nature portfolio

Peer Review File

Confounding factors and biases abound when predicting molecular biomarkers from histological images

Corresponding Author: Dr Muhammad Dawood

This manuscript has been previously reviewed at another journal that is not operating a transparent peer review scheme. The manuscript was considered suitable for publication without further review at Nature Biomedical Engineering.

Version 0:

Decision Letter:

Dear Dr Dawood,

Thank you again for submitting to *Nature Biomedical Engineering* your manuscript, "Buyer Beware: confounding factors and biases abound when predicting omics-based biomarkers from histological images". The manuscript has been seen by 3 experts, whose reports you will find at the end of this message.

You will see that the reviewers appreciate aspects of the work. However, they articulate concerns about the degree of support for some of the claims and about the advance that the work represents over relevant published studies, and provide useful suggestions for improvement. We hope that with substantial further effort you can address the criticisms, increase the strength of the evidence, and convince the reviewers of the merits of the study. In particular, we would expect that a revised version of the manuscript provides:

* clear discussion and clarifications of all statistical questions from the reviewers

* clear discussion of the clinical implications and any limitations of the work

* inclusion of more references and relevant discussion as suggested by Reviewer 1.

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

Please follow the following recommendations:

*Please submit your revised manuscript in a word doc or LaTeX format.

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

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

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

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

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

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

We expect that you will be able to resubmit the manuscript within 25 weeks of receiving this message. If this is the case, you will be protected against potential scooping. Otherwise, we will be happy to consider a revised manuscript as long as the significance of the work is not compromised by work published elsewhere or accepted for publication at *Nature Biomedical Engineering*.

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

Best wishes,

Barbara Cheifet
Editor
Nature Biomedical Engineering

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

Explanation of Study

The manuscript by Dawood et al, titled, "Buyer Beware: confounding factors and biases abound when predicting omics-based biomarkers from histological images" examines, in detail, pitfalls associated with using basic (H&E stains) whole digital slide images from various registries to determine the robustness of relatioships among WSI findings and specific biomarkers (e.g. mutations, protein expression (ER/PR), etc.) without accounting for the broad spectrum of other classifiers that can potentially influence various predictive models. The authors trained SlideGraph and CLAM to predict specific clinical biomarkers of interest using both deep features and self-supervised features. The authors then examined 3 factors: 1) bias due to interdependency amng biomarkers in the training set; 2) bias due to histopathological grades of the tumors; and 3) bias due to tumor mutational burden (TMB) of a cancer patient.

Brief Summary of Results

Using mostly appropriate methodology, the authors found: 1) interdependencies exist among molecular factors with respect to their influence on digital images; 2) predictive accuracy of WSI-based models varies across patient subgroups with different histopathological grades and mutational burdens; and 3) a need exists to re-visit model training protocols to consider biomarkers interdependencies at all stage from problem definition fo usage guidelines.

Major Technical Questions or Criticisms

• In general, the message provided by this manuscript is an important, perhaps even critical, one: predictive models based on processing and ML from H&E-stained WSI need to be viewed with some skepticism. The authors support this contention using a robust data analysis from existing publically-available resources. However, herein also lies the main criticism of the manuscript: more work is needed to provide a balanced perspective using different modeling approaches, especially those used by other authors who claim acceptable results when examining multi-parameter modeling. For example, McCaw et al (McCaw et al. Machine learning enabled prediction of digital biomarkers from whole slide histopathology images. doi: https://doi.org/10.1101/2024.01.06.24300926) have seen robust prediction models that take into account the multiple variables, including molecular feeder data. Other examples exist. It is critical, therefore, for the authors to re-review the literature and determine if there results are: a) unique to their use of commercially available modeling approaches used by them; or b) a true problem with modeling, in general, and other fundamental advancements are needed to advance the field. This reviewer recognizes that the field is advancing very rapidly making it difficult to keep up with the broad spectrum of literature avaiable. However, this criticism/suggestion is critical to address to determine the relative impact on the field.
• The authors conclude that a key limitation in the assessment is the generalisability of histology-based predictors. They suggest that if the association between histology grade and receptor status weakens in different data sets, model performance may become erratic. Similar to the above remarks, the authors need to determine whether this perceived limitation is based on their specific modeling approach or is a generalizable problem that limits the potential for eventual success using a different modeling approach to multiparameter data.

Minor Technical Questions or Criticisms
• The authors should discuss the practical applicability of multiparameter modeling with (lees so) or without (more so) flaws. For example, despite excellent modeling prediction, one would wonder about the clinical applicability of the H&E WSI modeling approach when the actual tests (e.g ER receptor staining, mutational profile of tumors) that provide direct data are readily available? For example, it seems doubtful that oncologists would accept an ER receptor positivity modeling prediction accuracy of 90% when one could actually conduct the standardized and widely accepted IHC approach. Why accept 90% when 100% is easily attainable. The authors should address this practical issue. In essence, where do these modeling approaches fit into the real world? Would it best fit for front line clinical management of individual patients or more

suited to ptoential screening studies for pharma application of newer agents, etc.

Any missing citations to relevant literature
See above criticism with specific reference. However, more references are available that should be thoroughly investigated by the authors.

Any stylistic issues or recommendations
None, other than more thorough evaluation of the literature and algorithmic approaches to problem solving

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

This paper presents two main contributions:

An in-depth analysis of covariant factors aiming to understand interdependencies among biomarkers. This is an observational study that explores the co-occurrence and dependencies between biomarkers across samples.

An investigation of two deep learning methods for predicting these biomarker interdependencies from histopathological images. This, too, is observational, focusing on how variable imputation techniques perform in the presence of co-occurring biomarkers.

The results are promising, and the validation partially supports the authors' claims. However, several important issues should be addressed to strengthen the study:

Heterogeneity in study design: The analysis spans pan-tumor populations with genetic and mutational variability in both training and testing cohorts. The authors assume that statistical behavior across these diverse populations is sufficiently consistent to justify conclusions about co-mutation predictability. However, this assumption is not rigorously validated. Differences in statistical behavior and predictability may be driven more by population heterogeneity than by underlying biological mechanisms. Some mutations, although measured routinely, may not be relevant for all tumor phenotypes, raising concerns about the generalizability of findings.

Limited depth in the deep learning component: The second contribution lacks novelty. The paper essentially concludes that deep learning models do not perform/generalize equally well for predicting combinations of mutations from histopathological images. Although an AUC of 0.7 may appear acceptable from a research standpoint, performance is inconsistent and often falls below clinically actionable thresholds. More importantly, the study misses the opportunity to incorporate the insights from biomarker dependencies (explored in the first part) into the modeling design. Questions regarding the considered architectures could be also raised, and weather limitation of performance is due to the complexity of the problem; that is inherit to the number of mutations to be predicted over a constant number of training samples in each sub-category.

Unclear clinical relevance: The clinical implications of the findings are not clearly articulated. The paper would benefit from a stronger discussion on how the results could inform clinical trial design, improve patient stratification, or guide therapeutic decision-making. While some claims are made about potential utility, they are not substantiated with actionable insights or pathways for clinical translation. The study is definitely interesting and worth further pursuing but at this stage it remains more a theoretical investigation that a concrete actionable outcome that could be used in the clinical workflow.

Recommendations for improvement:

1/ Refine the study design to better isolate the effects of tumor phenotype from population heterogeneity, ensuring that observed patterns in prediction performance are biologically meaningful and not confounded. Performing the study in a consistent sub-population with biological relevance could be an additional experiment.
2/ Derive actionable clinical insights that could assist clinicians in identifying relevant phenotypes or patient subgroups and inform more effective care pathways or trial designs. This is not present in the paper.
3/ Enhance the deep learning component by developing methods that incorporate biomarker co-dependencies directly into the model architecture or training strategy, thereby improving both prediction performance and clinical relevance.

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

This is an excellent paper. It is nice to see a critical examination of AI where fundamental issues are overlooked amidst the hype. The results presented clearly show that the prevalent thinking on biomarker prediction is wrong. If some issues are addressed this paper should be required reading for anyone working in this field.

Major comments:
-The statistical testing procedures are not clearly communicated. It seems like a diagram illustrating the predictor and stratifier variables, and the permutation testing procedure would be helpful since readers without a bioinformatics background will not be familiar with this approach (most readers I guess). Table 1 defines the procedure but is difficult to

follow, and a diagram would help.
-Page 4 paragraph 2 - How the swap between predictor and stratifier is used is not clear.
-Section 2.6 - what is "sufficiently high level of accuracy"? It seems like an extremely high accuracy is necessary to alter how these cases are handled clinically (e.g. treatment or triaging use of molecular testing) but this is ignored in papers on biomarker prediction. An honest answer here would be a welcome dose of reality.
-Engineers and computer scientists are probably the most important audience for this result, and they may be unfamiliar with the concept of epistasis. They would benefit from a brief explanation of why mutations co-occur or are mutually exclusive, their functional relationship, at an appropriate level of detail for them.

Minor comments:
-Should mention that the co-occurrence or mutual exclusivity of genetic alterations has been investigated extensively, including as part of TCGA studies. This is often the main point of "waterfall" plots included in these papers characterizing the genetic alterations in a cancer type. Several bioinformatics methods have been developed to examine these patterns to better understand functional relationships between genes. That doesn't take away from the novelty of this study at all - it would enhance the impact to make this link.
-Figure S2 needs its own legend defining the colors.
-The bar plots are a little crowded and hard to read.
-The choice of AUC as the metric of evaluation should be justified. See https://pmc.ncbi.nlm.nih.gov/articles/PMC9006654/

Version 1:

Decision Letter:

Dear Dr. Dawood,

My apologies for our slow decision on your manuscript. We were waiting for a second reviewer but decided to move forward based on the review that came in.

Thank you for submitting your revised manuscript "Buyer Beware: confounding factors and biases abound when predicting omics-based biomarkers from histological images" (NBME-24-2796A). Having consulted with one of the original reviewers, I am pleased to write that we shall be happy to publish the manuscript in Nature Biomedical Engineering, pending minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Biomedical Engineering offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't.** Failure to state your preference will result in delays in accepting your manuscript for publication.

Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Author names using non-Roman characters
Nature Portfolio journals can support presentation of author names using non-Roman characters in the HTML version of the article. If you wish to, please include author names in parentheses after the Roman-character spelling; see example online here. Currently supported scripts are: Arabic, Chinese, Cyrillic, Devanagari, Greek, Hebrew, Hangul, Japanese and Persian. You will be asked to verify the rendering is correct at proof stage.

Sincerely,
Rita

Rita Strack, Ph.D.
Chief Editor
Nature Biomedical Engineering

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

The revisions have addressed my concerns. Description of is significantly improved, particularly in terms of readability and referencing relevant works from other fields.

Version 2:

Decision Letter:

Dear Dr Dawood,

I am happy to inform you that your manuscript, "Confounding factors and biases abound when predicting molecular biomarkers from histological images", has now been accepted for publication in *Nature Biomedical Engineering*.

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

Our Articles are published about 40-60 days after the acceptance date (we recommend that you inform your institutional press office of this timeframe), and you will be notified of the actual publication date a few days in advance. Articles can be published any working day of the week, and are pushed live shortly after 10 am London time.

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

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

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

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

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

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

Best wishes,
Rita

Rita Strack, Ph.D.
Chief Editor
Nature Biomedical Engineering

Response to Editors/Reviewers' Comments

nBME-24-2796-T: Buyer Beware: confounding factors and biases abound when predicting omics-based biomarkers from histological images

The authors would like to thank the reviewers and the editors for their time and insightful feedback. We have carefully addressed all the comments and revised the manuscript accordingly. A point-by-point response to the comments is provided below (our response in blue). Changes made in the manuscript in response to the comments/suggestions are shown in green font.

Editor Comments:

E1.l: Clear discussion and clarifications of all statistical questions from the reviewers
Response to E1.1: We have addressed all comments from the reviewers, including the statistical questions raised in **R3.1** and **R3.2**. Please refer to our detailed responses to those items for clarification.

E2.1: Clear discussion of the clinical implications and any limitations of the work
Response to E1.2: In response to this comment, we have revised the Abstract and Discussion to make these points explicit, and we now cross-reference the supporting analyses in the Results (grade/TMB stratification, baseline grade predictors, and foundation-model features).

Clinical implications have now been made explicit as outlined below:

1. **CPath approaches are not a stand-alone replacement for molecular testing**: Because performance varies substantially across patient subgroups defined by co-dependent variables (e.g., grade, TMB, and other biomarkers), WSI-based predictors, at their current level, may not be used as a substitute for genomic assays; they are better positioned for triage/screening or supplementary decision support with confirmatory testing as needed. For example, our results demonstrate that model predictions for MSI status from whole slide images (WSIs) are not independent of BRAF mutation status, with subgroup analyses showing substantial AUROC drops when stratifying by BRAF and a similar pattern is noted when predicting BRAF status while stratifying by MSI status (Figures 2, 4 and S7). This reflects the well-established frequent co-occurrence (biological relationship) between the two biomarkers: MSI-high colorectal cancers frequently harbour BRAF V600E mutations. This cooccurrence may influence the generation of predictors, picking up both signals. While MSI-stable tumours more rarely have BRAF mutations and the biological implication and aetiology are different. Crucially, however, MSI-H and BRAF mutations have distinct therapeutic implications. MSI-H is a strong predictor of response to immune checkpoint inhibitors such as pembrolizumab or nivolumab, whereas BRAF V600E mutations are targeted using BRAF and MEK inhibitors in combination with EGFR blockade. Combinations of immunotherapy and BRAF inhibitors are currently being tested for the double mutant. A model that cannot disentangle MSI-H from BRAF status may therefore achieve high aggregate AUROC but lacks clinical utility, since confusing the two would

misguide treatment selection. This example underscores the broader need for bias-aware evaluation: predictors must be assessed not only for overall accuracy but also for their ability to distinguish correlated biomarkers with divergent therapeutic pathways (Morris et al., 2025). We now state this directly in the **Abstract** and expand in the **Discussion**.

2. **Bias-aware evaluation is required even with external validation:** We show that external AUROC can rise or fall due to cohort-specific associations (e.g., ER–grade coupling), not true biomarker learning; therefore, deployments should report grade- and TMB-stratified metrics and subgroup calibration, not only aggregate AUROC. We added this guidance to the **Discussion** and pointed to the empirical evidence in Figure 4, Figure 5, and Figure 7.

3. **Design implications for studies and trials:** The findings of our work are also applicable to clinical trial design, particularly when studying the molecular landscape of a disease phenotype or drug response in patients who are positive or negative for a specific molecular factor. In designing such studies, it is important to include, where possible, samples in which the biomarker of interest varies independently or is not strongly correlated with other biomarkers. If enrolling such samples is not feasible, trial designers should at least acknowledge and report limitations of the study due to biomarker co-dependency. For instance, when studying a biomarker's role in disease progression or treatment response, it is important to account for interdependent biomarkers and assess effects within relevant subgroups. We believe that incorporating stratification analysis into study design, whether for model training, phenotype discovery, or clinical trials, can help uncover clearer biological signals. Although this may add complexity to cohort selection, it is important for reducing confounding and advancing precision medicine. We now provide concrete guidance to: (i) preserve variation in the target biomarker relative to correlated variables during enrolment; (ii) pre-specify stratification factors (e.g., grade, TMB, site, key co-mutations) and conduct prospective subgroup analyses; (iii) include a dependency-aware analysis plan (e.g., stratified permutation tests, subgroup CIs, comparison to simple clinical baselines such as grade-only models); and (iv) conduct per-stratum power calculations rather than only aggregate targets.

4. **Appropriate current use:** We clarify near-term use cases of WSI-based predictive models: hypothesis generation, preclinical/translational screening, and resource-constrained settings with explicit safeguards and clinician oversight (confirmatory testing, conservative thresholds, and post-deployment monitoring).

5. **Reporting Requirements:** We also suggest that ML predictors should report population characteristics as well as co-occurrence patterns within their training and validation/test cohorts so that clinicians and other users can understand the limitations of these methods before their use.

Changes in Manuscript highlighting clinical implications:

1. **Abstract**: now states the triage/supplementary role and cautions against stand-alone use; mentions subgroup variability and need for bias reporting.

2. **Discussion**: adds practical guidance for deployment (bias-aware evaluation; trial design) and clarifies the limits of external validation.
3. **Results**: we point reviewers to the evidence underpinning these implications such as grade/TMB confounding and baseline-grade comparisons (see **sections 2.4 – 2.7**).

Limitations (now structured and specific):

1. **Scope of data and labels:** Our analyses are limited to H&E WSIs and WSI-level (coarse) labels; we did not evaluate IHC slides or models trained with fine-grained spatial labels (e.g., spatial omics supervision). We now say this explicitly and frame it as future work.
2. **Foundation-model representations:** In the original submission before this revision, we did not analyse the impact of biases highlighted in this paper on computational pathology foundation models. We have now included TITAN WSI-level features and trained both single-output and multi-output predictors; their behaviour is consistent with our main conclusions (i.e., still vulnerable to grade/TMB confounding), reinforcing that the issue is architectural-agnostic. See **Methods** Section 4.4.3, **Results** Sections 2.3–2.5, and Figures S5–S8, which document these additions.
3. **Generalisability estimates:** Despite analysing 8,221 patients across multiple cancers and cohorts, larger prospective studies are needed to quantify clinical impact and to develop deployment guidelines. We now state this plainly.
4. **Coverage of biomarker combinations:** We note that fully learning disentangled genotype-phenotype mappings likely requires richer datasets that better cover all combinations of biomarker labels across clinical strata; curating such datasets is challenging and remains future work.
5. **Methodological proposals:** We propose directions (causal adjustment, counterfactual augmentation, invariance/robustness objectives) but do not claim clinical validation of these strategies here; we now state that explicitly.

Pointer to where to look (for the editor/reviewer):

Abstract; Discussion (paragraphs 2, 4, 5, 6 and 8); Methods section 4.4.3, 4.5 and 4.7; Results section (grade confounding, baseline-grade comparator); Figure 8; Supplementary Figures (S3 and Figure S5–S8); and Supplementary Tables (S1–S3).

References:

Morris, Van K., et al. "Phase 1/2 trial of encorafenib, cetuximab, and nivolumab in microsatellite stable BRAFV600E metastatic colorectal cancer." *Cancer Cell* (2025).

E1.3: Inclusion of more references and relevant discussion as suggested by Reviewer 1

Response to E1.3: In response to the reviewer's comment, we included additional references in the introductory paragraph of the manuscript. For details, please see our response to **R1.4**.

Reviewer's comments:

Reviewer #1

Explanation of Study

The manuscript by Dawood et al, titled, "Buyer Beware: confounding factors and biases abound when predicting omics-based biomarkers from histological images" examines, in detail, pitfalls associated with using basic (H&E stains) whole digital slide images from various registries to determine the robustness of relationships among WSI findings and specific biomarkers (e.g. mutations, protein expression (ER/PR), etc.) without accounting for the broad spectrum of other classifiers that can potentially influence various predictive models. The authors trained SlideGraph and CLAM to predict specific clinical biomarkers of interest using both deep features and self-supervised features. The authors then examined 3 factors: 1) bias due to interdependency among biomarkers in the training set; 2) bias due to histopathological grades of the tumors; and 3) bias due to tumor mutational burden (TMB) of a cancer patient.

Brief Summary of Results

Using mostly appropriate methodology, the authors found: 1) interdependencies exist among molecular factors with respect to their influence on digital images; 2) predictive accuracy of WSI-based models varies across patient subgroups with different histopathological grades and mutational burdens; and 3) a need exists to re-visit model training protocols to consider biomarkers interdependencies at all stage from problem definition fo usage guidelines.

Major Technical Questions or Criticisms

R1.1: In general, the message provided by this manuscript is an important, perhaps even critical, one: predictive models based on processing and ML from H&E-stained WSI need to be viewed with some skepticism. The authors support this contention using a robust data analysis from existing publically-available resources. However, herein also lies the main criticism of the manuscript: more work is needed to provide a balanced perspective using different modeling approaches, especially those used by other authors who claim acceptable results when examining multi-parameter modeling. For example, McCaw et al (McCaw et al. Machine learning enabled prediction of digital biomarkers from whole slide histopathology images. doi: https://doi.org/10.1101/2024.01.06.24300926) have seen robust prediction models that take into account the multiple variables, including molecular feeder data. Other examples exist. It is critical, therefore, for the authors to re-review the literature and determine if there results are: a) unique to their use of commercially available modeling approaches used by them; or b) a true problem with modeling, in general, and other fundamental advancements are needed to advance the field. This reviewer recognizes that the field is advancing very rapidly making it difficult to keep up with the broad spectrum of literature available. However, this criticism/suggestion is critical to address to determine the relative impact on the field.

Response to R 1.1: We thank the reviewer for urging a broader perspective. We acted on this in two ways:

1. **Literature update and positioning:** We expanded the Introduction to acknowledge recent multi-parameter and foundation-model work (e.g., McCaw et al. and TITAN) and to explicitly state the gap our study addresses: prior studies (including McCaw et al.) optimise aggregate accuracy over multiple prediction parameters but typically do not test whether performance is stable across patient subgroups defined by co-dependent biomarkers or clinicopathological variables, nor do they use permutation-based bias checks. We now make this research gap explicit in the **Introduction** and **Key Contributions**. (refs now include McCaw et al. and TITAN).

2. **Additional modelling families and results:** Beyond CLAM and $SlideGraph^{\infty}$ (two representative, open-source WSI paradigms utilising foundation model patch embeddings from CTransPath and ShuffleNet), we have now included analysis of alternative modelling approaches, including multi-parameter prediction based on the feature representation from a state of the art WSI-level multimodal foundation model (TITAN), which has been pretrained on 330,000 WSIs paired with pathology reports. We used TITAN-derived features to train both single-output and multi-output models for biomarker prediction. Details about the experimental setup are provided in **Section 4.4.3** of the revised manuscript. Across both modelling paradigms, the models' predictive performance declines when stratifying by the status of co-dependent biomarkers. For example, the accuracy of the ER predictor drops significantly in CDH1-mutant cases (see **Figure S5**, **Figure S6**, and **Figure S7**). These results mirror our observations over previously assessed weakly-supervised models, such as $SlideGraph^{\infty}$, and CLAM. Together, this suggests that the limitations we report are not specific to the choice of feature representation or modelling approach but instead reflect a broader challenge: performance of models trained to predict one biomarker degrades when evaluated within strata defined by co-dependent biomarkers, grade, or TMB.

Together, the literature review and added analyses support that our observations are not a quirk of any one commercial or single-task model, but a broader issue: *when training data contain strong co-dependencies, WSI predictors tend to capture aggregated phenotypes rather than biomarker-specific signals, leading to subgroup-dependent performance.* We frame this as a data-/evaluation-level problem and provide a practical template (stratification + permutation testing) that complements accuracy-focused modelling.

In line with the suggestions from the reviewer, we have also highlighted the clinical implications of this work (please see our response to editor comments above (**E2.1**)).

In response to this comment, we updated the **Introduction** (see paragraphs 3 and 4), **Key Contributions** (bullet point 4) and added **Additional Figures** (Figure S3, Figure S5, Figure S6 and Figure S7**)** to the **Supplementary Materials**. We also revised the **Methods** (see Section 4.4.3), **Results** (see Section 2.3 and 2.4) and **Discussion** detailing the experimental setup and results of single-output and multi-output modelling approaches.

R1.2: The authors conclude that a key limitation in the assessment is the generalisability of histology-based predictors. They suggest that if the association between histology grade and receptor status weakens in different data sets, model performance may become erratic. Similar to the above remarks, the authors need to determine whether this perceived limitation is based

on their specific modeling approach or is a generalizable problem that limits the potential for eventual success using a different modeling approach to multiparameter data.

Response to R1.2: We appreciate the reviewer's focus on generalisability. Our central claim is that instability arises when the association structure between the target biomarker and correlated variables (e.g., grade, TMB, other biomarkers) differs between development and deployment cohorts or does not reflect the underlying biology. This is not unique to a specific algorithm family or even the choice of specific biomarkers, as discussed in our response to **R1.1**.

To address the reviewer's concern whether this limitation is model-specific or reflects a broader challenge, we conducted additional experiments using two modelling approaches: a single-output logistic regression model and a multi-output/multiparameter MLP, both trained on TITAN-derived features. We found similar trends for these models, like the one seen for weakly supervised models $SlideGraph^{\infty}$ and CLAM. For example, in the TCGA-UCEC cohort, the AUROC of the TP53 predictor dropped from 0.83 (all cases) to 0.77 (high-grade only) with the logistic regression model, and from 0.86 (all cases) to 0.77 (high-grade only) with the multi-output MLP. In the CPTAC-UCEC cohort, where the grade–mutation association differs, performance declined more sharply: from 0.61 (all cases) to 0.53 (high-grade only) with logistic regression, and from 0.74 (all cases) to 0.60 (high-grade only) with the multi-output MLP. The sharp drop in performance from the cross-validation dataset to the external validation dataset, observed across different modelling approaches, indicates an inherent limitation of these methods: they are sensitive to changes in association structure between biomarkers and confounding variables, which can undermine the reliability of these methods.

These additional experiments further confirm that the limitation highlighted in this work is general: *when training data contain strong co-dependencies, WSI predictors tend to learn aggregated phenotypes rather than biomarker-specific morphology, yielding subgroup-dependent performance that varies with cohort composition.* Our additions (TITAN single- and multi-output models) support this interpretation and show that multi-parameter modelling is helpful but insufficient unless paired with bias-aware evaluation and, prospectively, designs that preserve/adjust association structure.

In response to this comment, we have now updated the **Supplementary Materials** (added Figure S8) showing the influence of bias on models trained using single-output and multi-output modelling approaches. Additionally, we revised the **Results** (Section 2.5) and **Discussion** sections detailing these experimental results.

Minor Technical Questions or Criticisms

R1.3: The authors should discuss the practical applicability of multiparameter modeling with (lees so) or without (more so) flaws. For example, despite excellent modeling prediction, one would wonder about the clinical applicability of the H&E WSI modeling approach when the actual tests (e.g ER receptor staining, mutational profile of tumors) that provide direct data are readily available? For example, it seems doubtful that oncologists would accept an ER receptor positivity modeling prediction accuracy of 90% when one could actually conduct the standardized and widely accepted IHC approach. Why accept 90% when 100% is easily

attainable. The authors should address this practical issue. In essence, where do these modeling approaches fit into the real world? Would it best fit for front line clinical management of individual patients or more suited to potential screening studies for pharma application of newer agents, etc.

Response to R1.3: We agree that, in their current form, H&E-based multiparameter predictors should not replace gold-standard assays (IHC/genomics) in routine care. In the revised **Abstract** and **Discussion** (paragraph 6) sections, we explicitly position them where they add value with safeguards: (i) triage/pre-screening with mandatory confirmatory testing, (ii) tissue-limited or retrospective cohorts where additional assays are infeasible, (iii) large-scale translational/drug-development studies to prioritise expensive modalities, and (iv) hypothesis generation about morphology-molecular links. Our results show that performance can vary across subgroups and cohorts due to co-dependent variables (e.g., grade, TMB, other biomarkers), and that this instability persists across modelling families (weakly supervised and foundation-model feature baselines, single- and multi-output), indicating the limitation is not model-specific. To support safe use, we provide bias-aware evaluation guidance, i.e., reporting subgroup-stratified metrics and applying permutation-based checks and highlight cases where the simple clinicopathological baselines approach ML performance. Thus, the paper clarifies where these models fit in real workflows and contributes concrete methods to evaluate and deploy them responsibly.

R1.4: Any missing citations to relevant literature
See above criticism with specific reference. However, more references are available that should be thoroughly investigated by the authors.

Response to R1.4: We thank the reviewer for this comment. Based on the reviewer's comment, we have now added additional references in the introductory paragraph of the revised manuscript.

R1.5: Any stylistic issues or recommendations
None, other than more thorough evaluation of the literature and algorithmic approaches to problem solving.

Response to R1.5: We thank the reviewer for this comment. In the revised manuscript, we have incorporated the results of additional experiments (see **R1.1** and **R1.2**) to broaden the set of algorithms and problem modelling approaches evaluated. We have also added further references and explicitly highlighted the research gap in the Introduction Section of the revised manuscript.

Reviewer #2

This paper presents two main contributions:
An in-depth analysis of covariant factors aiming to understand interdependencies among biomarkers. This is an observational study that explores the co-occurrence and dependencies between biomarkers across samples.

An investigation of two deep learning methods for predicting these biomarker interdependencies from histopathological images. This, too, is observational, focusing on how variable imputation techniques perform in the presence of co-occurring biomarkers.

The results are promising, and the validation partially supports the authors' claims. However, several important issues should be addressed to strengthen the study:

R2.1: Heterogeneity in study design: The analysis spans pan-tumor populations with genetic and mutational variability in both training and testing cohorts. The authors assume that statistical behavior across these diverse populations is sufficiently consistent to justify conclusions about co-mutation predictability. However, this assumption is not rigorously validated. Differences in statistical behavior and predictability may be driven more by population heterogeneity than by underlying biological mechanisms. Some mutations, although measured routinely, may not be relevant for all tumor phenotypes, raising concerns about the generalizability of findings.

Response to R2.1: We agree with the reviewer that there is no guarantee of consistency of associations between biomarkers or mutations between cohorts, and our study is designed to show that current practice often ignores this fact in machine learning model development. We do not introduce a new model. Instead, we use standard, representative methods as testbeds to demonstrate a general failure mode of WSI-based prediction of molecular biomarkers. First, we estimate biomarker interdependencies within each cohort using odds ratios and Fisher's exact tests with FDR control, and we then report model performance per cohort, thereby confirming the reviewer's concern of possible differences between cohorts in specific mutations/biomarker co-occurrences. We then evaluate the performance of models trained on TCGA on CPTAC as external validation under association shift rather than as pooled pan-tumour inference. Under the prevailing training practice for WSI-based predictors that optimises aggregate performance without adjusting for co-occurrence, we show that models tend to learn combined phenotypes driven by associations present in the training cohort. We further show that when those associations differ in the test cohort, generalisability deteriorates. For example, a TP53 model trained and validated in TCGA UCEC achieves an AUROC of 0.70 in high-grade cases but falls to 0.36 on CPTAC UCEC high-grade cases. This decline appears to result from the fact that TP53 mutations are more frequent in high-grade tumours in TCGA-UCEC but less frequent in CPTAC-UCEC. The same cohort conditional behaviour appears when conditioning on TMB and on other biomarkers, and it persists across weakly supervised models and TITAN feature baselines in both single-output and multi-output settings. Our contribution is to reveal and quantify this assumption-driven failure mode and to provide a bias-aware evaluation template rather than proposing a new method. For more details, we refer the reviewer to **Section 2.2** for biomarker interdependencies, and to **Sections 2.5** and **2.7** for analysis of how shifts in these associations influence model performance.

We also agree that not all routinely measured mutations have a consistent/visible morphologic footprint across tumour types or may not be clinically important at present. We do not claim universal predictability from H&E. Rather, our key point is that even when a biomarker appears predictable within a cohort, performance is confounded by the status of other biomarkers and degrades across cohorts because models capture cohort-specific association structures (co-occurrence with other biomarkers, grade, TMB) rather than biomarker-specific morphology or specific biological mechanisms. We have now revised **Section 1.1** of the **Supplementary Materials** to briefly describe the basis for biomarker interdependencies and their phenotypic manifestation in WSIs.

R2.2: Limited depth in the deep learning component: The second contribution lacks novelty. The paper essentially concludes that deep learning models do not perform/generalize equally well for predicting combinations of mutations from histopathological images. Although an AUC of 0.7 may appear acceptable from a research standpoint, performance is inconsistent and often falls below clinically actionable thresholds. More importantly, the study misses the opportunity to incorporate the insights from biomarker dependencies (explored in the first part) into the modeling design. Questions regarding the considered architectures could be also raised, and weather limitation of performance is due to the complexity of the problem; that is inherit to the number of mutations to be predicted over a constant number of training samples in each sub-category.

Response to R2.2: We thank the reviewer for this comment. We would like to clarify that the aim of our study is not to develop a novel deep learning method to predict a combination of mutations but to systematically assess the limitations of current WSI-based approaches for biomarker prediction, with a focus on their robustness in the presence of confounding molecular and clinicopathological factors arising from associations between their labels in the training data. To this end, we deliberately covered diverse inductive biases by using an attention-MIL model (CLAM), a graph-based model ($SlideGraph^{\infty}$) with different deep features, and, in this revision, foundation-model WSI features (TITAN) in both single-label (predicting one biomarker at a time) and multi-label (predicting multiple biomarkers simultaneously) settings (see **R1.1** and **R1.2**). Across all of these, the same pattern emerges: apparent cohort-level performance drops (even for cases where AUROCs are >0.7) and becomes inconsistent once we condition on biomarker and clinicopathologic dependencies, and does not recover when training a multi-output head that can exploit training-time co-occurrences as suggested by the reviewer. This architecture-agnostic result is the intended contribution and, as pointed out by the reviewer, it reflects the complexity of the prediction task of learning disentangled representations from entangled molecular phenotypes and labels in the training data. This issue cannot be fully overcome by scaling data alone, as that would require near-exhaustive coverage of co-mutation/biomarker configurations across the biologically plausible space, which is impractical with current cohorts.

Based on this comment, we have revised the text to: (i) state that our contribution is an evaluation framework demonstrating architecture-agnostic limitations under dependency-aware assessment; (ii) add results with TITAN and multi-label training showing the same phenomenon; and (iii) temper any clinical reading by underscoring that these AUROCs are below decision-making thresholds and are presented to illuminate entanglement and sample-complexity constraints, not to advocate deployment. Specifically, we revised the **Key Contributions** (added contribution 4) and **Discussion** (see paragraph 7) sections. For results and discussion about single-output and multi-output models, please refer to our response to **R1.1** and **R1.2**.

R2.3: Unclear clinical relevance: The clinical implications of the findings are not clearly articulated. The paper would benefit from a stronger discussion on how the results could inform clinical trial design, improve patient stratification, or guide therapeutic decision-making. While some claims are made about potential utility, they are not substantiated with actionable insights or pathways for clinical translation. The study is definitely interesting and worth further

pursuing but at this stage it remains more a theoretical investigation that a concrete actionable outcome that could be used in the clinical workflow.

Response to R3.3: We thank the reviewer for this important comment. Please see our response **E2.1** for actionable insights and clinical implications of this work.

Recommendations for improvement:

R2.4: Refine the study design to better isolate the effects of tumor phenotype from population heterogeneity, ensuring that observed patterns in prediction performance are biologically meaningful and not confounded. Performing the study in a consistent sub-population with biological relevance could be an additional experiment.

Response to R2.4: We agree that isolating tumour-intrinsic morphology from population heterogeneity is essential. Our revision now makes this isolation more explicit along three axes:

1. **Demography vs biomarker status:** We quantified demographic balance between biomarker-positive and biomarker-negative groups in TCGA-BRCA (age, sex, race) and observed small standardised mean differences (SMDs) for age/sex and at most moderate SMD for race (ER), whereas grade and PAM50 showed larger SMDs across ER/PR/TP53, consistent with genuine phenotypic differences between biomarker-defined groups (**Table S2**). This points to tumour phenotype and not broad demographics as the dominant axis of heterogeneity.
2. **Within-grade modelling:** We then trained separate models within grade strata using TITAN WSI-level features with logistic regression. Within-grade AUROCs are consistently lower than pooled models (**Table S1**). For example, in TCGA-BRCA, the TP53 predictor drops from 0.84 pooled to ~0.73 across grades; ER and PR also show pronounced drops when evaluated within low/medium grades (see Fig. 5). This pattern is exactly what we would expect if pooled performance is inflated by cross-grade correlations rather than biomarker-specific morphology.
3. **Demographic restriction:** To assess whether demography alone explains these effects, we repeated the grade-stratified analysis in White patients only (n=717). The same trend persists (**Table S3**), indicating that the observed drops are not driven by population demographics. For instance, the ER predictor achieved an AUROC of 0.85 in the full White subgroup, which dropped to 0.66 in grade 1 cases. Similar patterns were observed for the other biomarkers.

These analyses confirm that the performance differences are unlikely to be driven by patient demographics alone. In response to this comment, we added Table S1, Table S2 and Table S3 into the **Supplementary Materials** and discussed these results in **Section 2.5**.

R2.5: Derive actionable clinical insights that could assist clinicians in identifying relevant phenotypes or patient subgroups and inform more effective care pathways or trial designs. This is not present in the paper.

Response to R2.5: We thank the reviewer for this important comment and refer the reviewer to our response **E2.1** on how we have revised the paper on this basis.

R2.6: Enhance the deep learning component by developing methods that incorporate biomarker co-dependencies directly into the model architecture or training strategy, thereby improving both prediction performance and clinical relevance.

Response to R2.6: We appreciate the suggestion. The aim of this paper is evaluation rather than method development. We focus on diagnosing limitations in current WSI prediction by using architecture-agnostic, bias-aware, conditional evaluation. To avoid conflating evaluation with model innovation, we relied on widely used pipelines and also trained a multilabel head that predicts all biomarkers jointly from a shared WSI representation, thus incorporating co-dependencies directly into the model training. The central findings persist in this multilabel setting: pooled metrics often reflect phenotype or TMB entanglement, and performance drops within grade and TMB strata when evaluated conditionally. We clarify this scope in the revised **Discussion Section** (see paragraph 7) and point to future work on structured multilabel learning paired with conditional evaluation, nuisance conditioned heads, and invariance across Grade by TMB environments. We believe keeping this manuscript focused on problem diagnosis and practical evaluation guidance best serves the goals of this work by prompting the community to develop more effective solutions.

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

This is an excellent paper. It is nice to see a critical examination of AI where fundamental issues are overlooked amidst the hype. The results presented clearly show that the prevalent thinking on biomarker prediction is wrong. If some issues are addressed this paper should be required reading for anyone working in this field.

Major comments:

R3.1: The statistical testing procedures are not clearly communicated.

Response to R3.1: We thank the reviewer for this comment. We have now revised **Sections 2.1** and **4.7** to clearly communicate the statistical analysis and testing procedure.

R3.2: It seems like a diagram illustrating the predictor and stratifier variables, and the permutation testing procedure would be helpful since readers without a bioinformatics background will not be familiar with this approach (most readers I guess). Table 1 defines the procedure but is difficult to follow, and a diagram would help.

Response to R3.2: We thank the reviewer for this helpful comment. In response, we have added a new figure (**Figure 8**) illustrating the workflow of the permutation testing and stratification analysis used to assess bias in WSI-based biomarker predictors. Additionally, to accommodate this figure in the main text, we have moved the original **Table 1** to the supplementary materials and renamed it as **Table S4**.

R3.3: Page 4 paragraph 2 - How the swap between predictor and stratifier is used is not clear.

Response to R3.3: We thank the reviewer for flagging this out. In the revised **Section 2.1,** we clearly documented the statistical approach. To ensure consistency, we explicitly define two sets of variables for each model: the prediction variable, which is the biomarker the model is trained to predict, and stratification variables, which are biomarkers or clinicopathological features showing significant mutual exclusivity or co-occurrence with the prediction variable and may act as confounders.

R3.4: Section 2.6 - what is "sufficiently high level of accuracy"? It seems like an extremely high accuracy is necessary to alter how these cases are handled clinically (e.g. treatment or triaging use of molecular testing) but this is ignored in papers on biomarker prediction. An honest answer here would be a welcome dose of reality.

Response to R3.4: We agree that "sufficiently high" must be defined in clinical terms and applies equally to ML models and to clinicopathological predictors such as pathologist-assigned grade. We have revised **Section 2.6** to replace this phrase with language that ties accuracy to the intended clinical action. For triage or pre-screening, accuracy must be comparable to the safety requirements of that role and stable under conditional evaluation across grade and TMB. For assay replacement or therapy-defining decisions, accuracy must be comparable to the performance of the molecular gold standard in prospective, multi-site settings with subgroup invariance and calibration. We do not prescribe a single AUROC threshold because utility depends on prevalence, pathway design, and the cost of errors. Our results indicate that both grade-only predictors and current WSI models can capture signal but do not yet meet replacement-level requirements; their appropriate near-term role is limited to triage with confirmatory testing.

R3.5: Engineers and computer scientists are probably the most important audience for this result, and they may be unfamiliar with the concept of epistasis. They would benefit from a brief explanation of why mutations co-occur or are mutually exclusive, their functional relationship, at an appropriate level of detail for them.

Response to R3.5: We thank the reviewer for this comment. We have now revised **Section 1.1** of the Supplementary Materials and included a brief explanation of the concept of epistasis, detailing why mutations co-occur or are mutually exclusive, and their functional relationships in the context of tumour evolution.

Minor comments:

R3.6: Should mention that the co-occurrence or mutual exclusivity of genetic alterations has been investigated extensively, including as part of TCGA studies. This is often the main point of "waterfall" plots included in these papers characterizing the genetic alterations in a cancer type. Several bioinformatics methods have been developed to examine these patterns to better understand functional relationships between genes. That doesn't take away from the novelty of this study at all - it would enhance the impact to make this link.

Response to R3.6: We thank the reviewer for this helpful suggestion and fully agree. Co-occurrence and mutual exclusivity of somatic alterations have been extensively characterised in TCGA and related cohorts and are often summarised via Waterfall/OncoPrint visualisations and dedicated statistical frameworks. In the revision, we now make this link explicit in the

Introduction (contribution 1) and Methods by citing representative TCGA analyses and methodology for exclusivity/co-occurrence (e.g., MEMo (PMID: 21908773), Mutex (PMID: 25887147), DISCOVER (PMID: 27986087)), and we clarify that our heatmap/LOR + Fisher's exact testing aligns with these methods. Our contribution is orthogonal: rather than discovering new exclusivity modules, we leverage these known dependency structures as stratification variables and, via permutation testing, show how they induce confounding and apparent performance in WSI-based biomarker predictors.

R3.7: Figure S2 needs its own legend defining the colors.

Response to R3.7: We thank the reviewer for this comment. In the revised manuscript, we have added a legend to Figure S2 to define the colours.

R3.8: The bar plots are a little crowded and hard to read.

Response to R3.8: We thank the reviewer for this comment. The motivation behind showing the predictive performance of different ML models across different feature representations was to provide a comprehensive summary in a single plot. We have now revised the figure caption for improved clarity.

R3.9: The choice of AUC as the metric of evaluation should be justified.
See https://pmc.ncbi.nlm.nih.gov/articles/PMC9006654/

Response to R3.9: We thank the reviewer for pointing to this editorial. We agree with the paper that AUROC, being rank-based, can mask cohort-specific score shifts and therefore should not be equated with deployable, thresholded performance on external datasets. In our study, AUROC is primarily used as a threshold-free, rank-based statistic for bias detection, as it enables subgroup evaluation and stratified permutation testing. It also allows us to maintain comparability with existing literature and align with established benchmarking practices. We explicitly avoid interpreting AUROC as a measure of clinical utility and have already shown that within-stratum AUROC degrades in ways consistent with the editorial's concern. We have added a brief clarification in the Methods to make this intent explicit (copied below)

"We report AUROC in line with prior work, but use it exclusively as a threshold-free, rank-based test statistic in subgroup/permutation analyses to diagnose confounding; it is not taken to indicate operating-point performance on external cohorts. Deployability should be assessed with operating-point metrics [1-2]."

1. *US Food and Drug Administration. "Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations." Draft Guidance. January (2025).*
2. *Klontzas, Michail E., et al. "ESR Essentials: common performance metrics in AI—practice recommendations by the European Society of Medical Imaging Informatics." European Radiology (2025): 1-13.*