BRIEF COMMUNICATION   **OPEN**

Check for updates

Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology

Oliver Lester Saldanha[1,2,10], Chiara M. L. Loeffler[1,2,10], Jan Moritz Niehues[1,2], Marko van Treeck[1,2], Tobias P. Seraphin [iD][3], Katherine Jane Hewitt[1,2], Didem Cifci[1,2], Gregory Patrick Veldhuizen [iD][1,2], Siddhi Ramesh[4], Alexander T. Pearson [iD][5,6] and Jakob Nikolas Kather [iD][1,2,7,8,9] ✉

The histopathological phenotype of tumors reflects the underlying genetic makeup. Deep learning can predict genetic alterations from pathology slides, but it is unclear how well these predictions generalize to external datasets. We performed a systematic study on Deep-Learning-based prediction of genetic alterations from histology, using two large datasets of multiple tumor types. We show that an analysis pipeline that integrates self-supervised feature extraction and attention-based multiple instance learning achieves a robust predictability and generalizability.

The genotype of any solid tumor determines its phenotype, giving rise to a large variety of patterns in cancer histopathology. Deep learning (DL), a tool from the realm of artificial intelligence, can infer genetic alterations directly from routine histopathology slides stained with hematoxylin and eosin (H&E)[1,2]. Initial studies demonstrated this predictability in lung cancer[3], breast cancer[4], and colorectal cancer[5]. Subsequently, several "pan-cancer" studies showed that DL-based prediction of biomarkers is feasible across the whole spectrum of human cancer[6–10]. However, these studies were overwhelmingly performed in a single large cohort without externally validating the results on a large scale. This raises a number of potential concerns, as prediction performance can be heavily biased by batch effects occurring in such single multicentric datasets[11,12]. To move closer to clinical applicability, external validation of any DL system is paramount[13]. Recent technical benchmark studies have demonstrated that attention-based multiple instance learning (attMIL)[14] and self-supervised learning (SSL)[15,16] for pre-training of feature extractors[17,18] can improve performance and generalizability for computational pathology biomarkers, but these technical advances have not yet been systematically applied to mutation prediction in a pan-cancer approach.

We acquired two large, multi-centric datasets of cancer histopathology images with matched genetic profiling: the Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) dataset. We used all tumor types which were present in both datasets, namely: breast (BRCA; TCGA $N = 1066$, CPTAC $N = 122$), colorectal (CRC; TCGA $N = 534$, CPTAC $N = 105$), glioblastoma (GBM; TCGA $N = 397$, CPTAC $N = 96$), lung adeno (LUAD; TCGA $N = 566$, CPTAC $N = 110$, lung squamous (LUSC; TCGA $N = 484$, CPTAC $N = 108$), pancreatic (PAAD; TCGA $N = 179$, CPTAC $N = 140$ patients), and (uterine) endometrial cancer (UCEC; TCGA $N = 517$, CPTAC $N = 95$; Fig. 1A, B and Supplementary Fig. 1A, B). We aimed to use Deep Learning to predict all $N = 1068$

clinically relevant oncogenes and tumor suppressor genes (Fig. 1C and Supplementary Fig. 1C) in the OnkoKb[19] database. The number of genes analyzed decreased after excluding alterations of unknown significance and the definition of a minimum number of mutated cases. We trained the model to predict mutations in TCGA (Fig. 1A) and evaluated the performance on CPTAC (Fig. 1B). The primary endpoint was the mean (±standard deviation) area under the receiver operating characteristic curve (AUROC) of five replicate experiments. We benchmarked our methods against four other methods, as laid down in the "Methods"" section, and found that the combination of SSL + attMIL outperforms other tested approaches.

We found that in most tumor types, mutations in several genes were predictable from histology (Fig. 1D, E and Supplementary Data 1). In accordance with previous studies[20], endometrial cancer (UCEC) had the highest number of detectable mutations. $N = 13$ out of $n = 43$ analyzable genes had an AUROC of 0.60 or higher, of which 6 reached an AUROC of 0.70 or higher in the external validation cohort (Fig. 2A). Among these were *PTEN* mutations (AUROC 0.73 ± 0.03) involved in hereditary cancer;[21] *TP53* mutations (AUROC 0.72 ± 0.05), which is associated with poor prognosis[22] and *APC* (AUROC 0.72 ± 0.11), as a potential predictive marker for immunotherapy in endometrial cancer[23]. We identified 7 genes for which mutations were predictable (out of 16 analyzable genes) for colorectal cancer (CRC) with AUROCs of over 0.6 in the external validation cohort. (Fig. 2B). This included prognostic alterations, such as *BRAF* and *KRAS* mutations, which reached an AUROC of 0.66 ± 0.24 and 0.66 ± 0.03 respectively (Supplementary Data 1). In GBM, 4 out of 9 genes had an AUROC over 0.70 including the genes *IDH1* (AUROC 0.84 ± 0.06), *ATRX* (AUROC 0.70 ± 0.10), *TP53* (AUROC 0.70 ± 0.07) and *RB1* (AUROC 0.70 ± 0.07; Fig. 2C). These are of increasing relevance as classification of brain tumors are increasingly based on molecular markers and therefore of therapeutic importance. In the other

---

[1]Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany. [2]Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany. [3]Department of Gastroenterology, Hepatology and Infectious Diseases, University Hospital Duesseldorf, Medical Faculty at Heinrich-Heine-University Duesseldorf, Düsseldorf, Germany. [4]Pritzker School of Medicine, University of Chicago, Chicago, IL, USA. [5]Biological Sciences Division, University of Chicago, Chicago, IL, USA. [6]University of Chicago Comprehensive Cancer Center, University of Chicago, Chicago, IL, USA. [7]Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK. [8]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany. [9]Department of Medicine 1, University Hospital and Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany. [10]These authors contributed equally: Oliver Lester Saldanha, Chiara M. L. Loeffler. ✉email: jakob_nikolas.kather@tu-dresden.de
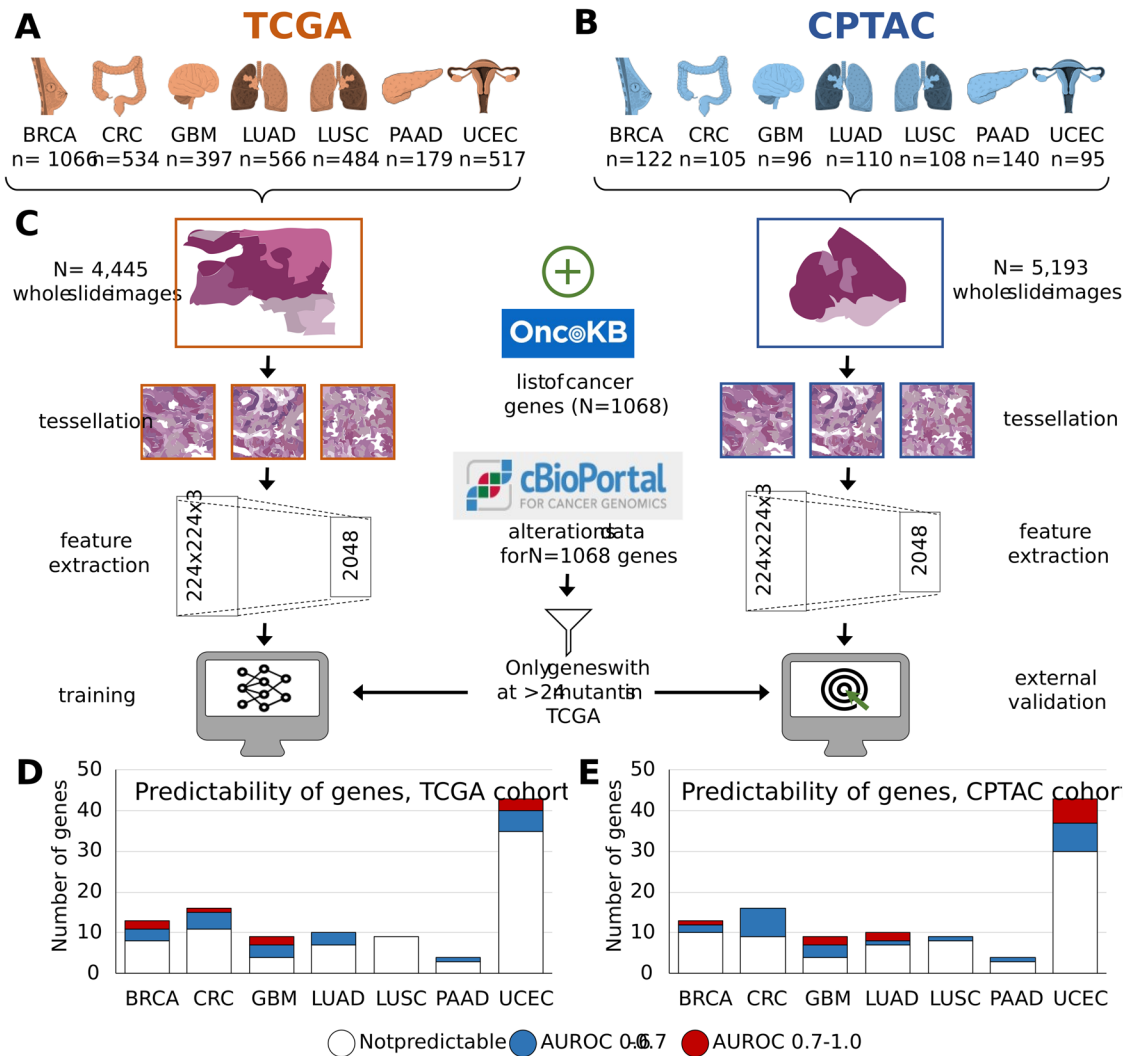
Fig. 1 Study Design with selection process of the genes. A, B Patient numbers for each tumor type in The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) dataset. **C** Flowchart showing the preprocessing steps for the training and validation cohort. Furthermore, an outline of the gene selection process. **D** Overview Area under the receiving operating curve (AUROC) results for internal cross-validation in TCGA. **E** Overview AUROC results for external validation on CPTAC. The plots are based on the original AUROC values with 5 decimal digits, while numbers in the manuscript text have been rounded to two decimal digits. (Icons were used from Servier Medical Art provided by Servier, licensed under a Creative Commons Attribution 3.0 unported license).

tumor types prognosis relevant mutations such as *EGFR* in LUAD (AUROC 0.76 ± 0.03) or *TP53* in BRCA (AUROC 0.72 ± 0.05) could be predicted with an AUROC over 0.70 (Fig. 2D, E). For the tumor type LUSC highest AUROCS of 0.61 ± 0.14 for *RB1* was achieved (Fig. 2F). Moreover the DL algorithm was able to detect CDH1 alterations in BRCA with an AUROC of 0.68 ± 0.17 (Fig. 2G–I). Compared to the other tumor types in our study, the tumors with the highest number of predictable alterations (UCEC, CRC, and BRCA) have a higher average tumor mutational burden[24]. We hypothesize that many morphological alterations are related to immune-mediated changes in the tumor microenvironment. Our method yielded interpretable spatial predictions (Fig. 2G), and unlike previous studies[25] provided separate heatmaps for attention (Fig. 2H) and classification (Fig. 2I and Supplementary Fig. 2). Previous studies have shown such heatmaps to be correlated to the underlying molecular ground truth on a spatial scale[26,27]. In summary, in our study, the use of the new combination of SSL + attMIL showed the best performance in comparison to the other techniques (Supplementary Fig. 3 and Supplementary Data 2), while a visual examination of correctly and wrongly classified cases suggested a

plausible distribution of model attention on the whole slide image (Supplementary Figs. 4A–O, 5A–O, 6A–O, and 7A–O).

A key limitation is that many clinically relevant genes were not analyzable due to having fewer than 25 mutants in TCGA or many alterations with unknown significance. Large-scale efforts are needed to create datasets with a sufficient size, which could be facilitated by federated[28] or swarm[29] learning. Since the early 2000s, studies have shown a link between genetic alterations and histological phenotypes[30], which DL can exploit[1,6,7]. While there is no biological reason why every frequent genetic alteration is actually manifest in histology, our results add to the growing amount of evidence which shows that many of these alterations are indeed determinable from H&E. This is also detached from the morphological subtype (Supplementary Fig. 3B) or other molecular alterations such as MSI or *POLE* mutational status (Supplementary Fig. 3C–E). Crucially, in contrast to previous studies, our pan-cancer mutation prediction models have been externally validated, thereby minimizing the risk of overfitting[11].

Our analysis shows that in almost all tumor types, cross-validated performance of SSL + attMIL is correlated to external
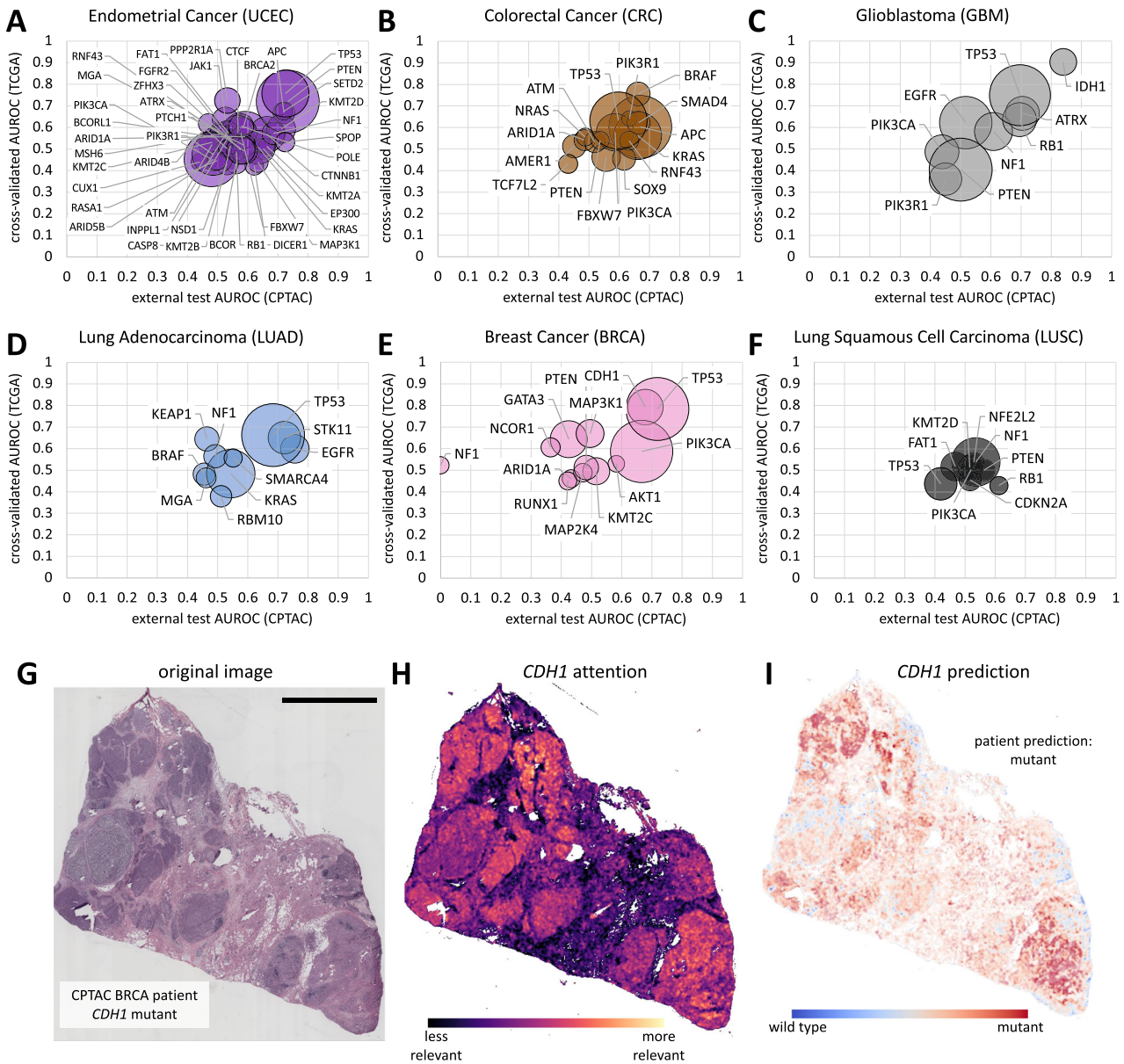
Fig. 2 Classification performance for all genes in internal and external validation. A–F Internal cross-validation and external validation Area under the receiving operating curve (AUROC) for six tumor types (PAAD in Supplementary Fig. 8). The bubble size scales with the number of mutant patients in the external validation cohort. All raw data are in Supplementary Data 1. **G–I** A representative patient from the CPTAC-BRCA cohort, with attention map and prediction maps for *CDH1* mutational status. The scale bar represents 5 mm.

validation performance (Fig. 2A–F and Supplementary Fig. 8) and outperforms the current state of the art. To our knowledge, this is the first time that a multiplexed biomarker prediction from H&E slides has been shown to generalize well. Our study identifies a number of clinically relevant candidate genes amenable to DL-based pre-screening as part of clinical routine practice, with the aim of identifying patients who are good candidates for confirmatory genetic testing.

METHODS

Ethics statement
All experiments were conducted in accordance with the Declaration of Helsinki. For this study, we used anonymized H&E-stained slides from public repositories.

Data acquisition and experimental design
Mutation data for was obtained from https://www.cbioportal.org/[31], excluding alterations of unknown significance and excluding all genes with fewer than $N = 25$ mutant cases in TCGA (Fig. 1C and Supplementary Fig. 1C). This resulted in 43 analyzable genes in endometrial cancer (UCEC), down to 4 analyzable genes in pancreatic cancer (PAAD, Fig. 2A–F and Supplementary Fig. 8). We then used our in-house open-source DL pipeline (https://github.com/KatherLab/marugoto) which uses the SSL-trained model RetCCL[32] to obtain 2048 features per tile and uses attMIL to make patient-level predictions[33,34].

Model architecture
The methods used in this paper follow a two-step approach: the first step is the feature extraction (transforming image tiles into

feature vectors) and the second step is slide aggregation (transforming a set of feature vectors obtained from a given pathology slide into a single prediction for that slide).

For feature extraction, we explore two different methods using a Resnet neural network which was pretrained in a different way. The first feature extraction model is based on a Resnet18 which was pre-trained on ImageNet[35]. We chose this model due to its broad use in the computational pathology research literature[14]. The second model is the Retrieval with Clustering-guided Contrastive Learning (RetCCL)[32] model, a Resnet50 backbone that was trained on a pathology dataset with Self Supervised Learning (SSL).

Also, for the aggregation, we explore two different methods: average pooling (avgPool), which was the dominant approach in the 2018–2020 research literature on clinical datasets[3,5,14,36]. This approach uses a simple multilayer perceptron similar to obtain a prediction for each tile, and then averages the predictions across all tiles for a given slide. The architecture of the multilayer perceptron (classifier network) is (512×256), (256×2). In contrast, the attention-based MIL model (attMIL)[33] has the following architecture: (512×256), (256×2) with an additional attention mechanism[37]. Finally, we use hyperbolic tangent (tanh) as an activation layer to obtain a prediction score.

The core method of our in-house image analysis pipeline "marugoto" is to combine an SSL feature extractor with an attMIL aggregation model. To benchmark both of the models against a baseline, we combine feature extraction and aggregation models in four different ways: ImageNet+attMIL, ImageNet+avgPool, SSL + attMIL, and SSL + avgPool. Lastly, we compared these new techniques to our previous in-house pipeline DeepMed[38] which implements the approach proposed by Coudray et al.[3]. In this approach, the deep layers of a Resnet18 are fine-tuned on image tiles, as described before[5]. We apply all five approaches to the mutation prediction task in colorectal cancer (train on TCGA-CRC, test on CPTAC-CRC) and compare the results (Supplementary Fig. 3A). Based on these, we choose SSL + attMIL for all other experiments.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

TCGA images are from https://portal.gdc.cancer.gov/, CPTAC images are from https://wiki.cancerimagingarchive.net/display/Public/CPTAC + Pathology+Slide+Downloads. Genetic data are available at https://www.cbioportal.org/.

CODE AVAILABILITY

Our pipeline is available under an open source license (https://github.com/KatherLab/marugoto).

REFERENCES

1. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
2. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat. Cancer* **3**, 1026–1038 (2022).
3. Coudray, N. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
4. Couture, H. D. et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *NPJ Breast Cancer* **4**, 30 (2018).
5. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
6. Fu, Y. et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
7. Kather, J. N. et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**, 789–799 (2020).
8. Arslan, S. et al. Deep learning can predict multi-omic biomarkers from routine pathology images: A systematic large-scale study. *bioRxiv* https://doi.org/10.1101/2022.01.21.477189 (2022).
9. Chen, R. J. et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878.e6 (2022).
10. Ghaffari Laleh, N., Ligero, M., Perez-Lopez, R. & Kather, J. N. Facts and hopes on the use of artificial intelligence for predictive immunotherapy biomarkers in cancer.*Clin. Cancer Res.* **29**, 316–323 (2023).
11. Howard, F. M. et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **12**, 4423 (2021).
12. Howard, F. M., Kather, J. N. & Pearson, A. T. Multimodal deep learning: An improvement in prognostication or a reflection of batch effect?. *Cancer Cell* **41**, 5–6 (2023).
13. Kleppe, A. et al. Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer* **21**, 199–211 (2021).
14. Laleh, N. G. et al. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med. Image Anal.* **79**, 102474 (2022).
15. Lai, Z. et al. A semi-supervised learning for segmentation of gigapixel histopathology images from brain tissues. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2021**, 1920–1923 (2021).
16. Amis, G. P. & Carpenter, G. A. Self-supervised ARTMAP. *Neural Netw.* **23**, 265–282 (2010).
17. Schirris, Y., Gavves, E., Nederlof, I., Horlings, H. M. & Teuwen, J. DeepSMILE: contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer. *Med. Image Anal.* **79**, 102464 (2022).
18. Ciga, O., Xu, T. & Martel, A. L. Self supervised contrastive learning for digital histopathology. *arXiv* https://doi.org/10.48550/arXiv.2011.13971 (2020).
19. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* **2017**, PO.17.00011 (2017).
20. Loeffler, C. M. L. et al. Predicting mutational status of driver and suppressor genes directly from histopathology with deep learning: a systematic study across 23 solid tumor types. *Front. Genet.* **12**, 806386 (2022).
21. Spurdle, A. B., Bowman, M. A., Shamsani, J. & Kirk, J. Endometrial cancer gene panels: clinical diagnostic vs research germline DNA testing. *Mod. Pathol.* **30**, 1048–1068 (2017).
22. Tresa, A. et al. Clinical profile and survival outcome of endometrial cancer with p53 mutation. *Indian J. Surg. Oncol.* **13**, 580–586 (2022).
23. Li, Y., Song, Y., Huang, J. & Wang, K. APC mutations as a predictive marker of endometrial cancer immunotherapy: a retrospective cohort study. *Lancet Oncol.* **23**, S9 (2022).
24. Chalmers, Z. R. et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9**, 34 (2017).
25. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
26. Loeffler, C. M. L. et al. Artificial intelligence–based detection of FGFR3 mutational status directly from routine histology in bladder cancer: a possible preselection for molecular testing? *Eur. Urol. Focus* https://doi.org/10.1016/j.euf.2021.04.007 (2021).
27. Schmauch, B. et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat. Commun.* **11**, 3877 (2020).
28. Lu, M. Y. et al. Federated learning for computational pathology on gigapixel whole slide images. *Med. Image Anal.* **76**, 102298 (2022).
29. Saldanha, O. L. et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat. Med.* https://doi.org/10.1038/s41591-022-01768-5 (2022).
30. Rosner, A. et al. Pathway pathology: histological differences between ErbB/Ras and Wnt pathway transgenic mammary tumors. *Am. J. Pathol.* **161**, 1087–1097 (2002).
31. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, I1 (2013).
32. Wang, X. et al. RetCCL: clustering-guided contrastive learning for whole-slide image retrieval. *Med. Image Anal.* **83**, 102645 (2023).
33. Ilse, M., Tomczak, J. M. & Welling, M. Attention-based deep multiple instance learning. *arXiv* https://doi.org/10.48550/arXiv.1802.04712 (2018).

34. Seraphin, T. P. et al. Prediction of heart transplant rejection from routine pathology slides with self-supervised deep learning. *medRxiv* https://doi.org/10.1101/2022.09.29.22279995 (2022).

35. Deng, J. et al. ImageNet: a large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition.* 248–255 (2009).

36. Echle, A. et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology* **159**, 1406–1416.e11 (2020).

37. Leiby, J. S., Hao, J., Kang, G. H., Park, J. W. & Kim, D. Attention-based multiple instance learning with self-supervision to predict microsatellite instability in colorectal cancer from histology whole-slide images. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2022**, 3068–3071 (2022).

38. van Treeck, M. et al. DeepMed: a unified, modular pipeline for end-to-end deep learning in computational pathology. *bioRxiv* https://doi.org/10.1101/2021.12.19.473344 (2021).

AUTHOR CONTRIBUTIONS

O.L.S., C.M.L.L., S.R., A.T.P., and J.N.K. conceived the experiments. M.V.T. developed the source codes for the analysis. J.M.N. and O.L.S. tested and corrected the codes. C.M.L.L. and T.P.S. curated the source data. K.J.H. quality-checked the source data. D.C. and G.P.V. quality-checked the genetic data. O.L.S. performed the experiments. O.L.S. interpreted the data and wrote the manuscript. All authors revised the manuscript draft, contributed to the interpretation of the data and agreed to the submission of this paper.

COMPETING INTERESTS

J.N.K. provides consulting services for Owkin, France; Panakeia, UK and DoMore Diagnostics, Norway. No other potential conflicts of interest are reported by any of the authors.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-023-00365-0.

Correspondence and requests for materials should be addressed to Jakob Nikolas Kather.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.