nature portfolio

Peer Review File

AI allows pre-screening of FGFR3 mutational status using routine histology slides of muscle-invasive bladder cancer

Open Access This 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 the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this 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 http://creativecommons.org/licenses/by/4.0/.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author): Expert in bladder cancer clinical research, genetics, genomics, and biomarkers

Overview
Bannire et al. present their work on AI-based analysis of H&E slides and subsequent identification of FGFR3 mutated UCs. Multiple cohorts with a total of 1113 patients from Erlangen, Germany and the TCGA project were split for training and validation of the FGFR3MUT algorithm. All samples were taken at initial diagnosis. The manuscript is well structured and written. Their key findings are the algorithm's performance in different validation cohorts, including patients with MIBC and mUC.
The authors' intent is to provide pathologists with a tool to increase the pre-test probability of detecting FGFR3 mutations via PCR or NGS without the cost of missing FGFR-altered patients. FGFR-targeted therapies are currently relevant in the mUC setting and may become relevant in the localized setting. Furthermore, the use of AI for pathology-based pattern recognition is a relevant field of progress to improve sample turn-around times and cost savings. Yet, similar work has been published previously. Overall, the presented manuscript is relevant for Nat Comms readers.

However, I would like the authors to address the following concerns:

Major
- The authors chose to report their test performance primarily as AUC. However, this is not useful in the presence of class imbalance. This is the case here, given the low number of FGFR3 events in the various cohorts (8-14%). Instead, the authors should report the PPV in the main text and replace the "rule out %" with the NPV. To add more clarity, the authors should also provide the number of pts who were called FGFR3 mutated by the algorithm in each cohort. Lastly, the number of saved molecular analyses should be provided, since it's the aim of the study.
With the current results, the algorithm clearly overestimates the presence of FGFR alterations. This, together with the variable performance in the different validation cohorts, may be indicative of overfitting. To further elucidate, the author should provide the performance metrics of the training set for comparison. Ideally, the authors should use an additional independent test set.
- Local staining procedures and reagents can cause batch effects when AI is applied. The authors describe accessing FFPE slides from the TCGA project in addition to their local samples. It's unclear to me if those are physical slides or digital images of them. In the case of physical slides, the author should state whether they were stained using the setup in Erlangen. If all H&E images were generated in Erlangen, the additional independent data test set (requested above) should be stained externally.
- The frequencies of FGFR3 alterations are low compared to previous publications. The authors should use a more comprehensive assay in a representative subset of patients to assess the number of false-negative "ground truth". Furthermore, the authors should provide information on the three FGFR3 regions amplified by their SNaPshot PCR.
- The authors don't provide any information on the presence or absence of cis and histological subtypes. If they were not excluded, it should be stated how they affected performance.
- Rightfully, the authors point toward the relevance of the TAR-210 system to moving FGFR-targeted therapy to the NMIBC space. However, in this context, validation on an NMIBC cohort is key. The authors

should ideally provide this, or at least the clinical rationale for MIBC enrichment whilst excluding UTUC cases. As is, the cohorts are not ideal for either treatment landscape (local vs. systemic).
- The authors should provide information on how patients/biopsies were selected. Given that FGFR3 mutations are associated with a macroscopic growth pattern, the authors should provide, which biopsies showed a papillary vs. solid growth, and whether the algorithm performed well in both groups. Furthermore, it would be interesting to see how much the algorithm is driven by other macroscopic features of the slide. The authors report several examples of the most predictive tiles. However, more information would be helpful to assess this. The authors should state which measures were taken to avoid over-fitting.
- In the paragraph ll139-158, the authors refer to the concept of FGFR3-driven tumours. However, this feature is never defined or referenced. In particular, the authors exclude FGFR wt tumours with high FGFR3 gene expression from the group which they refer to as being FGFR3-driven. Please elucidate.
- The authors should discuss their work in the context of previous, similar studies (https://doi.org/10.1002/cam4.4044, not cited; https://doi.org/10.1016/j.euf.2021.04.007, cited).
- Paragraph ll252-262: The discussion of the failed FGFR3-TACC3 fusions is unbalanced. A Hazard Ratio of 0.49 for survival with an upper-limit of the confidence interval that barely crosses unity in a subset analysis for a small cohort cannot be described as “patients … did not formally benefit”. In contrast the HR for survival in the FGRF mutated cohort is only 0.67. For what it’s worth, objective response rates are also higher for the translocation cohort (HR 4.18 vs. 3.89). As a clinician I can confirm that responses for FGFR3-TACC3 fusion patients are frequent and clinically meaningful. Furthermore, the authors should provide further references that fusions don’t alter the morphological and/or clinical phenotype to support their hypothesis of this being at fault for poor test performance. I would also argue that results (ROC/NPV/PPV etc) should include the inability to identify fusions, as (although rarer than point mutations) these are clinically very relevant and need to be identified.
- As a comment: It is this final point that may represent a fatal flaw. If the authors are able to develop a screening platform that is sensitive enough to rule out FGFR alterations, this would represent a time and cost savings as those patients would not have to undergo further testing. Given the different prevalence (and therefore pre-test probability) of FGFR alterations in localized vs. metastatic disease I would argue the test would need separate validation in those cohorts, but the critical issue is that if FGFR fusions are excluded, this is not a trivial exclusion and then even patients who’s slides test negative by this methodology would have to undergo formal testing for fusions.

Minor
- ll 46-47: surgery is a key element of NMIBC management and should be mentioned.
- ll 49-51: The recent OS data of EV/pembro should be considered more than just a limited improvement.
- In the introduction, the authors first refer to FGFR3 mutations (ll54) to introduce the role FGFR3 activation in UC tumorigenesis. Later, the broader term of alterations is introduced. I suggest initially referring to all alterations since fusions and amplification may also play an important role in FGFR-driven tumorigenesis.
- l74: It’s irrelevant to the manuscript that Janssen produces erdafitinib. Therefore, the brand name should be removed.
- ll72-76: The FDA approved erdafitinib after progression on chemotherapy, the EMA post IO. Hence, erdaftitinib can also be given as 2nd line, including in locally advanced pts.
- Gene names are not consistently italicized.

Reviewer #2 (Remarks to the Author): Early-career researcher co-reviewer

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.

Reviewer #3 (Remarks to the Author): Expert in machine learning and deep learning, computer vision, and cancer digital pathology

The authors present an interesting and timely work on prediction of gene expressions and mutation status from H&E stained digital pathology images on bladder cancer. The work is well presented, the experiments seem extensive and the results encouraging. As such, I am supportive of this work towards publication, but need to flag upfront that my evaluation of this paper is purely based on the digital pathology side and not on the bioinformatics part at all, since I come from a medical computer vision background, and not computational genomics. Having said that, I mainly have three technical comments to make.

First, CNN based deep classifiers like ResNet and autoencoder based deep segmentors like U-Net have been the gold standard till a few years back but are no longer so. Computer vision has shifted to vision transformer based foundation models. For segmentation we now have SAM, MedSAM and CellSAM. For deep feature extraction from digital pathology images we have pre-trained foundation models like Microsoft's GigaPath model (open source), Mahmood Lab's UNI model (open source) and Google's PathFoundation model (accessible through API). Why not use these directly for feature extraction instead of ResNet? And even if this can be justified somehow, at least a comparative result is needed with one of those models in order to be state-of-the-art.

The second comment is regarding model evaluation. I feel in order to have further confidence or robustness in the results some uncertainty quantification exercise is needed, for example conformal inference. This would also increase the chances of translation into practice from an AI safety and interpretability point of view.

Finally, how do we account for generalisability in the sense of staining and quality variations between H&E slides from different sources or scanners which has been a bane of AI for digital pathology? Again, this might be a shortcoming of the current method as the ResNet is trained from scratch on this dataset specifically, and hence might not work on an unknown test set from a different source, whereas the foundation models are already pre-trained on millions of pan cancer images from different sources for generalised feature extraction, please justify?

Response to reviewers

Common response

Dear Reviewers,

We sincerely thank you for your comprehensive review and valuable feedback on our manuscript. We are pleased that you recognized the relevance and timeliness of our work in the context of rapidly evolving MIBC and mUC treatment landscapes, as well as the well-structured presentation of our findings.

Your insights have significantly enhanced the quality of our manuscript. We have diligently addressed each of your comments and made the following key improvements:

- **Enhanced Model Performance**: We updated our model's feature extractor with a Vision Transformer, resulting in superior performance in both cross-validation and external validation.

- **Improved Result Presentation**: We have clarified our model's performance by including predictive values and highlighting the potential reduction in molecular tests.

- **Refined Discussion**: We streamlined our analysis in the discussion section to present a more focused and impactful message on FGFR3 mutations.

These revisions have substantially improved the clarity, accuracy, and relevance of our work. We believe these changes align with your expectations and address the concerns raised in your review. Please find below a point-by-point response to all your comments.

We appreciate your time and expertise in evaluating our research.

Thank you for your consideration.

Pierre-Antoine Bannier

On behalf of all co-authors

Response to reviewers

Reviewer 1 & 2

Major comments

Comment #1: "*The authors chose to report their test performance primarily as AUC. However, this is not useful in the presence of class imbalance. This is the case here, given the low number of FGFR3 events in the various cohorts (8-14%). Instead, the authors should report the PPV in the main text and replace the "rule out %" with the NPV. To add more clarity, the authors should also provide the number of pts who were called FGFR3 mutated by the algorithm in each cohort. Lastly, the number of saved molecular analyses should be provided, since it's the aim of the study. With the current results, the algorithm clearly overestimates the presence of FGFR alterations. This, together with the variable performance in the different validation cohorts, may be indicative of overfitting. To further elucidate, the author should provide the performance metrics of the training set for comparison. Ideally, the authors should use an additional independent test set."*

We presented the AUC because it is a widely recognized metric to evaluate machine learning model performance and allows for better comparison with competing approaches in the literature. As suggested by the reviewers, we have now included both the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV) in the main text (see paragraph "*The model saves on average 40% of molecular tests"* in the Results section), and revised the supplementary materials to exclude the "rule out %" column and replace it with the NPV. We added a table (Table 1) in the main text to highlight the number of saved molecular tests. To address concerns about overfitting, we added the performance of the model in cross-validation (AUC = 0.82 [0.74 - 0.90]) in the supplementary materials and mentioned it in the main text. We want to highlight that we did not observe a significant decrease in AUC, Sensitivity, and Specificity when comparing cross-validation results to those from external validation across three independent test sets, which hints that the model generalizes correctly to unseen cohorts.

Comment #2: *"Local staining procedures and reagents can cause batch effects when AI is applied. The authors describe accessing FFPE slides from the TCGA project in addition to their local samples. It's unclear to me if those are physical slides or digital images of them. In the case of physical slides, the author should state whether they were stained using the setup in Erlangen. If all H&E images were generated in Erlangen, the additional independent data test set (requested above) should be stained externally."*

Here is the clarification regarding the format and staining of the slides used in our study:

1. The slides from the TCGA project were utilized in a digital format, accessed directly from the TCGA slide archive. They were stained outside of our institution.
2. The MIBC I (2018-2022) and MIBC II cohorts were stained in two different laboratories using different staining procedures and reagents in Erlangen.

3. The NMIBC I (2023) cohort included slides stained partly in the Erlangen laboratory and partly in the Regensburg laboratory, using different staining procedures from Erlangen laboratory.
4. The mUC (2017) cohort was stained exclusively in the Erlangen laboratory.
5. The NMIBC II cohort was stained in the Erlangen laboratory.

In our cohorts, there is a large variety of H&E staining protocols coming from different laboratories, and also a variety in the digital scanners used (TCGA vs. the other cohorts). We clarified this part in the cohort description in the Methods section (paragraph *Datasets description*).

Comment #3: *"The frequencies of FGFR3 alterations are low compared to previous publications. The authors should use a more comprehensive assay in a representative subset of patients to assess the number of false-negative "ground truth". Furthermore, the authors should provide information on the three FGFR3 regions amplified by their SNaPshot PCR."*

We agree that this point should be further explored. There are few studies in the current literature that have systematically investigated the prevalence of activating *FGFR3* mutations in large patient cohorts. We have cited other sources in the introduction as evidence that the prevalence depending on tumor stage (NMIBC: pTa/pT1 50-80%; nmMIBC 10-15% and mUC ~15%; Gust et al, van Rhijn et al, Tomlinson et al, Ascione et al; newly added to references) overlaps very well with the prevalence described in the literature. Like our data, the data collected in the literature refer exclusively to activating and thus also therapeutically targetable and functionally relevant *FGFR3* mutations.
The SNaPshot method (van Oers et al, 2005, CCR ; Billerey et al, 2001, Am J Pathol) is based on the amplification of three gene regions in exons 7, 10 and 15 of the FGFR3 gene, in which all known activating (hot-spot) mutations of the *FGFR3* gene in urothelial carcinomas are localized. The method allows the detection of the following known hotspot point mutations with functionally proven activation of FGFR3 with very high sensitivity and at the same time very simple methodology: R248C, S249C, G372C (in NGS G370C), G382R (in NGS G380R), S373C (in NGS S371C), Y375C (in NGS Y373C), A393E, K652E/Q (in NGS K650E/Q), K652M/T. It is important to note that no additional known activating *FGFR3* mutations were described in the largest NGS-based urothelial carcinoma dataset to date (TCGA bladder cancer; exome sequencing based); in this respect, the result of this multiplex PCR can be considered comprehensive. In addition, the companion diagnostic assay of erdafitinib (Therascreen assay), for example, only covers significantly fewer activating mutations (R248C, S249C, G370C, Y373C), which, however, already cover >90% of the expected activating mutations in the spectrum of urothelial cancer.

Comment #4: *"The authors don't provide any information on the presence or absence of cis and histological subtypes. If they were not excluded, it should be stated how they affected performance."*

We presented the results on the different subgroups (with CIS / without CIS; histological subtypes; growth patterns) in the Results section (see paragraph *The model performed consistently across histological subtypes*). Furthermore, we added:

1. A “Concomitant CIS” line in the Supplementary Table 1 with the proportion of Concomitant CIS for each cohort,
2. A Table (Supplementary Table 9) with the histological subtype distribution for each cohort,
3. A Table (Supplementary Table 10) with the distribution of growth patterns for each cohort.
4. 3 Tables (Supplementary Table 12, 13 and 14) with the detailed results on the external validation cohorts.

Comment #5: *“Rightfully, the authors point toward the relevance of the TAR-210 system to moving FGFR-targeted therapy to the NMIBC space. However, in this context, validation on an NMIBC cohort is key. The authors should ideally provide this, or at least the clinical rationale for MIBC enrichment whilst excluding UTUC cases. As is, the cohorts are not ideal for either treatment landscape (local vs. systemic).”*

We agree that the selection of the cohort should be further elucidated. While UTUC are included in metastatic treatment settings (e.g., in THOR-Trial; or in our mUC cohort) the majority of patients suffering from metastatic UC are still patients with metastatic urothelial cancer of the urinary bladder. Thus, we believe that we are still covering the overall majority of patients that could be screened for FGFR3 inhibition in metastatic settings by using our large data set of nmMIBC/mMIBC. However, as it is well established that NMIBC and MIBC/mUC arise from different pathways associating with very different rates of activating FGFR3 mutations we also included NMIBC from beginning to delineate differences of potential morphological features in NMIBC versus MIBC/mUC. As outlined in the results and discussion sections including NMIBC even improved performance of our model in MIBC and mUC thus underlining that there are morphological features that are independent of tumor stage and indicate the presence of *FGFR3* mutations.

In addition, we now added a further NMIBC validation cohort consisting of n=109 NMIBC as these tumors are a current focus of FGFR3 inhibitor development. We reported the results of the external validation in the paragraph “*Training on NMIBC cases improves the model performance”*. The lower performance benchmarks in the NMIBC validation cohort can be attributed to the circumstance that our present model has been mainly trained for FGFR3 mutation detection in solid muscle-invasive and metastatic urothelial cancer. Although there might be shared features indicating the presence of FGFR3 mutations between NMIBC and MIBC/mUC, these are likely not identical. Thus, we believe that algorithms for the biologically different subset of NMIBC need to be developed separately solely for this specific purpose.

Comment #6: *“The authors should provide information on how patients/biopsies were selected. Given that FGFR3 mutations are associated with a macroscopic growth pattern, the authors should provide, which biopsies showed a papillary vs. solid growth, and whether the algorithm performed well in both groups. Furthermore, it would be interesting to see how much the algorithm is driven by other macroscopic features of the slide. The authors report several examples of the most predictive tiles. However, more information would be helpful to assess this. The authors should state which measures were taken to avoid over-fitting.”*

All patients from the different cohorts were collected and included consecutively as now outlined in the material and method section. In addition we added a sentence how tissue samples were selected: “Per sample a FFPE block with at least 30% tumor content (related to all tissue covered on the block) and at least an area of vital non-necrotic 0.5x0.5 cm tumor tissue area with surrounding tumor stroma was selected by an experienced board-certified uropathologist (M.E.).” . We added in Supplementary Table 13 a comparison of the model performance on the Solid and Papillary growth patterns. These results are mentioned in the main text in paragraph “*The model performed consistently across histological subtypes*”. We would like to highlight that the model showed consistent performance on histological subgroups, subgroups based on growth patterns and based on concomitant CIS over three different independent external validation cohorts. We think this is strongly indicative of the model being correctly fit on the data.

Regarding the assessment of macroscopic features, we detail in Figure 5 and in the associated paragraph *“Histopathology shows patterns linked to FGFR3 mutations”* the morphological features associated with the FGFR3-mutation and wild-type statuses.

Comment #7: *“In the paragraph ll139-158, the authors refer to the concept of FGFR3-driven tumours. However, this feature is never defined or referenced. In particular, the authors exclude FGFR wt tumours with high FGFR3 gene expression from the group which they refer to as being FGFR3-driven. Please elucidate.”*

We agree that this is an uncommon terminology and not an established concept. We decided to rephrase this paragraph and remove the term “FGFR3-driven” to avoid confusion. We rather described the analyzed groups descriptively (e.g., FGFR3 wild type with high FGFR3 expression).

Comment #8: *“The authors should discuss their work in the context of previous, similar studies (https://doi.org/10.1002/cam4.4044, not cited; https://doi.org/10.1016/j.euf.2021.04.007, cited).”*

We thank the reviewer for bringing up this point. We included these two studies in our discussion section, where we reflected and discussed their content in the view of our present results. Besides, we would like to highlight that we provided a comparison on the exact same cases as Loeffler et al. on TCGA (n=327) and scored 0.83 [0.76-0.88] on external validation, significantly improving on their reported 0.70 in cross-validation. In external validation on the TCGA dataset, our model achieved comparable performance to Velhamos et al., despite their results being from cross-validation rather than external testing. Furthermore, our model was more thoroughly validated on three external validation cohorts.

Comment #9: *“Paragraph ll252-262: The discussion of the failed FGFR3-TACC3 fusions is unbalanced. A Hazard Ratio of 0.49 for survival with an upper-limit of the confidence interval that barely crosses unity in a subset analysis for a small cohort cannot be described as “patients … did not formally benefit”. In contrast the HR for survival in the FGRF mutated cohort is only 0.67. For what it’s worth, objective response rates are also higher for the*

translocation cohort (HR 4.18 vs. 3.89). As a clinician I can confirm that responses for FGFR3-TACC3 fusion patients are frequent and clinically meaningful. Furthermore, the authors should provide further references that fusions don't alter the morphological and/or clinical phenotype to support their hypothesis of this being at fault for poor test performance. I would also argue that results (ROC/NPV/PPV etc) should include the inability to identify fusions, as (although rarer than point mutations) these are clinically very relevant and need to be identified."

We agree with your opinion on that point and apologize for the former unbalanced statement. We updated and rephrased the text passage especially highlighting that further investigations with fusion enriched cohorts need to be studied to elucidate whether these alterations could be identified with computational pathology methods, because these patients can benefit from FGFR3 inhibition tremendously. However, we cannot provide references stating that a different morphology is the reason for the poor performance; thus, we removed this speculative statement. We concluded that our inability to screen FGFR3 fusions could be caused by the very low number of fusion positive samples and needs to be further studied in upcoming studies.

Comment #10: *"As a comment: It is this final point that may represent a fatal flaw. If the authors are able to develop a screening platform that is sensitive enough to rule out FGFR alterations, this would represent a time and cost savings as those patients would not have to undergo further testing. Given the different prevalence (and therefore pre-test probability) of FGFR alterations in localized vs. metastatic disease I would argue the test would need separate validation in those cohorts, but the critical issue is that if FGFR fusions are excluded, this is not a trivial exclusion and then even patients who's slides test negative by this methodology would have to undergo formal testing for fusions."*

We agree with your opinion and have therefore reworded and weakened the statement in the last paragraph of the conclusion. We have also emphasized that further blinded validation and calibration experiments in larger cohorts - ideally in patients treated with erdafitinib or similar drugs - are needed. We also fully agree that, given sufficient power of a predictive tool for mutations, all patients must continue to be screened for fusions. However, a pre-screening tool for mutations with meaningful performance measures could significantly reduce costs, even if all patients always required a fusion test. This is because mutation and fusion testing usually require different analytes (DNA versus RNA) and different sequencing modalities when performed by NGS - which is currently one of the most commonly used methods for assessing FGFR3 alterations in predictive molecular pathology. Widespread use of the RNA-based Therascreen test (so far still CDA for erdafitinib), which allows parallel testing of fusions and mutations, is also not expected, as the Quiagen PCR equipment required for this test is not widely accessible and only very few mutations are covered by this test. Therefore, this test does not appear to be suitable for the investigation of FGFR(3) alterations, especially if other FGFR inhibitors are approved for urothelial carcinomas whose approval regulations are based on a broader mutation spectrum.

Minor comments

Comments #1, 3, 5: *"ll 46-47: surgery is a key element of NMIBC management and should be mentioned", "In the introduction, the authors first refer to FGFR3 mutations (ll54) to introduce the role FGFR3 activation in UC tumorigenesis. Later, the broader term of alterations is introduced. I suggest initially referring to all alterations since fusions and amplification may also play an important role in FGFR-driven tumorigenesis.", "ll72-76: The FDA approved erdafitinib after progression on chemotherapy, the EMA post IO. Hence, erdaftitinib can also be given as 2nd line, including in locally advanced pts."*

We have updated the introduction accordingly.

Comment #2: *"ll 49-51: The recent OS data of EV/pembro should be considered more than just a limited improvement."*

We have updated the introduction respectively and inserted a citation of the EV302 study.

Comment #4: *"l74: It's irrelevant to the manuscript that Janssen produces erdafitinib. Therefore, the brand name should be removed."*

We removed the mention of Janssen in the manuscript.

Comment #6: *"Gene names are not consistently italicized."*

We consistently italicized the gene names.

Reviewer 3

Comment #1: *"First, CNN based deep classifiers like ResNet and autoencoder based deep segmentors like U-Net have been the gold standard till a few years back but are no longer so. Computer vision has shifted to vision transformer based foundation models. For segmentation we now have SAM, MedSAM and CellSAM. For deep feature extraction from digital pathology images we have pre-trained foundation models like Microsoft's GigaPath model (open source), Mahmood Lab's UNI model (open source) and Google's PathFoundation model (accessible through API). Why not use these directly for feature extraction instead of ResNet? And even if this can be justified somehow, at least a comparative result is needed with one of those models in order to be state-of-the-art."*

We acknowledge the shift towards using vision transformer-based feature extractors in computational pathology, and more broadly in computer vision. Our decision to use a Wide ResNet50 trained via a self-supervised momentum contrastive method was guided by its superior performance in preliminary comparisons. Initially, we evaluated two ViT16-Small and ViT16-Base models, trained using the iBOT method on a diverse pan-cancer training set from 16 TCGA cohorts. However, the Wide ResNet50 consistently outperformed these ViT feature

extractors. In the meantime, we developed a new feature extractor named Phikon (Filiot et al.) trained on 40 million patches from 6093 slides and Bioptimus launched H0 (Saillard et al.). Both of these ViT-based feature extractors have been benchmarked on the prediction of FGFR3-mutation in MIBC H&E slides. The results are reported in the table below. Additionally, in response to the reviewer's suggestion, we have benchmarked the Mahmood Lab's UNI feature extractor and the SwinTransformer CTransPath extractor. We report the AUROC for the FGFR3MUT models using different feature extractors in Table 1.

After careful consideration, in the FGFR3MUT model, we decided to replace our Wide ResNet50 feature extractor with Bioptimus' H0 extractor. This choice was motivated by 1) the solid performance of the feature extractor in cross-validation and 2) the absence of data leakage (which might artificially inflate the results) with TCGA-BLCA. All results in the main text, the figures and the supplementary have been updated accordingly.

	Cohorts			
	Cross-validation (n=391)	mUC (n=96)	MIBC II (n=183)	TCGA MIBC (n=307)
Wide ResNet50 MoCo COAD (previous)	0.80 [0.78 - 0.83]	0.81 [0.70 - 0.90]	0.82 [0.73 - 0.90]	0.82 [0.74 - 0.88]
Bioptimus' H0 (new)	**0.82 [0.80 - 0.84]**	**0.85 [0.74 - 0.95]**	**0.89 [0.82 - 0.94]**	0.80 [0.70 - 0.86]
Phikon*	0.80 [0.78 - 0.83]	0.80 [0.65 - 0.91]	0.88 [0.82 - 0.93]	0.74 [0.66 - 0.83]
UNI	**0.82 [0.80 - 0.84]**	0.82 [0.70 - 0.91]	0.84 [0.74 - 0.92]	**0.83 [0.74 - 0.90]**
CTransPath*	**0.82 [0.80 - 0.84]**	**0.85 [0.70 - 0.95]**	0.81 [0.68 - 0.91]	**0.86 [0.79 - 0.91]**

***** *Data leakage with TCGA-BLCA*

Table 1: AUROC of FGFR3MUT models using different feature extractors. All feature extractors were evaluated under identical conditions, employing the multiple-instance learning model detailed in the main text. The same cases were utilized across all tests, to classify FGFR3-mutant versus wild-type cases.

Comment #2: *"The second comment is regarding model evaluation. I feel in order to have further confidence or robustness in the results some uncertainty quantification exercise is needed, for example conformal inference. This would also increase the chances of translation into practice from an AI safety and interpretability point of view."*

We thank the reviewer for this suggestion. We added an entire paragraph in the results section called *"The model demonstrates high confidence in its predictions"*, and added 3 figures (Figure 4c-e) to quantify the uncertainty of the model predictions. We articulated our demonstration around 2 points as the reviewer suggested.

First, we used a Deep Ensemble uncertainty quantification technique (described in the *Uncertainty quantification and conformal prediction* paragraph of the Methods section) to assess the variance of the predicted probabilities of the ensemble models. We show that the models in the ensemble largely agree on the predicted probabilities, especially as they are closer to 0 and 1.

Second, we conformalized the predictions in the three external validation cohorts (described in the *Uncertainty quantification and conformal prediction* paragraph of the Methods section) and computed prediction sets for the three external validation cohorts. We reported the expected cardinality of these prediction sets in the main text.

Comment #3: ***"****Finally, how do we account for generalisability in the sense of staining and quality variations between H&E slides from different sources or scanners which has been a bane of AI for digital pathology? Again, this might be a shortcoming of the current method as the ResNet is trained from scratch on this dataset specifically, and hence might not work on an unknown test set from a different source, whereas the foundation models are already pre-trained on millions of pan cancer images from different sources for generalised feature extraction, please justify?"*

We acknowledge the reviewer's concern regarding the critical challenge of staining and scanner variability in computational pathology. We would like to emphasize that our initial choice of feature extractor, a WideResNet50, was specifically trained in-domain using a self-supervised momentum contrastive approach (MoCo v2) on TCGA (excluding TCGA-BLCA). This dataset included more than 4 million patches, encompassing a broad spectrum of staining variations and scanners. In the development of the FGFR3MUT model, the WideResNet50 (68M parameters) trained in-domain was utilized for feature extraction, with its parameters frozen. We clarified this part in the manuscript.

Furthermore, we would like to highlight that we switched to a 1.8 billion parameter foundational model (Bioptimus' H0; Saillard et al.) trained on 500k pan-tumor whole-slide images and hundreds of millions of H&E tiles, enhancing the generalization capability of our FGFR3MUT model.

Finally, we believe that our extensive validation process, which included multiple cancer stages (NMIBC, MIBC, mUC) and varied staining protocols (coming from different laboratories at Erlangen hospital and on TCGA) and scanners (e.g., Aperio, 3D Histech), demonstrates the model's generalizability and robustness to batch effects, confirming that it is correctly fitted to the data.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

We thank the authors for their commitment to improving the submitted work. There are only a few minor issues remaining.

Discussion:
279: The sentence could be read that all TKIs target FGFR, please re-write.
The beginning of the discussion is a little too redundant compared to the introduction. It could be shortened to make it a more appealing read.

Fig 1:
- Not all details depicted are explained in the legend (eg. what the percentages refer to in the upper part of b)

Fig 2:
- The AUC provides enough space to add the PPVs and NPVs
- to increase readability, the authors could consider rearranging the tiles in d-f to have a case be represented in one row (or column)

Fig 3:
- To me, it is not clear to what the authors refer to with “figure in bold”
- The authors should use the canonical isoform nomenclature when referring to mutations (currently, a uses Y373C, b+c use Y375C). Please harmonize colours accordingly.

FIg 4:
- a - the metastasis symbol in the legend and the one in the figure differ

Fig 5:
- (T) and (NT) are not explained
- A consistent color scheme for FGFR wt and mut across all figures would be desirable
- b - the patient’s FGFR status could be provided

Table 1:
- First data column: n as a fraction of total should be provided
- Adding a column containing the % of “ground truth FGFRalt” per cohort would be interesting
- Please change the wording of the column header to “Potentially saved molecular tests”

Reviewer #2 (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.

Reviewer #3 (Remarks to the Author):

Authors have addressed my comments in details and I am mostly satisfied with the revised manuscript... a couple of final minor suggestions are that since the authors have included the suggested results on conformalised uncertainty quantification, maybe consider citing this recent paper on the topic in nature medicine: https://www.nature.com/articles/s41591-023-02562-7 and since they have used a self-supervised momentum contrastive method (MoCo) for training, consider citing this improved version of MoCo: https://link.springer.com/chapter/10.1007/978-3-030-87589-3_1, speaking of which how do we know that the "flavour" of Moco used for this work is the right one? There seems to be many variations of momentum contrastive learning out there in the existing literature.

Response to reviewers

Reviewer 1 & 2

Comment #1: *"Discussion: 279: The sentence could be read that all TKIs target FGFR, please re-write. The beginning of the discussion is a little too redundant compared to the introduction. It could be shortened to make it a more appealing read."*

We rephrased the first sentence to lift any ambiguities. Furthermore, we shortened the first paragraph of the discussion and removed redundant passages with the introduction.

Comment #2: *"Fig 1: Not all details depicted are explained in the legend (eg. what the percentages refer to in the upper part of b)"*

We clarified all the details. Notably, we clarified that the percentages in (b) referred to the proportion of FGFR3-mutant cases in each cohort.

Comment #3: "*Fig 2: The AUC provides enough space to add the PPVs and NPVs, to increase readability, the authors could consider rearranging the tiles in d-f to have a case be represented in one row (or column)"*

We added the NPVs and PPVs. Furthermore, we reorganized the panels d-f to have a case represented in one row. Now, a row shows a thumbnail of the slide (d), the corresponding heatmap (e) and the top predictive tiles (f).

Comment #4: "*Fig 3: To me, it is not clear to what the authors refer to with "figure in bold", the authors should use the canonical isoform nomenclature when referring to mutations (currently, a uses Y373C, b+c use Y375C). Please harmonize colors accordingly."*

We clarified the legend, and used the canonical isoform nomenclature.

Comment #5: "*Fig 4: a - the metastasis symbol in the legend and the one in the figure differ".*

We fixed it.

Comment #6: "*Fig 5:*
- (T) and (NT) are not explained
- A consistent color scheme for FGFR wt and mut across all figures would be desirable
- b - the patient's FGFR status could be provided".

We clarified in the legend what (T) and (NT) refer to. Besides, we changed the color scheme to make it consistent with all the other figures. We also provided the patient's FGFR status.

Comment #7: *"Table 1:*
- First data column: n as a fraction of total should be provided
- Adding a column containing the % of "ground truth FGFRalt" per cohort would be interesting
- Please change the wording of the column header to "Potentially saved molecular tests""

- We added the fraction of total in the first column.
- This has already been added in the main figure 1, we would rather keep Table 1 as simple as possible.
- We changed the wording.

Reviewer 3

Comment #1: *"A couple of final minor suggestions are that since the authors have included the suggested results on conformalised uncertainty quantification, maybe consider citing this recent paper on the topic in nature medicine: https://www.nature.com/articles/s41591-023-02562-7"*

We added the citation of this paper in the main text.

Comment #2: *"since they have used a self-supervised momentum contrastive method (MoCo) for training, consider citing this improved version of MoCo: https://link.springer.com/chapter/10.1007/978-3-030-87589-3_1, speaking of which how do we know that the "flavour" of Moco used for this work is the right one? There seems to be many variations of momentum contrastive learning out there in the existing literature."*

The revised version of the model no longer uses a feature extractor trained by momentum contrastive learning, but rather using a self-supervised learning method called DINOv2 (see the *"Preprocessing of whole-slide images"* section). We duly explained and cited the relevant work for this method.

To answer the reviewer's question regarding momentum contrastive learning, we used MoCo v2 (https://arxiv.org/pdf/2003.04297) to train the feature extractor presented in the first submission of the paper. This publication (https://arxiv.org/abs/2012.03583) explains the training of the feature extractor initially presented.