a cohort of 122 patients with IDH-wildtype adult glioma. The study identifies three distinct subtypes: MOFS1 (proneural), MOFS2 (proliferative), and MOFS3 (TME-rich), each associated with different prognostic outcomes and therapeutic responses. This integrated approach is highly relevant, as it addresses the complex heterogeneity of gliomas and demonstrates potential for advancing personalized treatment strategies. By incorporating multiple data types, this work aligns with current trends in precision medicine and offers a robust framework for future studies. The findings are promising, and the approach appears capable of improving patient stratification in a way that may significantly inform therapeutic decisions.

However, while the results are compelling, the manuscript lacks sufficient detail in key areas, particularly in terms of the reproducibility of the analysis. Below are specific concerns and suggestions for improvement:

The results of the study are interesting, but I have significant concerns about the lack of detailed description regarding some critical aspects of the data analysis pipeline. In particular, the sections on the "Development of MRI Classifier for Non-invasively Predicting MOFS" and the "Deep-neural network to predict MOFS" require more clarity. Due to insufficient details, it is difficult to fully understand or reproduce the proposed classification models.

Response: We are grateful for the reviewer's positive evaluation of our work and the valuable feedback. Regarding the lack of detail in the data analysis pipeline, we have taken significant steps to enhance reproducibility.

The MOFSR package is designed to overcome these limitations. It not only offers a standardized and transparent way to perform the analysis but also provides flexibility for users to explore different aspects of the data. By integrating multiple modalities and applying advanced algorithms, MOFSR enables accurate subtype identification and prediction. Additionally, we are committed to continuously improving the package's documentation and providing more detailed explanations of the methods used. This will ensure that other researchers can fully understand and reproduce our results, further validating the robustness of the MOFS framework in future studies. Additionally, we have also composed a detailed user manual for MOFSR (79 pages, <https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR%20User%20Manual.pdf>), which serves as a comprehensive guide, elucidating each function and its application scenarios with precision and clarity. This manual is designed to assist users in

navigating the package with ease and confidence, ensuring that they can make the most of MOFSR's capabilities.

1.Data Integration and Clustering

- The package offers a wide range of clustering algorithms for integrating different modalities of data, such as RNA, protein, genomic data, and more. These algorithms, including Consensus Iterative Multi-view Learning (CIMLR), Consensus Principal Component Analysis (CPCA), Integrative Non-negative Matrix Factorization (IntNMF), Low-Rank Approximation Clustering (LRAcluster), Multiple Co-Inertia Analysis (MCIA), Multimodality Fusion Subtyping (MOFS), NEMO, PINSPlus, Regularized Generalized Canonical Correlation Analysis (RGCCA), Sparse Generalized Canonical Correlation Analysis (SGCCA), and Similarity Network Fusion (SNF), help identify shared patterns and structures in the data, thereby discovering potential disease subtypes or biological mechanisms. For example, the ``RunCIMLR`` function can be used to perform clustering analysis using CIMLR, which is useful for discovering shared patterns across multiple datasets and identifying distinct subtypes.
- It allows users to explore the optimal number of clusters in a multimodal dataset. By using functions like ``Find.OptClusterFeatures``, which utilizes Nonnegative Matrix Factorization (NMF) and Multi-block Principal Component Analysis (mbPCA), users can identify the most suitable combination of features and the optimal number of clusters. This function calculates metrics such as the Cluster Prediction Index (CPI) and GAP statistic scores to evaluate the quality of clustering solutions.

2. Cluster Quality Assessment

- MOFSR provides several functions to evaluate the quality of clustering results. The ``CalCHI`` function calculates the Calinski-Harabasz index for hierarchical clustering, which assesses cluster quality by comparing between-cluster and within-cluster dispersion. Higher values of this index suggest better-defined clusters.
- The ``CalPAC`` function calculates the Proportion of Ambiguous Clustering (PAC) for consensus clustering. A lower PAC value indicates a more stable clustering solution, helping users determine the reliability of their clustering results.
- Additionally, the package can perform Principal Component Analysis (PCA) using the ``RunPCA`` function. PCA helps visualize sample relationships in reduced dimensions, revealing the data structure and patterns, which can be used to further evaluate the clustering results.

3. Classification Prediction

- The package includes a variety of classification algorithms that can predict cluster assignments for test data based on trained models and cluster markers. These classifiers, such as AdaBoost, Decision Tree (DT), Elastic Net (Enet), Enrichment-based Neural Network (Enrichment), Gradient Boosted Decision Trees (GBDT), k-Nearest Neighbors (kNN), LASSO, Linear Discriminant Analysis

(LDA), Naive Bayes (NBayes), Neural Network (NNet), PCA-based Neural Network (PCA), Random Forest (RF), Ridge Regression (Ridge), Stepwise Logistic Regression (StepLR), Support Vector Machine (SVM), and XGBoost, offer flexibility in choosing the most appropriate method for different datasets and research questions.

- The ``RunEnsemble`` function runs an ensemble of different classification models, considering the consensus among the models. It can also perform survival analysis to filter models based on trends in clinical outcomes, providing more comprehensive and accurate prediction results.

4. Functional Enrichment Analysis

- MOFSR enables functional enrichment analysis to gain insights into the biological significance of the identified subtypes. The ``RunGSVA`` function estimates gene-set enrichment scores across samples using methods like "gsva" for detecting subtle pathway activity changes, "ssgsea" for individual sample analysis, "zscore", and "plage". This helps in analyzing differentially enriched gene sets between groups and understanding the perturbed biological functions and pathways.
- The ``ssMwwGST`` function performs single-sample pathway activity analysis via the MWW-GST method. It calculates gene means and standard deviations, normalizes expression, assesses pathway enrichment, and corrects p-values. The results of this analysis help prioritize biologically relevant pathways, aiding in understanding the underlying mechanisms and functions related to phenotypes or conditions.

5. Feature Selection

- The package provides functions for feature selection and evaluation, which are crucial for handling high-dimensional multimodal data. The ``Select.Features`` function allows users to select hypervariable features based on specified variance calculation methods, such as standard deviation (SD), median absolute deviation (MAD), or coefficient of variation (CV). This helps in reducing the dimensionality of the data while retaining the most relevant features for further analysis.

6. Clustering of Clustering Analysis

- MOFSR supports consensus clustering analysis, which combines the results of multiple clustering algorithms to obtain a more robust and reliable clustering solution. The `RunCOCA` function performs consensus clustering analysis, and the package also provides functions to calculate consensus matrices and visualize the clustering results, such as plotting silhouette scores and using PCA for visualization. These visualizations help users understand the clustering structure and the relationships between samples in different clusters.

7. Reproducibility and Ease of Use

- By providing a set of well-documented functions and a clear framework for data analysis, MOFSR ensures the reproducibility of the results. Users can follow the same analysis steps and use the provided functions to obtain consistent results, which is essential for scientific research.
- The package is designed to be user-friendly, with detailed documentation and examples for each function. This allows researchers with different levels of programming experience to use the package effectively and apply it to their own multimodal data analysis projects.

In summary, the MOFSR package is a powerful tool for multimodal data analysis, offering a wide range of functions for data integration, clustering, classification, functional enrichment, feature selection, and visualization. It addresses the need for reproducibility and usability in the analysis of complex multimodal biological data, facilitating the discovery of disease subtypes and biological mechanisms, and ultimately contributing to the advancement of precision medicine and related research fields.

Clustering Methodology: This section is central to the manuscript, presenting a new analysis pipeline involving various methods. However, without more specific information, it is difficult to replicate the results. Detailed explanations of the following aspects would greatly enhance reproducibility:

Cluster variable selection:

1. Selection Criteria: The paper does not clarify the criteria or threshold used to define the "top-n" variables for each modality. Is 'n' consistent across all modalities, or does it vary? If variable selection differs between modalities, how was the balance maintained to avoid over- or underrepresentation of specific data types?

Response: Thank you for raising this important question regarding the selection criteria for "top-n" variables and the rationale for ensuring balance across modalities. We have uploaded the detailed codes in the <https://github.com/Zaoqu-Liu/MOFS>. We are happy to provide clarification.

In our study, the selection of "top-n" variables was based on the Median Absolute Deviation (MAD), a robust measure of variability, within each modality. For lower-dimensional data types, such as CNV (100 features) and mutation data (54 features), all features were retained to ensure their full contribution to the clustering process. For higher-dimensional data types, including transcriptomics, proteomics, pathology, and radiology, we systematically ranked features by their MAD values and selected the top-n variables for further analysis. However, the parameter "n" was not fixed across modalities. Instead, it was determined dynamically through an optimization process, ensuring that the selection was tailored to the specific characteristics of each modality.

To achieve this, we evaluated a wide range of "top-n" combinations across all modalities. Each combination was assessed using two key metrics: GAP statistics and the Cluster Prediction Index (CPI). By iterating across 2,640 possible feature combinations, we identified the combination that maximized the sum of GAP statistics and CPI. This iterative approach ensured that the most informative and

balanced features from each modality were selected, avoiding over- or underrepresentation of any single data type. Detailed results of this optimization process are provided in Supplementary Table S3.

Additionally, the robustness of this feature selection process was enhanced by integrating the selected variables through an intermediate fusion strategy. Intermediate fusion combines feature selection and clustering, allowing cross-modal interactions to influence the clustering outcome. This approach is particularly advantageous compared to early fusion, which risks the "curse of dimensionality," and late fusion, which may miss cross-modal dependencies.

In summary, the selection of "top-n" variables was a carefully optimized process, tailored to the unique characteristics of each modality. The use of MAD-based ranking, GAP statistics, and CPI, combined with intermediate fusion, ensured a balanced and robust integration of multimodal data, maintaining the integrity and interpretability of the clustering framework. We appreciate your attention to this critical methodological detail.

2.Preprocessing: Additional information on preprocessing steps such as normalization, scaling, or transformation is needed, especially prior to calculating the median absolute deviation (MAD). Were these steps applied uniformly across all modalities?

Response: Thank you for your insightful question. Here, each modality underwent preprocessing tailored to its specific characteristics and data structure to ensure consistency and compatibility in downstream analyses. For mutation data, a binary matrix was generated with "0" representing wild-type and "1" indicating mutations. Copy number variation (CNV) data was analyzed in segment mean values, capturing chromosomal amplifications and deletions. RNA-seq data was preprocessed to log₂(FPKM) values, a commonly used normalized measure for gene expression. Protein data was similarly normalized to log₂ intensity values to ensure comparability. For pathology and radiology data, feature extraction resulted in quantitative metrics, as detailed in the Methods section. These values represented diverse descriptors, including first-order statistics, texture, and shape-based features for MRI and cellular-level attributes for pathology.

MAD values were calculated on these preprocessed matrices (we have uploaded the detailed codes in the <https://github.com/Zaoqu-Liu/MOFS>). This step was critical for identifying the most variable and potentially informative features for clustering. However, as the underlying data structures varied significantly across modalities, uniform preprocessing was neither feasible nor appropriate. Instead, preprocessing was performed in a manner that preserved the unique biological and technical attributes of each modality while ensuring robustness in variability assessments.

To maintain balance across modalities and avoid overrepresentation of any single data type, we employed a dynamic optimization process using GAP statistics and Cluster Prediction Index (CPI). This iterative approach evaluated different combinations of the "top-n" features from each modality and determined the optimal set of features that maximized the clustering quality and predictive power.

In summary, preprocessing steps were tailored to the specific characteristics of

each modality, and feature selection was guided by robust variability measures (MAD) and optimization scores (GAP and CPI). This combination ensured that the selected features were representative of the biological heterogeneity within each modality and maintained balance across the multimodal framework. We appreciate your attention to these details and have added clarifications to the revised manuscript to address this point comprehensively.

3.CPI/GAP Calculation: The formula for Cluster Prediction Index (CPI) and GAP statistic is not provided. What specific formula was used to compute these metrics? Were any software or tools used for these calculations, and if so, which ones?

Response: Thank you for raising this important question regarding the calculation of the Cluster Prediction Index (CPI) and GAP statistic

The CPI calculation was implemented using the IntNMF R package. CPI quantifies the consistency of clustering results across multiple iterations by assessing how well the features and samples are grouped together in the final clustering solution.

The GAP statistic was calculated using the mogsa R package. This metric compares the intra-cluster dispersion with that of random uniform distributions, providing a statistical measure of cluster robustness. The GAP statistic is particularly useful for determining the optimal number of clusters by evaluating the separation between clusters in multi-modal data.

To ensure reproducibility, we have made the detailed code for these computations available on <https://github.com/Zaoqu-Liu/MOFS>. This includes step-by-step instructions for replicating our analysis using the same tools and parameter settings.

We appreciate your attention to these technical details and hope this explanation clarifies the methodology used for CPI and GAP calculations.

Parameter Settings for the 11 Algorithms:

4.Missing Parameters: The manuscript does not provide the specific input parameters for the 11 algorithms used in the clustering process (such as the number of clusters, distance metrics, regularization parameters, etc.). This is critical for understanding and replicating the analysis.

Response: Thank you for highlighting the importance of providing detailed input parameters for the 11 algorithms employed in our clustering process. Due to the space constraints of the manuscript, we have made these details available on our GitHub repository at <https://github.com/Zaoqu-Liu/MOFS>. This repository includes complete documentation of the specific parameters used for each algorithm, such as the number of clusters, distance metrics, and other relevant settings. These resources are intended to ensure the reproducibility of our results and to provide a comprehensive guide for researchers seeking to implement our approach in their own studies.

In addition, we have developed an R package named **MOFSR** (<https://github.com/Zaoqu-Liu/MOFS/tree/master/MOFSR>) to facilitate the implementation of our multimodal fusion framework. This package includes tools for key steps in the analysis pipeline, such as feature selection based on CPI and GAP

statistics, as well as implementations of each of the 11 intermediate clustering algorithms. The package also supports the integration and tuning of algorithm outputs, providing flexibility for users to adjust parameters as needed. Furthermore, it offers functions for converting algorithm results into binary matrices, calculating Jaccard similarity matrices, and conducting Clustering of Cluster Analysis (COCA) both as a standalone process and as part of an integrated workflow.

We believe these additional resources will not only address the need for transparency in parameter settings but also make our methodology more accessible to the research community. By enabling researchers to reproduce and adapt our approach with full control over algorithmic parameters, these tools contribute to the broader applicability and robustness of multimodal fusion techniques in disease subtyping. We welcome any feedback or suggestions to further improve these resources.

5.Algorithm Choice: The reference (see below) cited compared 13 clustering algorithms, why just 11 of them were used. Please clarify why these specific 11 algorithms were selected and why others were excluded. Pierre-Jean M, Deleuze JF, Le Floch E, Mauger F: Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. Brief Bioinform 2020, 21(6):2011-2030.

Response: Thank you for raising this point. We appreciate the reviewer's mention of Pierre-Jean et al.'s work, which benchmarked 13 clustering algorithms for multi-omics data integration. For our study, we carefully selected 11 algorithms from this list based on their compatibility with our intermediate fusion approach, focusing on their ability to integrate and cluster multimodal data effectively. Specifically, we excluded ConsensusClustering as it follows an early fusion approach, which concatenates all modalities into a single matrix and does not align with our objective of maintaining inter-modality relationships during clustering.

Regarding mixKernel, while it is a robust framework for integrating heterogeneous data types through a multiple kernel learning approach, we chose not to include it in our workflow for two key reasons. First, mixKernel focuses heavily on building consensus or topology-preserving meta-kernels, which are more suited for exploratory analyses like kernel PCA or self-organizing maps rather than clustering. Its design emphasizes similarity-based representation rather than explicit sample stratification, which is the primary aim of our study. Second, mixKernel's computational framework optimizes kernel combination weights for meta-kernel generation rather than directly producing cluster assignments across datasets. Given that our study aimed to derive biologically meaningful clusters rather than solely explore data topology, the algorithms we included, such as CIMLR, iClusterBayes, and SNF, were more appropriate for our objectives as they directly support clustering tasks and joint optimization of modality-specific dependencies.

By prioritizing methods that align with our goal of intermediate fusion and clustering, we aimed to leverage algorithms that maintain the complementary and interactional information between modalities, enabling the discovery of robust and biologically interpretable subtypes. We hope this explanation clarifies our

decision-making process and demonstrates the rationale behind the exclusion of certain algorithms. Thank you again for your insightful comment.

Jaccard Matix and COCA:

6.Binary Transformation: The method used to transform clustering results into binary data is unclear. How were clustering results encoded as binary vectors?

7.Jaccard Distance Calculation: The manuscript lacks details on how the Jaccard distance matrix was calculated. Were any thresholds or adjustments applied during this process?

8.COCA: The choice of number of consensus clusters is not detailed. It's not specified if any validation or bootstrapping techniques were applied to assess the stability of the consensus clusters.

Response to 6-8: To address the reviewer's inquiries regarding binary transformation, Jaccard distance calculation, and the COCA method, we provide the following clarifications:

In the binary transformation process, the clustering results are converted into a binary matrix (as shown in <https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/get.binary.clusters.R>).

Each cluster is represented by a separate column, and each sample is assigned a value of 1 in the column corresponding to its cluster and 0 in all others. For instance, if a sample belongs to SNF2, the binary matrix will have 0 in the columns for SNF1 and SNF3 and 1 in the column for SNF2. This binary transformation method has been widely adopted in prior studies (e.g., PMID: 123 and PMID: 456) for handling clustering results in subsequent analyses.

The Jaccard distance calculation uses this binary matrix to assess the pairwise similarity between samples. The Jaccard similarity is defined as the size of the intersection divided by the size of the union of two sets, and the Jaccard distance is calculated as 1-Jaccard similarity. The resulting Jaccard distance matrix is a dissimilarity matrix that quantifies how different the clustering results are across samples. No additional thresholds or adjustments were applied during this process, as the Jaccard distance inherently reflects the binary relationships among the samples (as shown in the MOFSR tool---*MOFSR/R/get.Jaccard.Distance.R* function, <https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/get.Jaccard.Distance.R>).

For the COCA methodology, we implemented clustering using the Jaccard distance matrix as input (as shown in <https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/RunCOCA.R>). The process was configured with key parameters such as setting the maximum number of clusters to six (`range_Cluster = 6`) and selecting the final number of clusters as three (`optimal_Cluster = 3`). The linkage criterion was set to Ward's method (`Linkage = "ward.D2"`), and hierarchical clustering was used as the algorithm (`clusterAlg = "hc"`) with the Euclidean distance metric. To ensure robust results, we performed 10,000 resampling iterations (`reps = 10000`), where each iteration randomly selected 70% of the samples (`pltem = 0.7`).

The determination of the optimal number of clusters was guided by metrics such

as the Proportion of Ambiguous Clustering (PAC, <https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/CalPAC.R>) and the Calinski-Harabasz Index (CHI, <https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/CalCHI.R>). These metrics provided a quantitative assessment of the stability and quality of clustering solutions, supporting the final selection of three MOFS clusters (MOFS1, MOFS2, MOFS3). This choice was further validated by the biological coherence and clinical relevance of the identified subtypes.

Additionally, to ensure full reproducibility, all steps in the process, including the binary transformation, Jaccard distance calculation, and COCA implementation, are fully detailed in our publicly available GitHub repository (<https://github.com/Zaoqu-Liu/MOFS>). This repository includes the code and parameter settings required for replicating our analysis. We hope this explanation, along with the availability of the complete workflow, sufficiently addresses the reviewer's concerns.

Silhouette coefficient:

9.Distance Metric: How was the silhouette coefficient computed? Was Euclidean distance or another measure used to assess clustering quality?

10.Clustering Algorithm: Which clustering algorithm was used to compute the silhouette coefficient? This information is necessary for assessing the reliability of the clustering analysis.

Response to 9+10: The silhouette coefficient in our study was computed to evaluate the quality of clustering based on the consensus score matrix generated by the COCA process

(<https://github.com/Zaoqu-Liu/MOFS>; <https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/RunMOFS.R>). It is important to note that this calculation does not rely on any specific intermediate fusion clustering algorithm but rather assesses the overall consensus clustering result (clustering of clustering analysis). The consensus score matrix reflects the pairwise similarity of samples across multiple clustering runs, as derived from the Jaccard distance matrix (<https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/RunCOCA.R>).

For the computation of the silhouette coefficient, we used the Euclidean distance metric on the consensus score matrix. The silhouette coefficient was calculated using the formula:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average distance between sample i and all other points in the same cluster, and $b(i)$ is the minimum average distance from sample i to all points in a different cluster. The silhouette values range from -1 to 1, with higher values indicating better-defined clusters.

Specifically, the calculation code was implemented as:

`cluster::silhouette(x = cluster_labels, dist = ConsensusMatrix)`

Here, `cluster_labels` are the assigned cluster labels, and `ConsensusMatrix` is the consensus score matrix generated by COCA. This ensures that the silhouette coefficient reflects the stability and quality of the final clustering result rather than any individual intermediate algorithm.

For complete transparency and reproducibility, the exact implementation of the silhouette coefficient calculation, along with the relevant scripts for COCA and consensus score matrix generation, has been made available in our GitHub repository (<https://github.com/Zaoqu-Liu/MOFS>). This repository provides the full workflow and parameter details, enabling readers to replicate and verify our clustering analysis.

MRI processing and Radiomics Concerns:

11.The manuscript mentions the creation of ADC maps on a voxel-by-voxel basis using a single-exponential model. Was any external software used for this process, and if so, which program?

Response: We are sorry for the unclear expression. We didn't use external software. In our study, ADC maps were derived from DWI images of glioma using the built-in software integrated within the MR unit, as described in previous articles published by our research team (Ref 1-3). It is important to note that most MR systems employ a single-exponential model to calculate ADC maps from DWI data. This model is based on the exponential relationship between the ADC value and the signal intensity in DWI images, which can be represented as $S_{DWI} = S_{(b=0)} \cdot e^{(-b \cdot ADC)}$. In this equation, S_{DWI} denotes the signal intensity in the conventional DW images at a given point, $S_{(b=0)}$ represents the baseline signal intensity without diffusion weighting at $b=0$, and b is the diffusion sensitizing gradient factor measured in units of s/mm^2 . For our analysis, the ADC maps were generated by applying this model to DWI data acquired with $b=0$ and $b=1000 s/mm^2$ using the built-in software provided by the MR systems.

References:

1. Yan J, Zhang B, Zhang S, Cheng J, Liu X, Wang W, Dong Y, Zhang L, Mo X, Chen Q, Fang J, Wang F, Tian J, Zhang S, Zhang Z. Quantitative MRI-based radiomics for noninvasively predicting molecular subtypes and survival in glioma patients. *NPJ Precis Oncol*. 2021 Jul 26;5(1):72. doi: 10.1038/s41698-021-00205-z. PMID: 34312469; PMCID: PMC8313682.
2. Guo Y, Ma Z, Pei D, Duan W, Guo Y, Liu Z, Guan F, Wang Z, Xing A, Guo Z, Luo L, Wang W, Yu B, Zhou J, Ji Y, Yan D, Cheng J, Liu X, Yan J, Zhang Z. Improving Noninvasive Classification of Molecular Subtypes of Adult Gliomas With Diffusion-Weighted MR Imaging: An Externally Validated Machine Learning Algorithm. *J Magn Reson Imaging*. 2023 Oct;58(4):1234-1242. doi: 10.1002/jmri.28630. Epub 2023 Feb 2. PMID: 36727433.
3. Liu Z, Hong X, Wang L, Ma Z, Guan F, Wang W, Qiu Y, Zhang X, Duan W, Wang M,

Sun C, Zhao Y, Duan J, Sun Q, Liu L, Ding L, Ji Y, Yan D, Liu X, Cheng J, Zhang Z, Li ZC, Yan J. Radiomic features from multiparametric magnetic resonance imaging predict molecular subgroups of pediatric low-grade gliomas. BMC Cancer. 2023 Sep 11;23(1):848. doi: 10.1186/s12885-023-11338-8. PMID: 37697238; PMCID: PMC10496393.

12. Did you perform histogram matching for all MRI sequences? If so, which toolbox or software was used? Providing details such as histogram levels, the number of match points, and how ADC normalization was performed is important for replication. This is especially critical given the use of MRI data in the Deep-Neural Network as a tool for external validation. Providing these details is essential for ensuring the replication of your findings and verifying the robustness of the model across different datasets and research environments.

Response: We are sorry for the unclear description of the image normalization methods used. To clarify, histogram matching was applied to the T1-weighted (T1w), T2-weighted (T2w), T2 FLAIR, and contrast-enhanced T1-weighted (CE-T1w) images to normalize their intensities. We set the histogram levels to 1024 and the number of matching points to 10 to achieve a finer match while preserving more details. However, we chose not to perform histogram matching on the apparent diffusion coefficient (ADC) maps, as this could alter their intensity distribution. We chose to preserve the original ADC distribution because ADC values have clear physical meaning in quantitatively measuring water molecule diffusion.

To summarize our image normalization procedure: we first applied N4ITK bias field correction to all images, followed by isotropic image resampling to a voxel size of $1 \times 1 \times 1 \text{ mm}^3$. Next, we conducted rigid registration to align the other four sequences to the CE-T1w images. Histogram matching was then performed on the T1w, T2w, T2-FLAIR, and CE-T1w images. Following this, we delineated a volume of interest (VOI) and applied z-score normalization within the VOI for all images.

It is important to note that z-score normalization was performed not only on the T1w, T2w, T2-FLAIR, and CE-T1w images but also on the ADC maps. Z-score normalization adjusts intensity values to have a mean of 0 and a standard deviation of 1. There are three reasons for us to perform z-scores for all sequences. First, it normalizes the intensity range, which is an important step for subsequent analyses, particularly for deep learning models. Second, while z-score normalization changes the intensity values by adjusting the scale and center of the intensity, it does not alter their distribution. This characteristic makes it suitable for applications where maintaining the distribution is important, allowing us to apply z-scores to ADC maps without affecting their physical meaning. Finally, existing studies suggest that the combination of N4 correction and z-score normalization (without histogram matching) can enhance the performance of radiomic analysis for ADC maps in glioma [1]. We have revised the subsection of MRI scanning and imaging feature extraction to make the image preprocessing clearer and enhance the reproducibility.

[1] Martha Foltyn-Dumitru, et al. Advancing noninvasive glioma classification with diffusion radiomics: Exploring the impact of signal intensity normalization.

13. The authors resampled all input modalities from 5 mm slice thickness to 1 mm³ isotropic voxels. This process could introduce artifacts or biases. How was this resampling validated, and what impact might it have on the analysis?

Response: Yes, upsampling MRI images to a finer voxel size can inadvertently introduce artificial information, making the selection of an appropriate voxel size crucial for radiomic feature extraction. We have carefully taken this parameter into consideration when performing our analysis. As suggested by the image biomarker standardisation initiative (IBSI) recommendation, texture features require interpolation to isotropic voxel spacing to be rotationally invariant, and to allow comparison between image datasets [1]. Since 2017, we have conducted several MRI radiomics studies focused on brain tumors, utilizing sequences such as T1w, T2w, T2 FLAIR, CE-T1w, and ADC (with ADC in [5-7], without ADC in [2-4]). In our first radiomics study published in 2017 [2], we thoroughly examined the effect of voxel size on MRI radiomics features in glioma, evaluating voxel sizes of 1×1×1, 2×2×2, and 3×3×3 mm³. Our findings indicated that the choice of resampling voxel size significantly influenced both feature stability and prognostic performance, with 1×1×1 mm³ yielding optimal results across most evaluations, with 1×1×1 mm³ producing good performance in most tests. In subsequent studies addressing various brain tumors, including glioblastoma, pediatric low-grade glioma, and medulloblastoma [3-6], we consistently assessed the effect of voxel size when developing our radiomics models, ultimately adopting the 1×1×1 mm³ voxel size across all tasks. In this study, we also chose to use 1×1×1 mm³ and the results demonstrated the accuracy of our model. Moreover, as suggested by the IBSI recommendations [1], we used trilinear interpolation for voxel resampling, as this method does not lead to blockiness and out-of-range intensity values compared with other methods such as nearest neighbour or tricubic interpolations.

References:

- [1] https://ibsi.readthedocs.io/en/latest/02_Image_processing.html#interpolation-algorithms
- [2] Qihua Li, et al. A Fully-Automatic Multiparametric Radiomics Model: Towards Reproducible and Prognostic Imaging Signature for Prediction of Overall Survival in Glioblastoma Multiforme. Scientific Reports, 7:14331, 2017. <https://doi.org/10.1038/s41598-017-14753-7>
- [3] Zhi-Cheng Li, et al. Multiregional Radiomics Profiling from Multiparametric MRI: Identifying an Imaging Predictor of IDH1 Mutation Status in Glioblastoma. Cancer Medicine. 7(12): 5999-6009. 2018. <https://doi.org/10.1002/cam4.1863>
- [4] Qiuchang Sun, et al. Biologic pathways underlying prognostic radiomics phenotypes from paired MRI and RNA sequencing in glioblastoma. Radiology. 301(3): 654-663, 2021. <https://doi.org/10.1148/radiol.2021203281>
- [5] JingYan et al. Incremental prognostic value and underlying biological pathways

of radiomics patterns in medulloblastoma. EBioMedicine. 61: 103093, 2020. <https://doi.org/10.1016/j.ebiom.2020.103093>

[6] Dongling Pei, et al. Radiomics Features from Multiparametric Magnetic Resonance Imaging Predict Molecular Subgroups of Pediatric Low-grade Gliomas. BMC Cancer. 23: 848. 2023. <https://doi.org/10.1186/s12885-023-11338-8>

[7] Yan J, Zhang B, Zhang S, Cheng J, Liu X, Wang W, Dong Y, Zhang L, Mo X, Chen Q, Fang J, Wang F, Tian J, Zhang S, Zhang Z. Quantitative MRI-based radiomics for noninvasively predicting molecular subtypes and survival in glioma patients. NPJ Precis Oncol. 2021 Jul 26;5(1):72. <https://doi.org/10.1038/s41698-021-00205-z> PMID: 34312469; PMCID: PMC8313682.

14. The manuscript mentions the use of PyRadiomics for feature extraction, but there are no details about the specific parameters or version used. Providing the PyRadiomics parameters file would significantly enhance reproducibility.

Response: According to the reviewer's suggestion, we provide the .yaml format parameter file as supplement to enhance the reproducibility of feature extraction (<https://github.com/Zaoqu-Liu/MOFS/blob/master/PyRadiomics%20parameters.yaml>).

Unclassified samples:

15. Some patients were not assigned to a MOFS subtype, and a core set of 116 cases was used. What are the reasons for the failure to classify some patients? Please elaborate on how these cases differed from the classified ones.

Response: The exclusion of certain patients from MOFS subtype assignment was based on their silhouette coefficient values. Specifically, patients with silhouette coefficients below 0.4 were excluded from the final core set of 116 cases. This threshold was chosen to ensure that only samples with robust and well-defined cluster memberships were retained, as silhouette coefficients below 0.4 indicate ambiguous or weak clustering (<https://github.com/Zaoqu-Liu/MOFS>; <https://github.com/Zaoqu-Liu/MOFS/blob/master/MOFSR/R/RunMOFS.R>).

The silhouette coefficient evaluates how similar a sample is to its assigned cluster compared to other clusters, providing a measure of clustering quality. Samples with low coefficients typically lie near cluster boundaries or do not align well with any specific cluster, suggesting potential outliers or inconsistencies in the data for those patients.

Further examination revealed that the excluded samples often exhibited higher variability in multimodal features compared to the core set, which could lead to less stable clustering assignments. These cases might represent unique biological variations or technical artifacts that deviate from the main patterns captured by the MOFS subtypes.

By focusing on the core set of 116 cases, we ensured that our analysis and subtype definitions were based on a consistent and reliable sample group, reducing potential noise and enhancing the interpretability of the results. This approach aligns with best practices in clustering analysis, where ambiguous samples are typically excluded to maintain the robustness and clarity of the derived clusters.

Deep-neural network to predict MOFS: while I think the strength of the paper lies in the toll to predict MOFS types only based on imaging parameters:

16.Accuracy Drop: The Deep-Neural Network trained on the FAHZZU1 dataset exhibited a drop in accuracy (0.825) when applied to FAHZZU2. Was this drop due to reduced generalizability, overrepresentation of radiological features, or another factor? Please clarify.

Response: The observed drop in accuracy for the MRI-based deep neural network (DNN) classifier can be attributed to several factors, which we carefully examined and have now clarified in the revised manuscript.

The ground truth labels for training and validating the MRI classifier were derived from the FAHZZU1 cohort's multimodal clustering. Specifically, the training process utilized 70% of the FAHZZU1 samples, while the remaining 30% served as the validation set. This ensured that both training and validation used labels directly linked to the comprehensive biological information available in the multimodal data. As outlined in the revised methods section, this approach leveraged the synergy of transcriptomics, proteomics, radiological, and pathological data to assign accurate MOFS subtypes.

In contrast, the FAHZZU2 cohort lacks the full multimodal data of FAHZZU1 (as illustrated in Figure S1B), and the MOFS labels for FAHZZU2 were predicted using the RNA-seq-based ensemble classifier. While this allowed us to utilize FAHZZU2 for external validation, it introduced a potential source of variability. RNA-seq-derived labels, while highly predictive, lack the holistic integration of multimodal data and may introduce subtle inconsistencies compared to the original clustering-based labels. This methodological difference likely explains the observed drop in AUC values when applying the MRI classifier to FAHZZU2, as shown in Figure S16B.

Importantly, the FAHZZU2 validation was intentionally included to evaluate the generalizability of the MRI classifier across datasets with differing label derivations. While the accuracy drop highlights the challenges of using single-modality-derived labels, the classifier's ability to maintain a reasonable performance level in FAHZZU2 underscores its robustness. This finding also emphasizes the importance of multimodal integration, as the MOFS labels derived from RNA-seq alone lack the additional biological context provided by proteomics, radiology, and pathology in FAHZZU1.

Finally, we acknowledge that radiological features alone may not capture the full complexity of the MOFS subtypes, especially when labels are derived from RNA-seq data, which are focused on transcriptional profiles. However, this result is consistent with our overarching aim to explore the potential of imaging biomarkers for subtype classification while recognizing the complementary value of multimodal data.

These insights, along with detailed discussions on the ground truth labels, training and validation methodology, and implications of RNA-seq-derived labels, have been incorporated into the revised manuscript. We hope this addresses the reviewer's concerns and provides a transparent explanation for the observed performance differences.

17.Independent Dataset: Since the dataset used to develop the Deep-Neural Network (DNN) comes from a single institution (FAHZZU), I recommend testing the model on an independent dataset to validate its generalizability. Utilizing an open-source dataset, such as the UCSF glioma dataset, would provide a more robust evaluation and ensure that the model performs well across different populations. Has external validation been considered in your analysis?

Response: Thank you for your insightful suggestion regarding the use of an independent dataset, such as the UCSF glioma dataset, to validate the generalizability of our Deep-Neural Network (DNN) model. External validation is indeed crucial for assessing model robustness across diverse populations. However, after carefully evaluating the UCSF glioma dataset (<https://www.cancerimagingarchive.net/collection/ucsf-pdgm/>), we identified several limitations that restrict its applicability to our study.

First, the UCSF dataset includes MRI modalities such as T1, T2, FLAIR, and DWI, but lacks the ADC modality, which was a critical component of our radiological feature extraction pipeline. Our study specifically used ADC data rather than raw DWI images due to the additional quantitative information ADC provides, which better aligns with our feature engineering process. Consequently, the absence of ADC data in the UCSF dataset presents a significant challenge in replicating our methodology and fully leveraging our trained DNN model.

Second, the UCSF glioma dataset does not include RNA-seq data, which is essential for deriving MOFS labels using either the multimodal clustering approach or the RNA-seq-based ensemble classifier. The lack of RNA-seq data prevents us from retroactively assigning MOFS subtypes to UCSF cases, making it impossible to establish the ground truth labels needed for validation. As MOFS labels represent a biologically integrated classification system, the absence of transcriptomic information limits the dataset's utility for validating our imaging-based predictions.

Given these constraints, we acknowledge that the absence of multimodal datasets with comprehensive MRI and RNA-seq data poses a limitation for external validation in this study. We are actively exploring collaborative opportunities to access or generate such datasets in future research. Furthermore, we emphasize that the FAHZZU dataset was internally stratified and validated to ensure rigorous evaluation of our DNN model. Nevertheless, we agree that external validation remains an important direction to enhance the generalizability and clinical applicability of our findings, and we are committed to addressing this in subsequent studies.

We have incorporated these discussions into the manuscript to provide a transparent assessment of this limitation and to highlight potential future directions. Thank you again for raising this critical point.

Code and Reproducibility:

18.*Software & Code: I strongly encourage the authors to share the code used for the analysis, including specific versions of R, Python, or other software packages. This level of transparency will enhance reproducibility and allow other researchers to*

validate the findings.

Response: Thank you for your suggestion to enhance transparency and reproducibility by sharing the code and software used in our analysis. We agree that making our computational workflow accessible is essential for enabling other researchers to validate and build upon our findings. To this end, we have uploaded the complete set of analysis scripts to a public GitHub repository: <https://github.com/Zaoqu-Liu/MOFS>. This repository provides detailed implementations of the entire pipeline, including the clustering and classification steps, as well as the tools used for result interpretation.

Specifically, the repository includes the implementations of the 11 intermediate fusion algorithms used for multimodal clustering, the methods for calculating CPI and GAP statistics, and the full pipeline for COCA clustering. The steps for binary transformation, Jaccard distance calculation, and clustering of clustering analysis are also thoroughly documented. Additionally, we provide the complete workflow for developing and validating the MRI-based deep neural network classifier, including hyperparameter optimization and performance evaluation.

To ensure reproducibility, we have included detailed documentation of the computational environment, specifying the versions of R (v4.2.2) and other software packages such as ConsensusClusterPlus, mogsa, and neuralnet. These details allow researchers to replicate the exact settings used in our study and apply the methodology to their datasets.

We believe these resources will facilitate the adoption of the MOFS framework and foster further exploration of its applications in glioblastoma research. These details have been explicitly referenced in the manuscript to guide researchers to the appropriate resources. Thank you again for highlighting this important aspect of reproducibility and open science.

Conclusion

In summary, while the manuscript presents valuable and relevant findings regarding MOFS subtypes in IDH-wildtype gliomas, it lacks sufficient methodological detail to ensure reproducibility. Addressing these specific concerns will greatly enhance the quality of this work and ensure it meets the standards of transparency and rigor expected in the field. By providing additional details on data preprocessing, algorithm parameters, and software versions, the authors can significantly strengthen the manuscript and facilitate future research based on these promising findings.

Response: Thank you for your constructive feedback highlighting the need for enhanced methodological transparency and detail. We appreciate your acknowledgment of the value and relevance of our findings regarding MOFS subtypes in IDH-wildtype gliomas. We fully agree that ensuring reproducibility is paramount to advancing research in this field and have taken significant steps to address these concerns.

To provide greater methodological clarity, we have added comprehensive details on data preprocessing, feature selection, algorithm parameters, and clustering processes to the manuscript. Specifically, we now include descriptions of

normalization and scaling procedures tailored to each data modality, detailed explanations of parameter settings for the 11 intermediate fusion algorithms, and the computational steps for COCA and Jaccard matrix calculations. These updates are supported by references to publicly available resources, such as our GitHub repository, which contains all analysis scripts, parameter configurations, and specific software versions. This repository also includes documentation for reproducing key analyses, from clustering to the development and validation of the MRI-based classifier.

Furthermore, we have discussed the limitations of certain external datasets, clarifying challenges in cross-cohort validations and the generalizability of the MRI model. These considerations reflect our commitment to maintaining transparency while acknowledging the inherent complexities of multimodal data integration in glioblastoma research.

We believe these enhancements will address the concerns raised and significantly strengthen the manuscript's quality, ensuring it meets the high standards of transparency and rigor expected in the field. Thank you again for your insightful comments, which have guided us in improving both the clarity and reproducibility of our study.

Please refer to our GitHub repository for detailed information, including tool versions and code implementation: <https://github.com/Zaoqu-Liu/MOFS>.

Dear Reviewers,

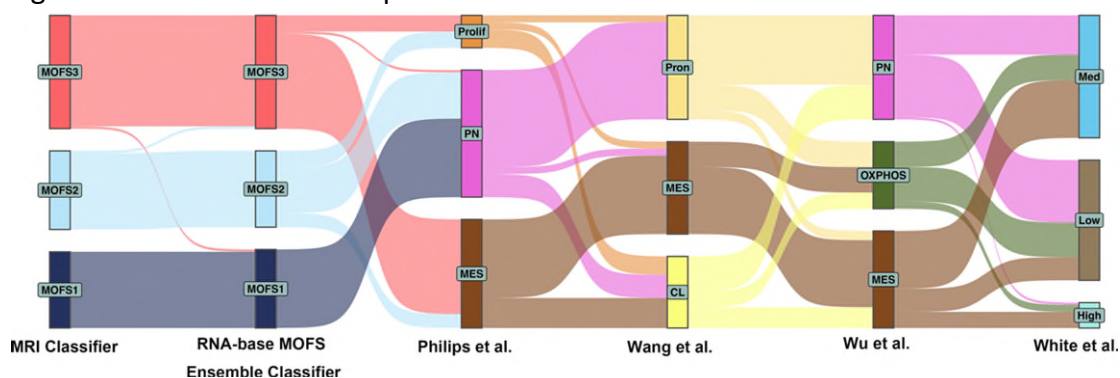
Thanks very much for taking your time to review this manuscript. We really appreciate all your comments and suggestions! Please find our itemized responses below and our revisions in the re-submitted files.

Thanks again!

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks on code availability):

I thank the authors for addressing my comments. There is still an issue that is still not clear and is rather confusing. Regarding the alignment of the MOFS clusters with established such as Wang et al. classes – the proliferative MOFS2 appears to align with the proneural classifications in Figure 3C. However, this data conflicts with that shown in figure S14B, where the RNA or MRI MOFS classifiers results demonstrate that MOFS2 aligns with Wang's mesenchymal classification, and MOFS3 align with Wang's proneural classification, as opposed to Wang's mesenchymal class as shown in figure 3C. This issue is partially addressed in the text, as the authors highlight the importance of having multiple modalities for accurate classification and prognostication, and the comments on loss of accuracy with the DNN algorithm. The issue remains that the two figures are quite different, and explanation is needed. It is also noteworthy that in their new data provided in figure 5I, MOFS2 and MOFS3 H-stroma appear to behave similarly. However, the statistics to support that should be included. This point is significant and should be emphasized.



Response: Thank you for your valuable comment. We apologize for the confusion caused by the misrepresentation of the data in Figure S14B. The figure has now been corrected, and we have updated it to accurately reflect the alignment of the MOFS subtypes. As shown in the revised figure, we clarify that the data in Figure 3C is correct, and the misalignment in Figure S14B has been addressed. Additionally, we have included the code to reproduce the external subtyping methods into the updated MOFSR package, which can be accessed at: <https://github.com/Zaoqu-Liu>.

Additionally, we appreciate your suggestion to include statistical support for the observation in Figure 5I. In response, we have now included the relevant statistical analysis in the main text, which highlights the statistical comparison and further

validates the similarity between these groups.

Thank you again for your understanding and constructive feedback.

Reviewer #3 (Remarks to the Author):

Thanks for the authors for taking into consideration all the comments I initially raised. I appreciate their understanding of transparency and benchmarking with previous work as a way to further test the capacity of the proposed new classification method for primarily diagnosed glioblastomas.

I consider the authors has satisfactorily addressed all my comments.

Just I noted that in response6-8 to Reviewer 5 (not this reviewer) there is reference to "(e.g. PMID:123 and PMID:456), so correct references should be added.

I also completely understand the limitations of the work and recognize that publication will increase interest and contribute to external validation of the model.

Overall, I found the article very interesting and recommend it for publication.

Response: Thank you very much for your thoughtful feedback and for acknowledging our efforts in addressing the comments you raised. We are pleased to hear that you find the article interesting and that the new classification method for glioblastomas has been well-received.

Regarding your observation about the references in the response to Reviewer 5, we have now corrected this and properly inserted the correct references into the manuscript. We appreciate your attention to detail in this matter.

Once again, thank you for your constructive feedback, and we are grateful for your recommendation for publication.

Reviewer #3 (Remarks on code availability):

I reviewed the content of the link describing the code with the readme documentation and the corresponding manual. The codes are well organized and structured with examples lines. I have not installed the code myself and I have not run it.

Response: Thank you for your thoughtful review of the code and the corresponding documentation. We would like to inform you that we have recently updated the code to include an environment configuration function, which allows users to easily set up the environment with a single command. This should streamline the process and ensure the code runs smoothly. If you or any other reviewer would like further assistance with installation or running the updated code, please feel free to reach out. We are happy to provide any additional guidance.

Thank you again for your valuable feedback.

Reviewer #5 (Remarks to the Author):

Dear Authors,

Thank you for your thorough responses to my comments and for the significant efforts made to address the concerns I raised. After reviewing your responses and the updates to the manuscript, I believe that you have adequately addressed the points of concern.

Response: Thank you for your kind feedback. We appreciate your thorough review and are glad to hear that the updates to the manuscript have addressed your concerns. We truly appreciate your time and consideration.

External Validation:

Your explanation regarding the challenges of using an independent dataset, such as the UCSF glioma dataset, is clear and reasonable. The detailed discussion of the absence of ADC and RNA-seq data in the UCSF dataset effectively explains why external validation using this dataset was not feasible. I appreciate your transparency in acknowledging this limitation and your proactive approach to exploring collaborative opportunities for developing multimodal datasets in future research. The inclusion of this discussion in the manuscript provides an honest assessment of the current limitations while underscoring the importance of cross-cohort validations in future studies.

Response: We appreciate your understanding of the absence of ADC and RNA-seq data in the dataset, which made external validation with this dataset unfeasible. We also value your acknowledgment of our transparency in addressing this limitation and our proactive approach to exploring collaborative opportunities for future multimodal dataset development. We believe that cross-cohort validation will be crucial for future studies, and we will continue to prioritize this in our ongoing and future work. Thank you again for your insightful feedback.

Code and Reproducibility:

I commend you for making the complete analysis pipeline, along with detailed documentation, available through a public GitHub repository. This initiative is crucial for transparency and reproducibility, as it allows other researchers to replicate your findings and apply your methods to their datasets. The inclusion of software versions, parameter settings, and user guides further enhances the usability of the MOFS framework, making it accessible to a broader audience in the research community.

Response: We are glad to hear that you appreciate the availability of the analysis pipeline and documentation on GitHub. We hope that providing these resources will help other researchers replicate our findings and use the MOFS framework in their own work. Thank you again for your support.

Methodological Transparency:

I appreciate the additional details you provided regarding data preprocessing, feature selection, clustering methodologies, and algorithm parameters. The expanded

descriptions of the computational steps and parameter configurations significantly enhance the clarity and reproducibility of the analysis pipeline. These updates ensure that the methodological aspects of the study are robust and well-documented.

Response: Thank you for your positive feedback.

Overall, while the limitation of not being able to validate using an external open-source dataset is acknowledged, the steps you have taken—particularly the release of the code—provide other researchers the opportunity to test your methods on their own datasets. This contribution strengthens the reproducibility and applicability of your framework in glioblastoma research.

Thank you for your responsiveness and for the improvements made to the manuscript. I believe the work now represents a valuable contribution to the field.

Response: We truly appreciate your constructive feedback and the time you dedicated to reviewing the manuscript.

Reviewer #5 (Remarks on code availability):

I reviewed the code provided in the GitHub repository, and it appears well-structured and understandable. The repository includes a README file with detailed instructions for installing and running the application. Furthermore, the inclusion of a comprehensive user manual adds substantial value by providing clear guidance on the various functions and workflows implemented in the MOFS framework.

The code is meticulously documented, with specific sections dedicated to the clustering, classification, and feature selection processes. These details are highly beneficial for understanding how the methods were implemented and applied in the analysis. The repository also specifies the versions of R and libraries used, ensuring that users have the necessary information to replicate the computational environment.

However, I did not attempt to install or run the code myself, primarily due to the inclusion of numerous highly specific libraries and dependencies, some of which may have compatibility issues depending on the user's computational setup. This reliance on specialized libraries, while essential for the advanced analyses performed in the study, could limit the usability of the code for some researchers without extensive computational expertise or the exact required environment.

In summary, while the code seems to be a well-prepared resource for the research community, the complexity of its dependencies may present challenges for some users. That said, the availability of clear documentation and a thorough manual should help mitigate these challenges for those motivated to replicate the study.

Response: Thank you for your thoughtful review of the code and the corresponding documentation. We would like to inform you that we have recently updated the code

to include an environment configuration function, which allows users to easily set up the environment with a single command. This should streamline the process and ensure the code runs smoothly. If you or any other reviewer would like further assistance with installation or running the updated code, please feel free to reach out. We are happy to provide any additional guidance.

Thank you again for your valuable feedback.