

# 1 **Direct prediction of Homologous Recombination Deficiency** 2 **from routine histology in ten different tumor types with** 3 **attention-based Multiple Instance Learning: a development** 4 **and validation study**

5 Chiara Maria Lavinia Loeffler (1,2,3), Omar S.M. El Nahhas (2), Hannah Sophie Muti (2, 4),  
6 Tobias Seibel (1), Didem Cifci (1), Marko van Treeck (1,2), Marco Gustav (2), Zunamys I. Carrero (2),  
7 Nadine T. Gaisa (7,8), Kjong-Van Lehmann (7,8), Alexandra Leary (9), Pier Selenica (10), Jorge S.  
8 Reis-Filho (10), Nadina Ortiz Bruechle (7,8\*), Jakob Nikolas Kather (2,3,5,6\*)

9  
10  
11 \*shared last authorship

12 + correspondence to [jakob-nikolas.kather@alumni.dkfz.de](mailto:jakob-nikolas.kather@alumni.dkfz.de)

13  
14 (1) Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany

15 (2) Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical  
16 University Dresden, Dresden, Germany

17 (3) Department of Medicine I, University Hospital and Faculty of Medicine Carl Gustav Carus,  
18 Technische Universität Dresden, Dresden, Germany

19 (4) Department for Visceral, Thoracic and Vascular Surgery, University Hospital Carl Gustav Carus,  
20 Technical University Dresden, Dresden, Germany

21 (5) Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of  
22 Leeds, Leeds, United Kingdom

23 (6) Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg,  
24 Heidelberg, Germany

25 (7) Institute of Pathology, University Hospital RWTH Aachen, Aachen, Germany

26 (8) Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf (CIO ABCD), Germany

27 (9) Gynecological Cancer Unit, Department of Medicine, Institut Gustave Roussy, Villejuif, France

28 (10) Experimental Pathology, Department of Pathology, Memorial Sloan Kettering Cancer Center,  
29 New York, NY, USA

## 30 Abstract

31 **Background:** Homologous Recombination Deficiency (HRD) is a pan-cancer predictive biomarker that  
32 identifies patients who benefit from therapy with PARP inhibitors (PARPi). However, testing for HRD is  
33 highly complex. Here, we investigated whether Deep Learning can predict HRD status solely based on  
34 routine Hematoxylin & Eosin (H&E) histology images in ten cancer types.

35 **Methods:** We developed a fully automated deep learning pipeline with attention-weighted multiple  
36 instance learning (attMIL) to predict HRD status from histology images. A combined genomic scar HRD  
37 score, which integrated loss of heterozygosity (LOH), telomeric allelic imbalance (TAI) and large-scale  
38 state transitions (LST) was calculated from whole genome sequencing data for n=4,565 patients from  
39 two independent cohorts. The primary statistical endpoint was the Area Under the Receiver Operating  
40 Characteristic curve (AUROC) for the prediction of genomic scar HRD with a clinically used cutoff  
41 value.

42 **Results:** We found that HRD status is predictable in tumors of the endometrium, pancreas and lung,  
43 reaching cross-validated AUROCs of 0.79, 0.58 and 0.66. Predictions generalized well to an external  
44 cohort with AUROCs of 0.93, 0.81 and 0.73 respectively. Additionally, an HRD classifier trained on  
45 breast cancer yielded an AUROC of 0.78 in internal validation and was able to predict HRD in  
46 endometrial, prostate and pancreatic cancer with AUROCs of 0.87, 0.84 and 0.67 indicating a shared  
47 HRD-like phenotype is across tumor entities.

48 **Conclusion:** In this study, we show that HRD is directly predictable from H&E slides using attMIL within  
49 and across ten different tumor types.

50  
51 **Keywords:** Homologous Recombination Deficiency, Deep Learning, DNA repair mechanism, artificial  
52 intelligence, molecular pathology, pan cancer study.

53

## 54 **Background**

55 Homologous recombination repair (HRR) is a DNA repair mechanism that ensures genomic integrity  
56 after DNA double-strand breaks (DSB), which occur regularly during the cell cycle (1). Homologous  
57 recombination deficiency (HRD) results in defective DNA break repair, increased somatic copy number  
58 alterations and genomic instability, driving malignant transformation and causing cancer (2). Poly(ADP-  
59 Ribose)-polymerase (PARP) plays pivotal roles in base excision repair of single strand DNA breaks  
60 (SSDBs), which is a compensatory DNA repair mechanism in the context of HRD. In the setting of  
61 homologous recombination (HR) proficiency PARP inhibition results in the accumulation of unrepaired  
62 SSDBs. These can eventually convert to DSBs, which can be repaired via HR thus maintaining  
63 genomic integrity and cell viability. However in the case of a HRD tumor, PARP inhibition-induced DSBs  
64 are no longer repaired, resulting in direct cytotoxicity.. This phenomenon of *synthetic lethality* is the  
65 reason why HRD is an important biomarker to select patients for PARP inhibitor (PARPi) treatment in  
66 several tumor types, especially in breast, ovarian, prostate and pancreatic cancer (3–6). Prevalences of  
67 HRD varies according to the genomic definition of HRD and among tumor types, ranging from 0% in  
68 thymoma or thyroid cancer up to 70% in ovarian cancer(7). The use of PARPi has led to improved  
69 disease-free survival in multiple clinical trials by increasing platinum sensitivity in ovarian (OV) and  
70 breast cancer (BRCA), and other tumor types (8,9).

71  
72 The success of PARPi therapy is mainly limited by the challenge of diagnosing HRD. Many different  
73 test strategies are available. The most robust test for HRD are oncogenic mutations in the Breast  
74 Cancer genes 1 and 2 (*BRCA1/2*) (10,11). However, this approach excludes patients without *BRCA1/2*-  
75 related deficiencies in the HR pathway (12). Moreover, other mechanisms such as epigenetic  
76 modifications, germline and somatic mutations of genes related or non related to the HRR pathway may  
77 cause HRD (13). Unfortunately, non-*BRCA* HR mutations have not been reliably shown to predict HRD  
78 or PARPi benefit in the clinic. Certain patterns of mutations, like the single base substitution 3 (SBS3)  
79 are also associated with a defective HR and therefore a potential biomarker (14,15). Finally, another  
80 strategy for detecting HRD is to look for the consequence of HRD rather than the cause. This approach  
81 uses whole genome sequencing single nucleotide polymorphism (SNP) array data to identify loss of

82 heterozygosity (LOH), telomeric allelic imbalance (TAI) and large-scale state transitions (LST), also  
83 defined as a genomic instability score (GIS). This combined score has been validated in randomized  
84 clinical trials as predictive of PARPi benefit (16–18). Biologically, this methods provides a more  
85 comprehensive assessment of genomic instability caused by HRD, rather than scores exclusively  
86 based on mutation or HRR genes (Figure 1A). However, the GIS is not yet implemented in routine  
87 diagnostics in clinical workflows (11,12,19). Combining the different components of HRD using  
88 algorithms (e.g. scarHRD, HRDetect, CHORD) may be the gold standard to determine the genomic  
89 “scar” associated with HRD (20–22). A non-DNA-based way to determine HRD is using a functional test  
90 such as the RAD51 focus formation assays (23,24). U.S. Food and Drug Administration (FDA)-  
91 approved genetic tests for HRD typically rely on a combination of alterations in *BRCA1/2* genes and  
92 LOH (FoundationOne CDx, Foundation Medicine, Inc., Cambridge, MA) or GIS (myChoice CDx, Myriad  
93 Genetics Laboratories, Inc., Salt Lake City, UT) (10,11). However defining cut-off values for  
94 stratification between positive and negative cases is difficult (7,25). Taken together, the HRD testing  
95 landscape is highly complex. Many different tests coexist and they are not perfectly concordant. There  
96 is a high clinical need for a cheap, fast and standardized HRD test which captures a breadth of  
97 biological processes and not just alterations in individual genes. In this study, we hypothesized that the  
98 tumor phenotype as observed on histological whole slide images (WSI) of tumors reflects the HRD  
99 status and can be used to diagnose HRD.

100  
101 Deep Learning (DL) is an artificial intelligence (AI)-based technology which has emerged as a powerful  
102 method to quantitatively mine data from histological WSI of tumors in the last five years. DL enables us  
103 to detect genetic alterations directly from histopathological image data (26–28). Specifically, DL has  
104 been shown to detect single mutations(29,30), as well as phenotypic manifestation of DNA instability  
105 mechanisms such as microsatellite instability (MSI), just by processing scanned WSI of tumor tissue  
106 stained with H&E (31,32). Today, several DL systems to predict genetic alterations and clinical  
107 outcomes have received regulatory approval and are available for routine diagnostic use in Europe and  
108 the USA (33,34). Some smaller pilot studies have shown encouraging data for DL-based prediction of  
109 HRD from H&E WSI (35,36). However, HRD is a pan-cancer biomarker and DL has not been  
110 systematically used to diagnose HRD across tumor types directly from routine H&E pathology slides.

111  
112 Therefore, in the present study, we developed a DL system to predict HRD status directly from H&E  
113 pathology slides. We used the state-of-the-art technology “attention-based Multiple Instance Learning”  
114 (attMIL) in a weakly supervised experimental setup, using no spatial labels or manual annotations  
115 whatsoever (28) to train the DL system, we used the calculated scarHRD, one of the most  
116 comprehensive HRD scores which integrates a variety of genomic changes (Figure 1B). We trained and  
117 evaluated the DL classifiers by cross-validation in a large cohort of n=4,113 patients from The Cancer  
118 Genome Atlas (TCGA), comprising 10 types of solid tumors. The models were then externally validated  
119 on four cancer types in an independent validation dataset (n=474) in a tumor-wise and cross-cancer  
120 experimental approach (Figure 1C). Taken together, our experimental results provide direct evidence  
121 that HRD is detectable from routine histology in different types of cancer with DL.

## 122 **Methods**

### 123 **Data Acquisition**

124 In total data from 5,155 patients of 10 tumor types from The Cancer Genome Atlas (TCGA) and 573  
125 patients from five tumor types from the Clinical Proteomic Tumor Analysis Consortium (CPTAC, Figure  
126 1C) were obtained from <https://www.cbioportal.org/>. Accordingly, the cancer types included in the  
127 present study were breast invasive carcinoma (TCGA-BRCA n=1,058), colorectal cancer (TCGA-CRC  
128 n=580), glioblastoma (TCGA-GBM n=420, CPTAC-GBM n=99), liver hepatocellular carcinoma (TCGA-  
129 LIHC n=364), lung adenocarcinoma (TCGA-LUAD n=536, CPTAC-LUAD n=111), lung squamous cell  
130 carcinoma (TCGA-LUSC n=497; CPTAC-LSCC n=109), ovarian cancer (TCGA-OV n=520), pancreatic  
131 adenocarcinoma (TCGA-PAAD n=177; CPTAC-PDA n=153), prostate adenocarcinoma (TCGA-PRAD  
132 n=488) and endometrial carcinoma (TCGA-UCEC n=515, CPTAC-UCEC n=101, Supplementary Figure  
133 1A,B). Image data and clinical data were available in TCGA-BRCA for n=1005, TCGA-CRC for n=496,  
134 TCGA-GBM for n=232, CPTAC-GBM for n=99, TCGA-LIHC for n=348, TCGA-LUAD for n=460,  
135 CPTAC-LUAD for n=106, TCGA-LUSC for n=451, CPTAC-LSCC for n=108, TCGA-OV for n=90,  
136 TCGA-PAAD for n=173, CPTAC-PDA for n=139, TCGA-PRAD for n=391, TCGA-UCEC for n=467 and  
137 CPTAC-UCEC for n=99, therefore leaving us in total with n=4,565 patients for the analysis (Figure 1C,

138 Supplementary Figure 1A,B). Moreover, some figures were created using <https://www.cbioportal.org/>  
139 (37,38). For additional experiments on *BRCA1/2* mutational status we retrieved data from Riaz et al  
140 previously published paper (39). Estrogen receptor data for the subgroup analysis was only available  
141 for n=661 patients in the TCGA-BRCA cohort.

142

## 143 **Image Preprocessing**

144 WSIs were downloaded for the above mentioned cohorts from the GDC Portal  
145 (<https://portal.gdc.cancer.gov/>) and The Cancer Imaging Archive (TCIA,  
146 <https://www.cancerimagingarchive.net/>). Initially, the images were tessellated into patches with an edge  
147 length of 256  $\mu\text{m}$  and a resolution of 224x224 pixels. Secondly, the patches for each cohort were color  
148 normalized using the Macenko spectral matching technique(40) to enforce a standardized color  
149 distribution across cohorts. To train the prediction models, we used our in-house open-source DL  
150 pipeline “marugoto” (<https://github.com/KatherLab/marugoto>) consisting of a self-supervised learning  
151 (SSL) model using a pre-trained ResNet50 architecture with ImageNet weights, fine-tuned pan-cancer  
152 on approximately 32.000 WSI to extract a 2048-dimensional feature vector for each patch per patient  
153 (41). To obtain patient-level predictions, 512x2048 feature matrices (MIL bags) are constructed by  
154 concatenating 512 feature vectors selected at random per patient and fed into an attMIL framework with  
155 the following architecture: (512x256), (256x2) with a subsequent attention mechanism (Figure 1B).  
156 (42,43)

## 157 **Calculation of HRD Scores**

158 For the patient-wise calculation of HRD, single nucleotide polymorphism (SNP) data, generated by the  
159 Allele-Specific Copy number Analysis of Tumors (ASCAT) algorithm, was downloaded from the  
160 Genomic Data Commons (GDC) Portal: <https://portal.gdc.cancer.gov/> (accessed 06/15/2022). In  
161 CPTAC, the respective data was only available for the CPTAC-3 cohort. The HRD score was calculated  
162 using the scarHRD (<https://github.com/sztup/scarHRD>), as described in previous studies (20,44).  
163 ScarHRD uses whole genome sequencing data in the form of SNP arrays to calculate the three  
164 subscores LOH, LST and TAI. The sum of these subscores makes up the patient-wise HRD score

165 (Figure 1A). The cut-offs of the different subscores have been previously defined by Abkevich et al. for  
166 LOH, Popova et al. for LST and Birkbak et al. for TAI (16–18). Adding up the LOH, LST and TAI scores,  
167 patients can be divided into HRD high (HRD-H) and HRD low (HRD-L) at a cut-off of 42 (7). All patients  
168 in the CPTAC-GBM cohort were HRD-L. Hence, we excluded them from further analysis  
169 (Supplementary Figure 1A,B).

## 170 **Experimental Design**

171 In our study, we performed three main experiments (Figure 1B). To assess the baseline predictability of  
172 HRD from routine histology, we first trained an HRD classifier in a within-cohort approach using five-  
173 fold-cross-validation within each of 10 tumor entities mentioned above in the TCGA cohorts (internal  
174 validation). This was achieved by randomly splitting each cohort on the level of patients, creating non-  
175 overlapping training and test sets for model training. The ratio for splitting the training and testing set  
176 was 80:20 of the entire dataset, and the training and validation set was split 75:25 of the training set.  
177 Thus, the absolute split for training, internal validation and internal testing was 60, 20 and 20,  
178 respectively. Five different models were trained until each part was used as a test set once. Thus, no  
179 data leakage from the test set to the training set occurred. This process was repeated individually for  
180 each cancer type in the TCGA cohorts. A weighted cross-entropy loss function was used to assist the  
181 model with the imbalanced dataset. Secondly, we deployed the five models trained in the first  
182 experiments on the same tumor type from the CPTAC cohorts as an external validation. By utilizing this  
183 approach, we circumvent any potential claims of selecting the model with the highest AUROC in the  
184 external validation. Lastly, we trained an HRD classifier on the TCGA-BRCA cohort, which had the  
185 highest number of patients, and deployed it on all other TCGA cohorts (CRC, GBM, LIHC, LUAD,  
186 LUSC, PRAD, PAAD, OV, UCEC) as well as on all CPTAC cohorts (LUAD, LSCC, PDA, UCEC). In our  
187 study, we aimed to evaluate the performance of the models using the AUROC, which is commonly used  
188 for assessing the accuracy of binary classification tasks. Our primary statistical endpoint was the  
189 AUROC +/- 95%-confidence interval (CI) and Area under the precision recall curve (Supplementary  
190 Table 1). To further assess the performance of each model, we used a two-sided t-test to compare the  
191 patient-level prediction scores between the HRD-H and HRD-L patient groups as defined by the ground  
192 truth and report the p-values, assuming a significance level of 0.05 as statistically significant, without



193 correction for multiple testing (Supplementary Table 1). As a final step to obtain a more profound  
194 understanding of the TCGA-BRCA cohort, we uploaded our custom HRD-H and HRD-L ground truth  
195 and predicted subgroups in cbioportal to examine the characteristics of these cases in the TCGA-BRCA  
196 PanCancer Atlas cohorts.

197

## 198 **Explainability**

199 To visualize the output of our model, we created high resolution heat maps that show the spatial  
200 distribution of our model's attention and prediction scores on the WSI. Therefore, using RetCCL  
201 convolutional neural network image feature vectors for 32x32 pixel fields were extracted from the WSI.  
202 We then calculated attention and classification scores for each image region and normalized them  
203 across the distribution of scores within each patient cohort. Based on these scores, color heatmaps for  
204 each patient, with red indicating high attention or a positive classification and blue indicating low  
205 attention or a negative classification were generated. To ensure interpretability of the underlying  
206 morphology together with the attention and classification scores, we separately reconstructed the final  
207 attention and classification heatmaps by blending the raw color heatmaps with the image features. This  
208 approach allows us to interpret the output of our model in a way that is easy to understand and provides  
209 insight into the underlying morphology of the tumor.

210

## 211 **Results**

### 212 **HRD is predictable from histology with attMIL**

213 First, we investigated whether DL could predict HRD status from H&E types within 10 different types of  
214 cancer from the TCGA cohort. We used cross-validation on the level of patients to train and test an  
215 attMIL-based DL model within each cohort. In our dataset, the prevalence of HRD ranged from 3% in  
216 glioblastoma (GBM) up to 63% in OV (Supplements Figure 1C). We found that in five out of 10 cancer  
217 types, the mean prediction AUROC was above 0.6, and the 95% CI of the fold-wise HRD prediction  
218 AUROCs remained above the null hypothesis of 0.5. Among these, HRD prediction reached statistical



219 significance with a p-value below 0.05 in three cancer types: endometrial cancer (UCEC, AUROC  
220 0.79+/-0.04, p=0.0008), breast cancer (BRCA, AUROC 0.78+/-0.02, p<0.0001) and lung  
221 adenocarcinoma(LUAD, AUROC 0.66+/-0.05, p=0.02; Figure 2A). AUPRC values are reported in the  
222 Supplementary Table 1. Prediction of HRD was not possible in LUSC, LIHC, GBM, as their prediction  
223 AUROCs did not exceed the baseline (0.55+/-0.04 0.56+/-0.14, 0.58+/-0.38) with CIs above the null  
224 hypothesis or p-values below 0.05 (Supplementary Figure 2 A-J, Supplementary Table 1). For the  
225 tumor types PAAD, OV and PRAD, the AUROCs ranged from 0.58+/-0.22 to 0.6+/-0.09 to 0.76+/-0.22.  
226 Together, these data demonstrate that DL can predict HRD status from histology images alone in  
227 several tumor types.

## 228 **HRD is predictable from H&E histology with attMIL in an independent test set**

229 A step that is germane to the successful development of DL models is external validation with WSIs  
230 from patient cohorts which are completely independent from the training set (45). Hence, for our  
231 external validation experiments, we deployed the classification models obtained from the cross-  
232 validation training on TCGA to analyze cohorts from the CPTAC dataset corresponding to the same  
233 cancer type. External validation cohorts in CPTAC were available for endometrial cancer (UCEC),  
234 pancreatic cancer (PDA), lung adenocarcinoma (LUAD), and lung squamous cell carcinoma (LSCC). In  
235 these external validation experiments, we noted that the prediction performance was higher compared  
236 to internal validation experiments. Once again, the best performance was obtained in UCEC, with an  
237 AUROC of 0.93+/-0.07, p=0.01. In LUAD, the performance increased in the external validation, yielding  
238 an AUROC of 0.73+/-0.11 and a significant p-value of 0.03. In the case of PAAD/PDA, where the  
239 internal validation was unsuccessful (internal validation AUROC 0.58+/-0.22), the external validation  
240 resulted in an improved AUROC reaching 0.81+/-0.14, albeit with a p-value of 0.07. Meanwhile, in  
241 LUSC/LSCC, no improvement in performance was observed in the external validation set compared to  
242 the internal training set (AUROC 0.57+/-0.01, p=0.23, Figure 2A, Supplementary Figure 2 K-N).  
243 Together, these data show that DL-based classifiers of HRD status generalize beyond the training  
244 cohort.

## 245 **A HRD classifier trained on BRCA detects HRD across various types of cancer**

246 Lastly, we aimed to investigate if HRD-related morphological features in one cancer type can help to  
247 predict HRD status in another cancer type. This would point to a shared set of morphological features  
248 across cancer types, potentially allowing us to develop a pan-cancer pathology-based prediction  
249 system for HRD status. To test this, we applied our trained HRD classifiers in a cross-cancer  
250 experimental design. We used the breast cancer cohort TCGA-BRCA to train the HRD classification  
251 model because this cohort had the highest number of patients. Subsequently, we deployed this model  
252 on all other cohorts obtained from the TCGA and CPTAC datasets. Surprisingly, the BRCA-based  
253 model was able to significantly predict HRD from non-BRCA tissue in UCEC, PRAD and PAAD. For  
254 those three cohorts, the external deployment of a BRCA-based model resulted in higher prediction  
255 AUROCs than the respective internal validation experiments, reaching AUROCs of  $0.70\pm 0.02$ ,  
256  $p < 0.001$  in TCGA-UCEC,  $0.84\pm 0.07$  and  $p = 0.004$  in TCGA-PRAD  $0.67\pm 0.03$ ,  $p = 0.2$  in TCGA-PAAD,  
257  $0.87\pm 0.1$   $p = 0.05$  in CPTAC-UCEC and  $0.65\pm 0.02$   $p = 0.26$  in CPTAC-PDA, respectively (Figure 2B).  
258 In the tumor types LUAD and OV, the AUROCs remained with  $0.62\pm 0.03$  for TCGA-LUAD,  $0.66\pm$   
259  $0.06$  for CPTAC-LUAD and  $0.61\pm 0.03$  in TCGA-OV in a similar range to the internal validation results  
260 (Supplementary Figure 3A-M). Together, these data show that a classifier trained on breast cancer can  
261 predict HRD status from histology in other tumor types, indicating a shared “HRD morphology” between  
262 tumor types.

## 263 **Molecular and histomorphological characterization of TCGA-BRCA HRD-H and** 264 **HRD-L cases**

265 Finally, we investigated which molecular and morphological patterns were associated with ground truth  
266 and DL-predicted HRD status. We used the TCGA-BRCA cohort to analyze this in detail, as this was  
267 the largest cohort. We observed that in the HRD-H subgroup, 45% were classified as basal-like breast  
268 cancers, 11% as HER2-enriched, 15% as Luminal A, and 26% as Luminal B. In contrast, only 7% of the  
269 cases in the HRD-L subgroup were basal-like, 7% were HER2-enriched, 64% were Luminal A, and 18%  
270 were Luminal B (Figure 3A) (46). Within our predicted groups, we observed a similar distribution among  
271 the BRCA subtypes (Figure 3B).

272 To reassure that our model predicts HRD detached from phenotypic differences of estrogen receptor  
273 negative (ER-) vs. ER-positive (ER+) breast cancers we calculated the receiving operating curve (ROC)  
274 and precision recall curve (PRC) for the subgroups: ER+/HER2+, ER+/HER2-, ER-/HER2+, ER-/HER2-  
275 , indicating that HRD was also predictable with AUROCs of 0.66+/-0.3, 0.8+/-0.09, 0.72+/-0.43 and  
276 0.62+/-0.11 (Supplementary Figure 4A-H). Our analysis of the mutational landscape of both HRD-H and  
277 HRD-L ground truth revealed that *TP53* had the highest alteration frequency with 67% in the HRD-H  
278 ground truth group, significantly higher than 20% in the HRD-L group, following alterations in the *TTN*  
279 (26% vs. 14%) gene. In contrast, the most enriched alterations in the HRD-L group were observed in  
280 the genes *PIK3CA* (39%) followed by *CDH1* (16%), *GATA3* (14%) and *MAP3K1* (11%), whereas the  
281 prevalences in the HRD-H group of *PIK3CA*, *CDH1*, *GATA3* and *MAP3K1* were 19%, 2%, 6% and 1%,  
282 respectively (Figure 3C). For the HRD-H prediction subgroup alteration frequencies for *TP53*, were  
283 significantly higher at 77% (Figure 3D). Such divergences were not as noticeable in the HRD-L  
284 prediction group. These findings suggest that there are notable differences in alteration frequencies  
285 between the two subgroups, which are consistent across both the ground truth and prediction data.  
286 Moreover, we compared the HRD-H prediction score to the alteration status of somatic and germline  
287 mutations in the *BRCA1/2* genes, whereupon we saw that there was a significant difference between  
288 the mutant and wild-type cases for *BRCA1* germline and *BRCA2* somatic mutations (Figure 3E).  
289 Methylation data indicated that the HRD-H group had most of its methylation alterations in the N-shore  
290 portion of the *BRCA1* promoter region, whereas those in the HRD-L group were mainly located in the S-  
291 shore portion (Supplementary Figure 4I). Lastly, we proceeded to investigate the histomorphological  
292 patterns associated with the presence of HRD through whole slide prediction heatmaps in CPTAC-  
293 UCEC (Figure 4A-C). Our findings revealed that high grade, fibrosis, hemorrhage and lymphocytic  
294 infiltration are consistent features predictive of HRD across various tumor types, as shown in Figure 4  
295 for BRCA and UCEC, particularly in the top predicted HRD-H tiles for the top three patients. Fibrosis  
296 was observed in HRD-positive cases, particularly in BRCA (Figure 4D). Moreover, hemorrhagic  
297 necrosis especially adjacent to tumor tissue and tumor stroma was consistently observed as highly  
298 predictive areas in the true HRD-H cases across various cancer types. (Supplementary Figure 5,6). In  
299 summary, these data show that known HRD morphology characteristics were found in our DL based  
300 top predicted HRD-H patients.

## 301 Discussion

302 HRD has recently emerged as an important pan-cancer biomarker for targeted treatment in solid tumors  
303 (11,47). The assessment of HRD clinically, albeit indicated for all patients with gynecological tumors,  
304 remains challenging. This is due to the given availability of different methods with limited agreement  
305 and whose logistic complexities and inherent costs pose significant hurdles for their adoption. In this  
306 light, a pan-cancer test of HRD by DL-based image analysis on pathology slides could be a useful pre-  
307 screening tool and reduce the load of genetic tests.

308  
309 In this study, we demonstrated that DL can predict HRD status from H&E histology in different tumor  
310 types in both within-cohort and external validation experiments. Surprisingly, our findings revealed that  
311 a BRCA-based classifier could detect HRD from H&E slides across different tumor entities. As  
312 expected, the HRD prediction was significantly lower in tumors with a low prevalence of HRD. Our  
313 classifier has identified histomorphological characteristics such as hemorrhagic necrosis at tumor  
314 margins, lymphocyte infiltration, fibrosis, and high tumor cell density that are associated with HRD in  
315 BRCA(36). These findings validate the efficacy of our classifier. Moreover, despite having trained our  
316 classifier solely on BRCA, its consistent identification of HRD-associated morphological patterns across  
317 different tumor entities reiterates the value of our tool for broader applications. Compared to previous  
318 studies, we here show a pan-cancer DL-based prediction of a more comprehensive HRD score  
319 calculated from LOH, TAI, and LST as ground truth directly from H&E tumor slides. (35,36)

320  
321 Our morphological analysis showed that UCEC or PAAD, achieved better predictive results compared  
322 to LUSC or LIHC, a trend previously observed in pan-cancer studies (30,48). In general, tumors with a  
323 complex structure, such as adenocarcinomas are morphologically susceptible to genetic alterations  
324 than solid tumors growing in rather syncytial patterns. HRD-positive tumors barely resemble glandular  
325 tissue anymore, which might be their main distinctive feature and therefore a potential explanation for  
326 this constellation. Additional studies with larger patient cohorts would be required to confirm this. A  
327 closer look at the TCGA-BRCA subgroups revealed that predicted HRD-H is more common in triple-  
328 negative breast cancer, which is known for its poor prognosis and resistance to conventional

329 chemotherapy. In line with their ground truths, the majority of those patients were predicted to be HRD-  
330 positive by our classifier (Figure 3A,B) (46). Furthermore, clear molecular pathological differences were  
331 found in the two subgroups. Specifically, the HRD-H subgroup is characterized by *TP53* alterations,  
332 while the HRD-L subgroup has a higher frequency of *PIK3CA* alterations, suggesting an interactive  
333 effect between the *TP53* mutated cases and HRD-H patients (49,50). This is particularly true for  
334 *BRCA1* mutated cancers, where HRD-H was predicted significantly better than in *BRCA1* wildtype  
335 cases (Figure 3E) (51).

336  
337 Recently, the EMA and FDA granted the first approval to use PARPi therapy for HRD positive ovarian  
338 cancer patients. Clinical trials with promising interim data are also underway for other tumor entities and  
339 further approvals are expected in the future. Despite the evident link between HRD and *BRCA1/2*  
340 mutations, it is now well established that the total number of HRD-positive patients significantly exceeds  
341 the total number of *BRCA*-mutated patients in various cancer types (22,52). The patients who fall into  
342 this diagnostic gap can be identified with comprehensive HRD testing, as proposed in our study.<sup>41</sup> HRD  
343 testing would thus complement *BRCA1/2* testing as a biomarker test for PARPi use, such as with AI-  
344 based screening methods as applied here. Moving diagnostic routines towards phenotype-based  
345 instead of inconsistent, alteration-based HRD detection methods might extend our ability to identify  
346 patients who may benefit from PARPi and enroll them in clinical trials. Our study provides a proof of  
347 concept that there is indeed a pan-cancer preserved HRD morphology in histology slides which could  
348 potentially serve as an HRD marker. Prospective trials are needed to evaluate an AI-based HRD score  
349 as a biomarker to guide treatment decisions, potentially in a two-step approach leading to lower  
350 sequencing requirements and cost reduction.

## 351 Limitations

352 Our study has several limitations. Firstly, the sample sizes of our cohorts, particularly the CPTAC  
353 dataset, are relatively small. Moreover, the variation within the distribution of HRD prevalences between  
354 tumor types can result in class imbalances. Although the effect of imbalanced datasets on the accuracy  
355 of our classifiers was addressed via weighing techniques during the model training process, this could  
356 still have an effect on the statistical power of the results, as well as the generalisability of our models to

357 a larger population. We observed higher AUROCs in the external validation cohort, which may be  
358 attributed to the smaller size and higher class imbalance in the test set. Further studies with larger  
359 patient cohorts are required to validate our findings. Furthermore, the quality of the data from the TCGA  
360 and CPTAC cohorts may vary, thus potentially impacting the accuracy of our predictions.

## 361 **Conclusion**

362 Our findings provide evidence that DL has the potential to not only contribute but improve diagnostic  
363 HRD testing, potentially saving time and costs as well as improving outcomes for patients by identifying  
364 subgroups who may benefit from targeted therapy. Current clinical practices face challenging factors  
365 such as high cost, time consumption, lack of availability, and inconsistency in HRD status screening  
366 methods. These logistic, analytic and financial challenges contribute to the partial identification of  
367 cancer patients who may benefit from PARPi therapy and to the limited genetic testing, which is further  
368 compounded by the panoply of HRD status assessment methods whose inter-assay concordance is  
369 limited. With the aid of AI, we have the opportunity to identify these subgroups and improve patient  
370 outcomes.

371

## 372 **List of Abbreviations**

373 **AI:** artificial intelligence

374 **ASCAT:** Allele-Specific Copy number Analysis of Tumors

375 **attMIL:** attention-weighted multiple instance learning

376 **AUROC:** Area Under the Receiver Operating Characteristic curve

377 **BRCA:** breast invasive carcinoma

378 **BRCA1/2:** Breast Cancer genes 1 and 2

379 **CI:** confidence interval

380 **CIOMS:** Council for International Organizations of Medical Sciences

381 **CPTAC:** Clinical Proteomic Tumor Analysis Consortium

382 **CRC:** colorectal cancer

- 383 **DL:** Deep Learning
- 384 **DSB:** DNA double-strand breaks
- 385 **ER-:** estrogen receptor negative
- 386 **ER+:** estrogen receptor positive
- 387 **FDA:** U.S. Food and Drug Administration
- 388 **GBM:** glioblastoma
- 389 **GDC:** Genomic Data Commons
- 390 **GIS:** genomic instability score
- 391 **H&E:** Hematoxylin & Eosin
- 392 **HR:** Homologous recombination
- 393 **HRD-H:** HRD high
- 394 **HRD-L:** HRD low
- 395 **HRD:** Homologous Recombination Deficiency
- 396 **HRR:** Homologous recombination repair
- 397 **LIHC:** liver hepatocellular carcinoma
- 398 **LOH:** loss of heterozygosity
- 399 **LSCC:** squamous cell carcinoma of the lung
- 400 **LST:** large-scale state transitions
- 401 **LUAD:** adenocarcinoma of the lung
- 402 **LUSC:** squamous cell carcinoma of the lung
- 403 **OV:** ovarian cancer (OV)
- 404 **PAAD:** pancreatic adenocarcinoma
- 405 **PDA:** pancreatic adenocarcinoma
- 406 **PARP:** Poly(ADP-Ribose)-polymerase
- 407 **PARPi:** Poly(ADP-Ribose)-polymerase inhibitor
- 408 **PRAD:** prostate adenocarcinoma
- 409 **PRC:** precision recall curve
- 410 **ROC:** receiving operating curve
- 411 **SBS3:** single base substitution 3



412 **SNP:** single nucleotide polymorphism

413 **SSDBs:** single strand DNA breaks

414 **SSL:** self-supervised learning

415 **TAI:** telomeric allelic imbalance

416 **TCGA:** The Cancer Genome Atlas

417 **TRIPOD:** Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

418 **UCEC:** endometrial carcinoma

419 **WSI:** whole slide images

420

421

## 422 **Declarations**

### 423 **Ethics statement**

424 The experiments in this study were carried out according to the Declaration of Helsinki and the  
425 International Ethical Guidelines for Biomedical Research Involving Human Subjects by the Council for  
426 International Organizations of Medical Sciences (CIOMS). The present study also adheres to the  
427 “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis”  
428 (TRIPOD) statement.<sup>20</sup> The Ethics Board at the Medical Faculty of Technical University Dresden (BO-  
429 EK-444102022) approved of the overall analysis in this study. The patient sample collection in each  
430 cohort was separately approved by the respective institutional ethics board.

### 431 **Data and Code availability**

432 The WSI, molecular and clinical data for TCGA and CPTAC cohorts are publicly accessible at  
433 <https://portal.gdc.cancer.gov/> and <https://www.cbioportal.org/> (accessed, 08 March 2022). Script for  
434 calculating the HRD score is available under <https://github.com/sztup/scarHRD> (accessed 06 June  
435 2022). All other source codes can be downloaded under <https://github.com/KatherLab/marugoto>. Our  
436 calculated HRD score is publicly available in Supplementary Table 2. Moreover, our custom TCGA-

437 BRCA HRD-H and HRD-L group can be accessed for the PanCancer Atlas cohort at  
438 <https://www.cbioportal.org/> (Supplementary Table 3).

### 439 **Competing Interests**

440 JNK reports consulting services for Owkin, France, Panakeia, UK and DoMore Diagnostics, Norway  
441 and has received honoraria for lectures by MSD, Eisai and Fresenius. JSRF reports a leadership (board  
442 of directors) role at Grupo Oncoclinicas, stock or other ownership interests at Repare Therapeutics and  
443 Paige.AI, and a consulting or Advisory Role at Genentech/Roche, Invicro, Ventana Medical Systems,  
444 Volition RX, Paige.AI, Goldman Sachs, Bain Capital, Novartis, Repare Therapeutics, Lilly, Saga  
445 Diagnostics, Swarm and Personalis. No other potential conflicts of interest are reported by any of the  
446 authors.

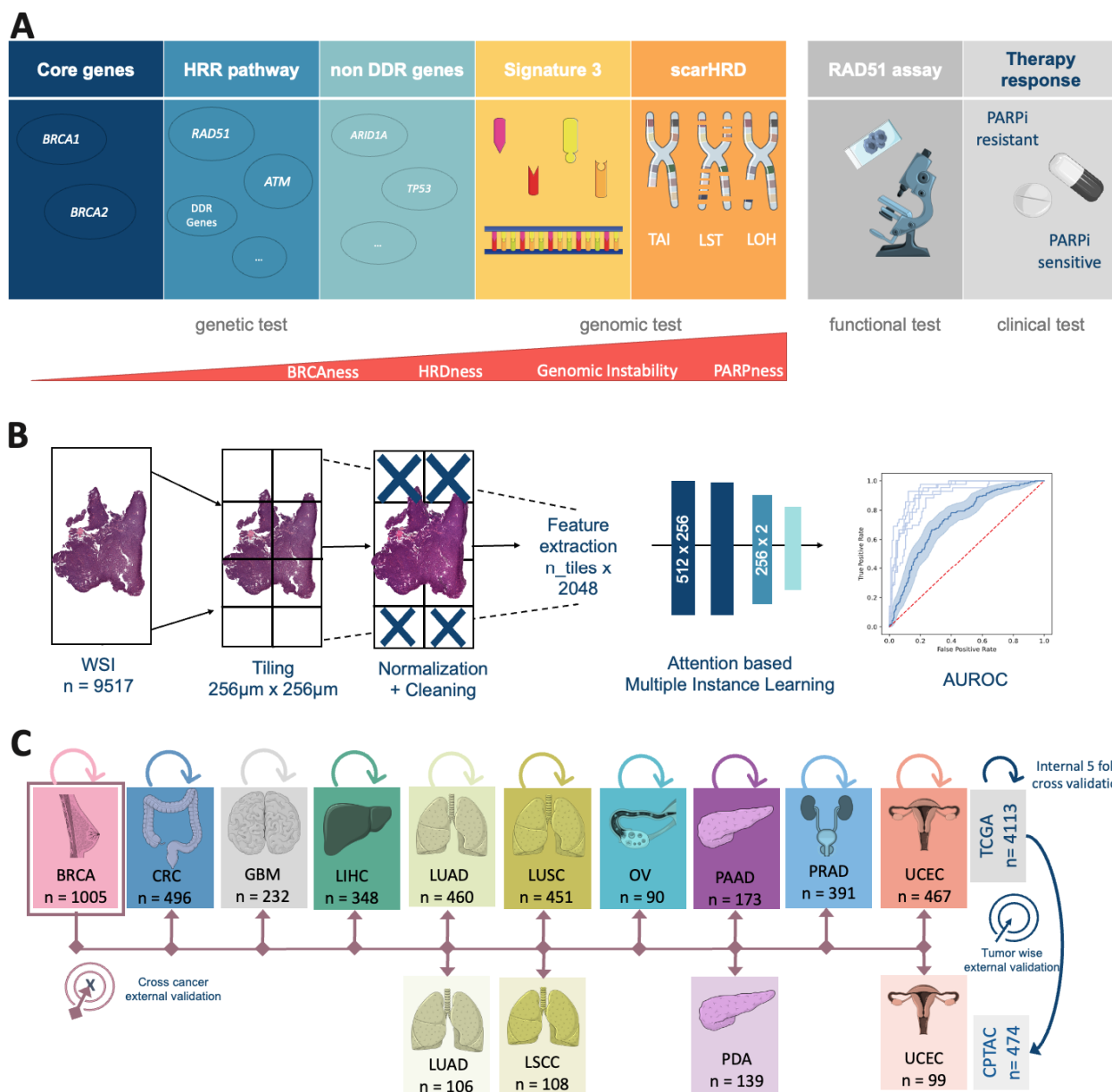
### 447 **Funding**

448 JNK is supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and  
449 the Max-Eder-Programme of the German Cancer Aid (grant #70113864), the German Federal Ministry  
450 of Education and Research (PEARL, 01KD2104C), and the German Academic Exchange Service  
451 (SECAI, 57616814). This research was supported by the National Institute for Health and Care  
452 Research (NIHR, NIHR213331) Leeds Biomedical Research Centre. The views expressed are those of  
453 the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social  
454 Care. JSRF is funded in part by the Breast Cancer Research Foundation, a Susan G Komen  
455 Leadership Grant, the NIH/NCI P50 CA247749 01 grant and by the NIH/NCI Cancer Center Core Grant  
456 P30-CA008748.

### 457 **Author Contribution**

458 CMLL, NOB, HSM, and JNK conceptualized the study. TS, CMLL, HSM and NOB curated the source  
459 data. MVT developed the source codes for the analysis. OSMEN, MG and CMLL conducted the  
460 experiments. CMLL interpreted the data and wrote the first draft of the manuscript. All authors revised  
461 the manuscript draft, contributed to the interpretation of the data and agreed to the submission of this  
462 paper.

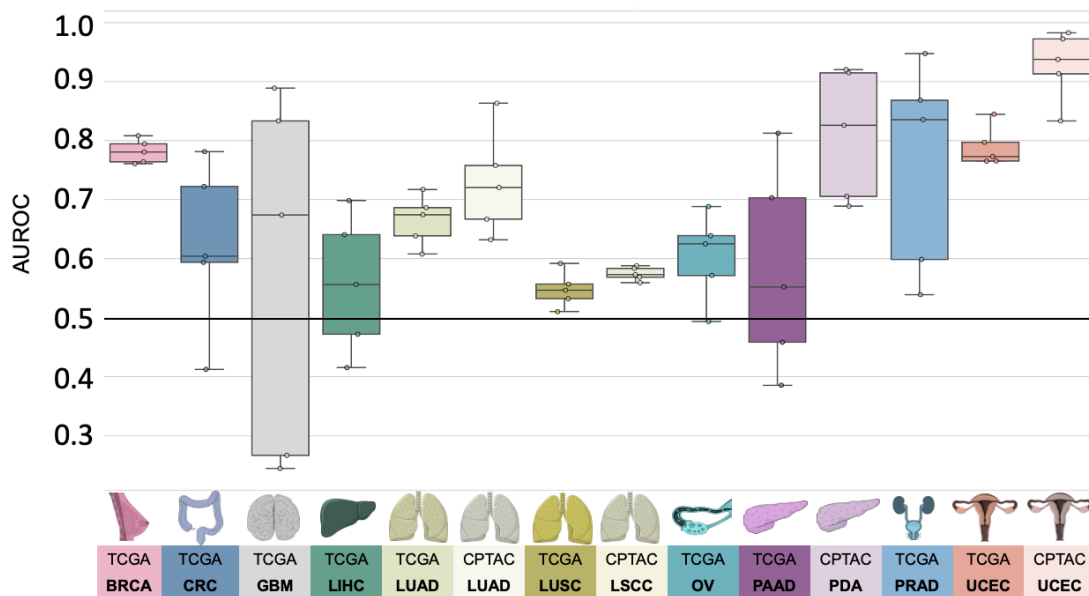
463 **Figures**



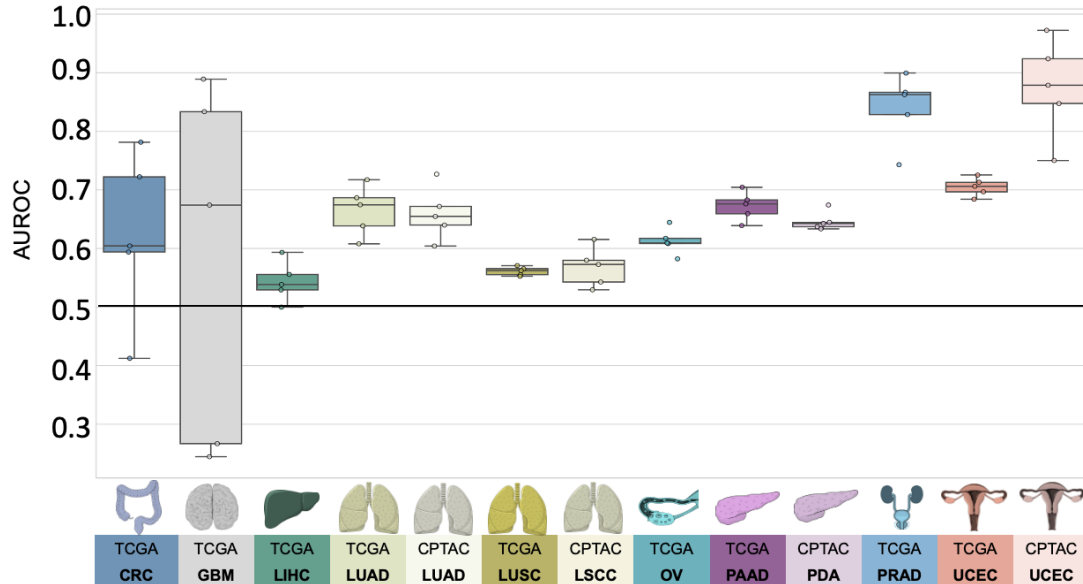
464  
 465 **Figure 1: Experimental Design and Study overview.** (A) Overview of the different Homologous  
 466 Recombination Deficiency (HRD) scores, their content and assessment methods. (B) Workflow of our  
 467 Deep Learning (DL) pipeline. A total of n=9517 Whole Slide Images (WSI) were processed and trained  
 468 with an attention-based Multiple Instance Learning (attMIL) approach. The statistical endpoint was the  
 469 Area under the receiving operating curve (AUROC). (C) Study design for the three main experiments  
 470 (Internal 5-fold cross-validation, tumor-wise external validation and cross-cancer external validation)  
 471 conducted and cohort overview for patients and tumor types included from The Cancer Genome Atlas

472 (TCGA, n=4113 patients) and Clinical Proteomic Tumor Analysis Consortium (CPTAC, n=474 patients).  
 473 Abbreviations: BRCA=breast cancer; CRC=colorectal cancer; GBM=glioblastoma; LIHC=liver cancer;  
 474 LUAD=lung adenocarcinoma; LUSC/LSCC=lung squamous cell carcinoma; OV=ovarian cancer;  
 475 PAAD/PDA=pancreatic adenocarcinoma; PRAD=prostate adenocarcinoma; UCEC=endometrial cancer;  
 476 HRR=Homologous recombination repair. (This Figure was partly generated using Servier Medical Art,  
 477 provided by Servier, licensed under a Creative Commons Attribution 3.0 unported license)  
 478

**A** Internal validation on TCGA and tumor-wise external validation on CPTAC

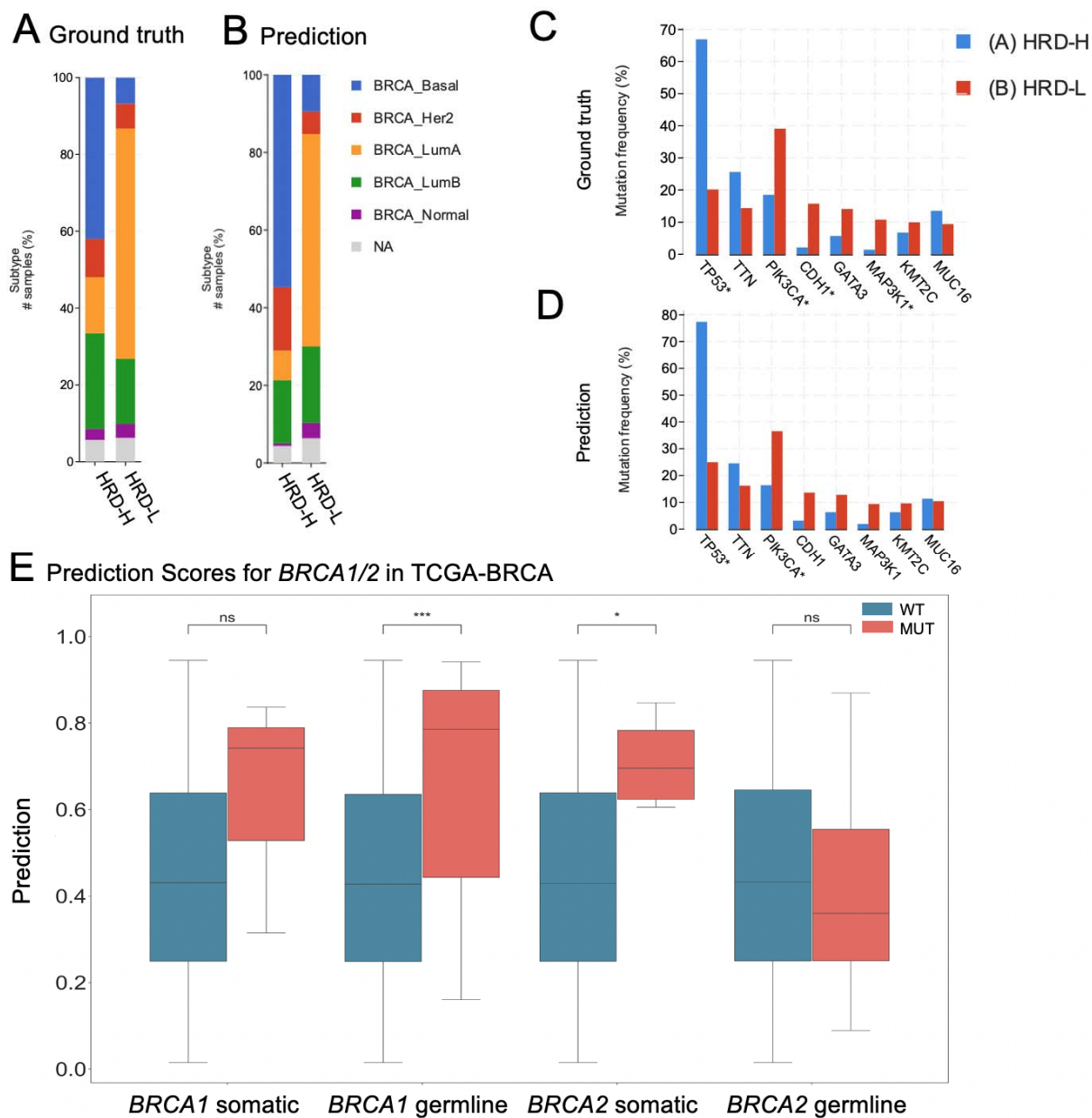


**B** Cross cancer external validation of TCGA-BRCA on all



479

480 **Figure 2: Comparison of Area under the receiving operating curve (AUROC) for internal and**  
481 **tumor wise external validation experiment models.** Boxplot displaying the distribution for the  
482 AUROC for (A) internal 5-fold cross-validation experiment of The Cancer Genome Atlas (TCGA) and  
483 tumor-wise external validation on the Clinical Proteomic Tumor Analysis Consortium (CPTAC); (B)  
484 AUROCs for the cross-cancer external validation experiment of the TCGA breast cancer cohort (TCGA-  
485 BRCA) on the TCGA and CPTAC cohort. The horizontal line indicates the median, whereas each box  
486 represents the interquartile range (IQR) between the first and third quartiles. The whiskers extend from  
487 the box to the minimum and maximum values, considering 1.5 times the IQR. Abbreviations:  
488 BRCA=breast cancer; CRC=colorectal cancer; GBM=glioblastoma; LIHC=liver cancer; LUAD=lung  
489 adenocarcinoma; LUSC/LSCC=lung squamous cell carcinoma; OV=ovarian cancer;  
490 PAAD/PDA=pancreatic adenocarcinoma; PRAD=prostate adenocarcinoma; UCEC=endometrial cancer  
491



492

493

494 **Figure 3: Molecular Characterization of The Cancer Genome Atlas breast cancer (TCGA-BRCA)**

495 **cohort.** (A) Distribution of breast cancer subtypes for the Homologous Recombination deficiency high

496 (HRD-H) and low (HRD-L) ground truth subgroups. (B) Distribution of the breast cancer subtypes for

497 the HRD-H and HRD-L Deep Learning (DL) predicted subgroups. (C) Alteration Frequency for several

498 genes of the HRD-H and HRD-L ground truth subgroups. (D) Alteration Frequency for several genes of

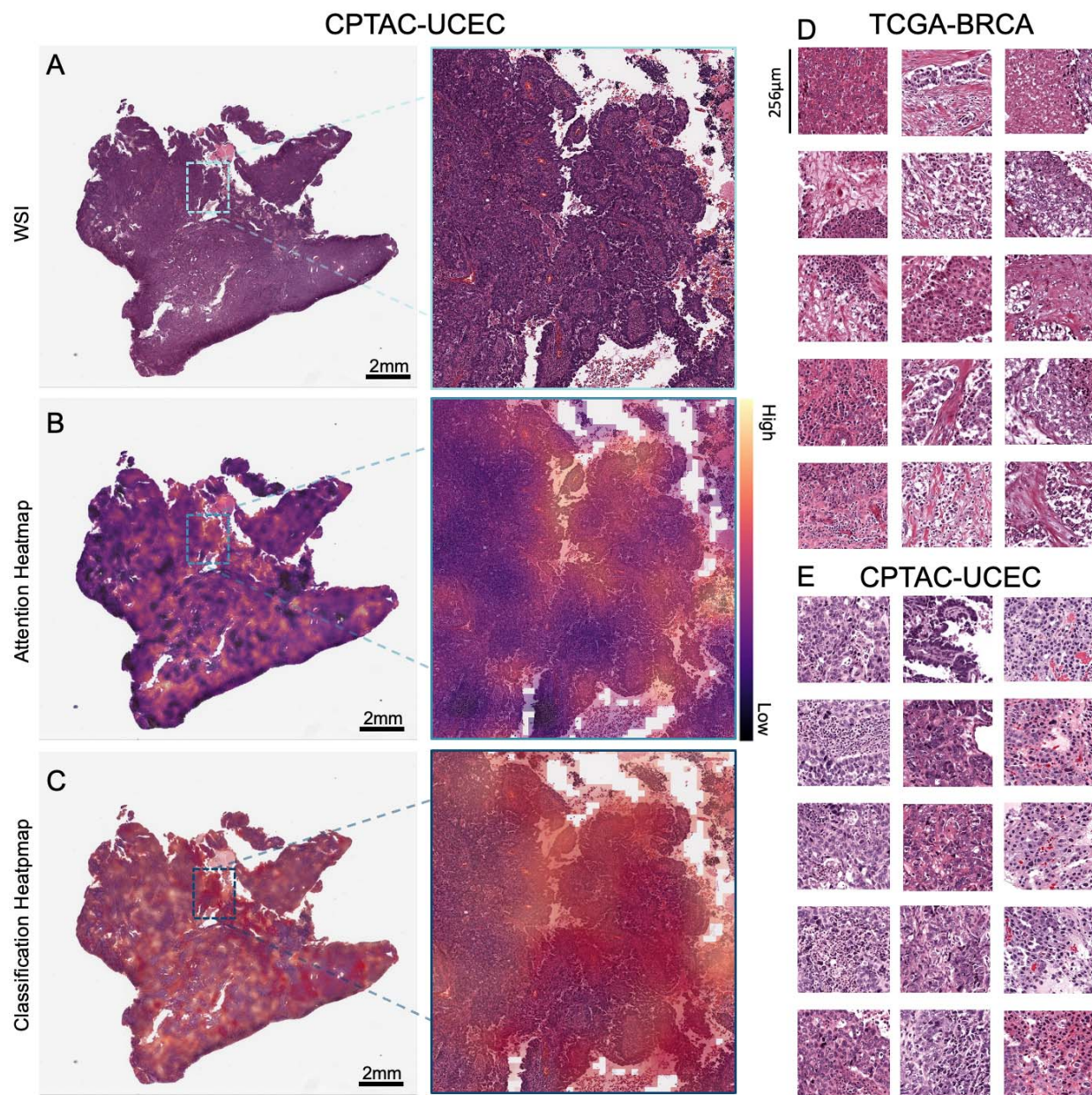
499 the HRD-H and HRD-L within cohort internal results prediction subgroups. (E) Grouped Boxplots

500 comparing the Homologous Recombination Deficiency high (HRD-H) prediction scores with the

501 mutational status (mutated=MLIT wildtype=WT) for the somatic and germline alterations of the



502 *BRCA1/2* genes. The central line represents the median value, while the box ranges between the first  
503 and third quartile (IQR) and the whiskers extend to the lowest and highest values within 1.5 times the  
504 IQR. The y-axis represents the Deep Learning (DL) HRD-H prediction values. An independent t-test  
505 was performed to calculate the p-values: ns:  $p \leq 1.00e+00$  \*:  $1.00e-02 < p \leq 5.00e-02$  \*\*:  $1.00e-03 <$   
506  $p \leq 1.00e-02$  \*\*\*:  $1.00e-04 < p \leq 1.00e-03$   
507



508

509



510 **Figure 4: Visualization of predicted Homologous Recombination Deficiency high (HRD-H) tumor**  
511 **samples.** (A) Whole slide image (WSI) of an HRD-H predicted patient (ID: C3L-00358-21) from the  
512 [Clinical Proteomic Tumor Analysis Consortium](#) (CPTAC) endometrial cancer (UCEC) cohort with  
513 magnification. (B) Attention heatmap for the same patient with magnification. (C) Classification  
514 Heatmap for the same patient with magnification. (D) Top predicted tiles for top three homologous  
515 recombination deficiency high (HRD-H) patients in The Cancer Genome Atlas (TCGA) breast cancer  
516 (BRCA). (E) Top predicted tiles for three HRD-H patients in the CPTAC-UCEC cohort.  
517

## 518 References

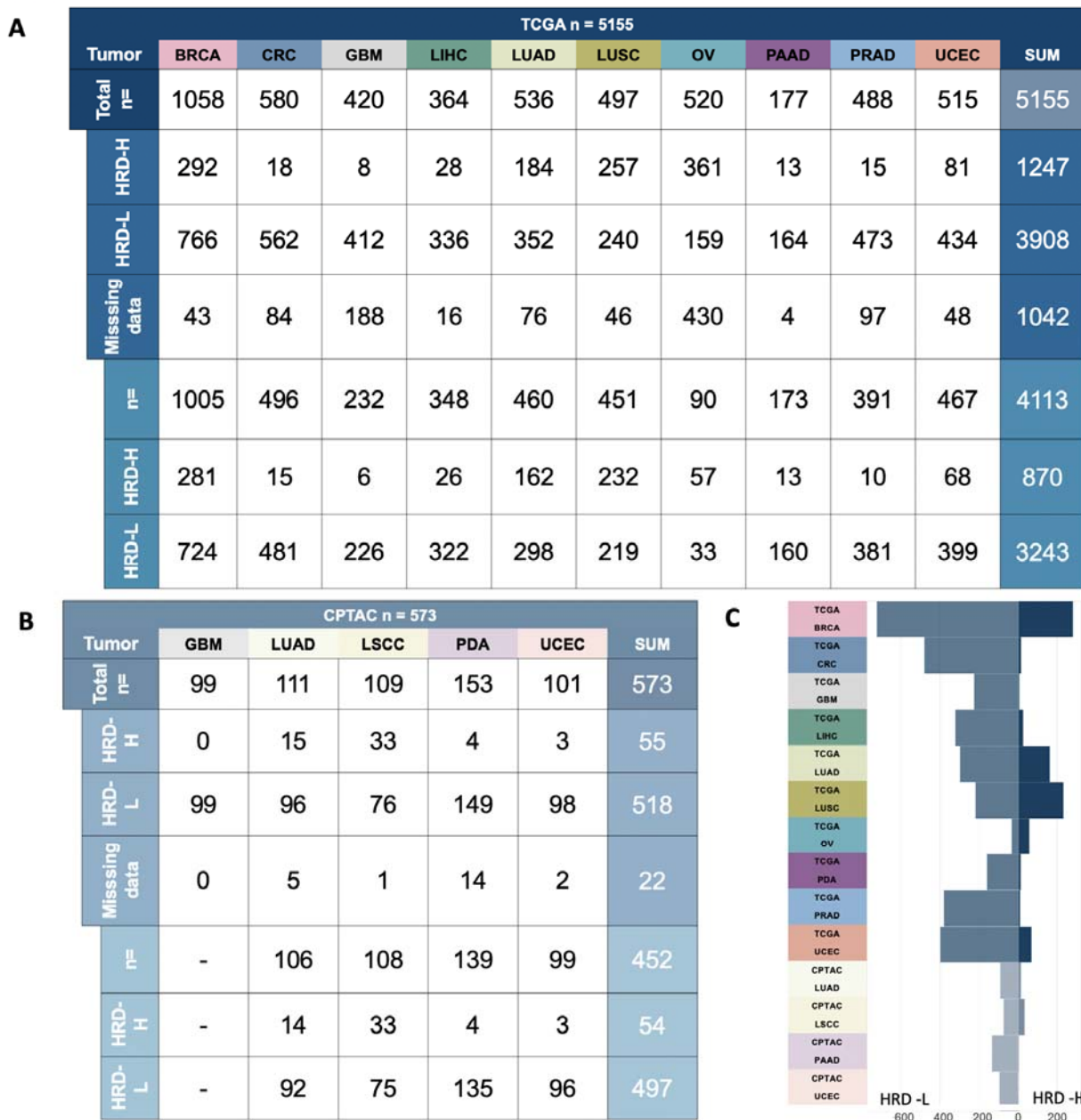
- 519 1. Frey MK, Pothuri B. Homologous recombination deficiency (HRD) testing in ovarian cancer clinical  
520 practice: a review of the literature. *Gynecol Oncol Res Pract*. 2017 Feb 22;4:4.
- 521 2. Hoeijmakers JH. Genome maintenance mechanisms for preventing cancer. *Nature*. 2001 May  
522 17;411(6835):366–74.
- 523 3. Rose M, Burgess JT, O'Byrne K, Richard DJ, Bolderson E. PARP Inhibitors: Clinical Relevance,  
524 Mechanisms of Action and Tumor Resistance. *Front Cell Dev Biol*. 2020 Sep 9;8:564601.
- 525 4. Dedes KJ, Wilkerson PM, Wetterskog D, Weigelt B, Ashworth A, Reis-Filho JS. Synthetic lethality  
526 of PARP inhibition in cancers lacking BRCA1 and BRCA2 mutations. *Cell Cycle*. 2011 Apr  
527 15;10(8):1192–9.
- 528 5. Leary A, Auguste A, Mesnage S. DNA damage response as a therapeutic target in gynecological  
529 cancers. *Curr Opin Oncol*. 2016 Sep;28(5):404–11.
- 530 6. Park W, Chen J, Chou JF, Varghese AM, Yu KH, Wong W, et al. Genomic Methods Identify  
531 Homologous Recombination Deficiency in Pancreas Adenocarcinoma and Optimize Treatment  
532 Selection. *Clin Cancer Res*. 2020 Jul 1;26(13):3239–47.
- 533 7. Takaya H, Nakai H, Takamatsu S, Mandai M, Matsumura N. Homologous recombination deficiency  
534 status-based classification of high-grade serous ovarian carcinoma. *Sci Rep*. 2020 Feb  
535 17;10(1):2757.
- 536 8. Tutt ANJ, Garber JE, Kaufman B, Viale G, Fumagalli D, Rastogi P, et al. Adjuvant Olaparib for  
537 Patients with BRCA1- or BRCA2-Mutated Breast Cancer. *N Engl J Med*. 2021 Jun  
538 24;384(25):2394–405.
- 539 9. Ledermann JA. PARP inhibitors in ovarian cancer. *Ann Oncol*. 2016 Apr;27 Suppl 1:i40–4.
- 540 10. Stewart MD, Merino Vega D, Arend RC, Baden JF, Barbash O, Beaubier N, et al. Homologous  
541 Recombination Deficiency: Concepts, Definitions, and Assays. *Oncologist*. 2022 Mar 11;27(3):167–  
542 74.
- 543 11. Miller RE, Leary A, Scott CL, Serra V, Lord CJ, Bowtell D, et al. ESMO recommendations on  
544 predictive biomarker testing for homologous recombination deficiency and PARP inhibitor benefit in  
545 ovarian cancer. *Ann Oncol*. 2020 Dec;31(12):1606–22.
- 546 12. Wagener-Rydzek S, Merkelbach-Bruse S, Siemanowski J. Biomarkers for Homologous  
547 Recombination Deficiency in Cancer. *J Pers Med [Internet]*. 2021 Jun 28;11(7). Available from:  
548 <http://dx.doi.org/10.3390/jpm11070612>
- 549 13. Fuh K, Mullen M, Blachut B, Stover E, Konstantinopoulos P, Liu J, et al. Homologous  
550 recombination deficiency real-time clinical assays, ready or not? *Gynecol Oncol*. 2020  
551 Dec;159(3):877–86.
- 552 14. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of  
553 mutational signatures in human cancer. *Nature*. 2020 Feb;578(7793):94–101.
- 554 15. Gulhan DC, Lee JJK, Melloni GEM, Cortés-Ciriano I, Park PJ. Detecting the mutational signature of  
555 homologous recombination deficiency in clinical samples. *Nat Genet*. 2019 May;51(5):912–9.
- 556 16. Abkevich V, Timms KM, Hennessy BT, Potter J, Carey MS, Meyer LA, et al. Patterns of genomic  
557 loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian  
558 cancer. *Br J Cancer*. 2012 Nov 6;107(10):1776–82.

- 559 17. Birkbak NJ, Wang ZC, Kim JY, Eklund AC, Li Q, Tian R, et al. Telomeric allelic imbalance indicates  
560 defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.* 2012 Apr;2(4):366–  
561 75.
- 562 18. Popova T, Manié E, Rieunier G, Caux-Moncoutier V, Tirapo C, Dubois T, et al. Ploidy and large-  
563 scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2  
564 inactivation. *Cancer Res.* 2012 Nov 1;72(21):5454–62.
- 565 19. Westphalen CB, Fine AD, André F, Ganesan S, Heinemann V, Rouleau E, et al. Pan-cancer  
566 Analysis of Homologous Recombination Repair-associated Gene Alterations and Genome-wide  
567 Loss-of-Heterozygosity Score. *Clin Cancer Res.* 2022 Apr 1;28(7):1412–21.
- 568 20. Sztupinszki Z, Diossy M, Krzystanek M, Reiniger L, Csabai I, Favero F, et al. Migrating the SNP  
569 array-based homologous recombination deficiency measures to next generation sequencing data  
570 of breast cancer. *NPJ Breast Cancer.* 2018 Jul 2;4:16.
- 571 21. Zhao EY, Shen Y, Pleasance E, Kasaian K, Leelakumari S, Jones M, et al. Homologous  
572 Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer.  
573 *Clin Cancer Res.* 2017 Dec 15;23(24):7521–30.
- 574 22. Nguyen L, W M Martens J, Van Hoeck A, Cuppen E. Pan-cancer landscape of homologous  
575 recombination deficiency. *Nat Commun.* 2020 Nov 4;11(1):5584.
- 576 23. Pellegrino B, Herencia-Ropero A, Llop-Guevara A, Pedretti F, Moles-Fernández A, Viaplana C, et  
577 al. Preclinical In Vivo Validation of the RAD51 Test for Identification of Homologous  
578 Recombination-Deficient Tumors and Patient Stratification. *Cancer Res.* 2022 Apr 15;82(8):1646–  
579 57.
- 580 24. Graeser M, McCarthy A, Lord CJ, Savage K, Hills M, Salter J, et al. A marker of homologous  
581 recombination predicts pathologic complete response to neoadjuvant chemotherapy in primary  
582 breast cancer. *Clin Cancer Res.* 2010 Dec 15;16(24):6159–68.
- 583 25. How JA, Jazaeri AA, Fellman B, Daniels MS, Penn S, Solimeno C, et al. Modification of  
584 Homologous Recombination Deficiency Score Threshold and Association with Long-Term Survival  
585 in Epithelial Ovarian Cancer. *Cancers [Internet].* 2021 Feb 24;13(5). Available from:  
586 <http://dx.doi.org/10.3390/cancers13050946>
- 587 26. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning  
588 model to predict RNA-Seq expression of tumours from whole slide images. *Nat Commun.* 2020  
589 Aug 3;11(1):3877.
- 590 27. Loeffler CML, Ortiz Bruechle N, Jung M, Seillier L, Rose M, Laleh NG, et al. Artificial Intelligence-  
591 based Detection of FGFR3 Mutational Status Directly from Routine Histology in Bladder Cancer: A  
592 Possible Preselection for Molecular Testing? *Eur Urol Focus.* 2022 Mar;8(2):472–9.
- 593 28. Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology:  
594 enhancing cancer research and clinical oncology. *Nat Cancer.* 2022 Sep;3(9):1026–38.
- 595 29. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational  
596 histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer.* 2020 Jul  
597 27;1–11.
- 598 30. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-cancer image-based  
599 detection of clinically actionable genetic alterations. *Nature Cancer.* 2020 Aug 1;1(8):789–99.
- 600 31. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict  
601 microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med.* 2019  
602 Jul;25(7):1054–6.

- 603 32. Muti HS, Heij LR, Keller G, Kohlruss M, Langer R, Dislich B, et al. Development and validation of  
604 deep learning classifiers to detect Epstein-Barr virus and microsatellite instability status in gastric  
605 cancer: a retrospective multicentre cohort study [Internet]. *The Lancet Digital Health*. 2021.  
606 Available from: [http://dx.doi.org/10.1016/s2589-7500\(21\)00133-3](http://dx.doi.org/10.1016/s2589-7500(21)00133-3)
- 607 33. Saillard C, Dubois R, Tchita O, Loiseau N, Garcia T, Adriansen A, et al. Blind validation of MSIntuit,  
608 an AI-based pre-screening tool for MSI detection from histology slides of colorectal cancer  
609 [Internet]. *bioRxiv*. 2022. Available from:  
610 <https://www.medrxiv.org/content/10.1101/2022.11.17.22282460.abstract>
- 611 34. Kleppe A, Skrede OJ, De Raedt S, Hveem TS, Askautrud HA, Jacobsen JE, et al. A clinical  
612 decision support system optimising adjuvant chemotherapy for colorectal cancers by integrating  
613 deep learning and pathological staging markers: a development and validation study. *Lancet*  
614 *Oncol*. 2022 Sep;23(9):1221–32.
- 615 35. Valieris R, Amaro L, Osório CAB de T, Bueno AP, Rosales Mitrowsky RA, Carraro DM, et al. Deep  
616 Learning Predicts Underlying Features on Pathology Images with Therapeutic Relevance for  
617 Breast and Gastric Cancer. *Cancers* [Internet]. 2020 Dec 9;12(12). Available from:  
618 <http://dx.doi.org/10.3390/cancers12123687>
- 619 36. Lazard T, Bataillon G, Naylor P, Popova T, Bidard FC, Stoppa-Lyonnet D, et al. Deep learning  
620 identifies morphological patterns of homologous recombination deficiency in luminal breast cancers  
621 from whole slide images. *Cell Rep Med*. 2022 Dec 20;3(12):100872.
- 622 37. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of  
623 complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. 2013 Apr  
624 2;6(269):11.
- 625 38. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics  
626 portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*.  
627 2012 May;2(5):401–4.
- 628 39. Riaz N, Bleuca P, Lim RS, Shen R, Higginson DS, Weinhold N, et al. Pan-cancer analysis of bi-  
629 allelic alterations in homologous recombination DNA repair genes. *Nat Commun*. 2017 Oct  
630 11;8(1):857.
- 631 40. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for  
632 normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on  
633 Biomedical Imaging: From Nano to Macro. 2009. p. 1107–10.
- 634 41. Wang X, Du Y, Yang S, Zhang J, Wang M, Zhang J, et al. RetCCL: Clustering-guided contrastive  
635 learning for whole-slide image retrieval. *Med Image Anal*. 2023 Jan 1;83:102645.
- 636 42. Leiby JS, Hao J, Kang GH, Park JW, Kim D. Attention-based multiple instance learning with self-  
637 supervision to predict microsatellite instability in colorectal cancer from histology whole-slide  
638 images. *Conf Proc IEEE Eng Med Biol Soc*. 2022 Jul;2022:3068–71.
- 639 43. Ilse M, Tomczak J, Welling M. Attention-based Deep Multiple Instance Learning. In: Dy J, Krause  
640 A, editors. *Proceedings of the 35th International Conference on Machine Learning*. PMLR; 10–15  
641 Jul 2018. p. 2127–36. (*Proceedings of Machine Learning Research*; vol. 80).
- 642 44. Rempel E, Kluck K, Beck S, Ourailidis I, Kazdal D, Neumann O, et al. Pan-cancer analysis of  
643 genomic scar patterns caused by homologous repair deficiency (HRD). *NPJ Precis Oncol*. 2022  
644 Jun 9;6(1):36.
- 645 45. Kleppe A, Skrede OJ, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning  
646 studies in cancer diagnostics. *Nat Rev Cancer*. 2021 Mar;21(3):199–211.
- 647 46. Ng CKY, Piscuoglio S, Geyer FC, Burke KA, Pareja F, Eberle CA, et al. The Landscape of Somatic

- 648 Genetic Alterations in Metaplastic Breast Carcinomas. *Clin Cancer Res.* 2017 Jul 15;23(14):3859–  
649 70.
- 650 47. Ngoi NYL, Tan DSP. The role of homologous recombination deficiency testing in ovarian cancer  
651 and its clinical implications: do we need it? *ESMO Open.* 2021 Jun;6(3):100144.
- 652 48. Loeffler CML, Gaisa NT, Muti HS, van Treeck M. Predicting Mutational Status of Driver and  
653 Suppressor Genes Directly from Histopathology With Deep Learning: A Systematic Study Across  
654 23 Solid Tumor .... *Frontiers in [Internet].* 2021; Available from:  
655 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8889144/>
- 656 49. Takamatsu S, Brown JB, Yamaguchi K, Hamanishi J, Yamanoi K, Takaya H, et al. Utility of  
657 Homologous Recombination Deficiency Biomarkers Across Cancer Types. *JCO Precis Oncol.*  
658 2022 May;6:e2200085.
- 659 50. Moukarzel LA, Ferrando L, Da Cruz Paula A, Brown DN, Geyer FC, Pareja F, et al. The genetic  
660 landscape of metaplastic breast cancers and uterine carcinosarcomas. *Mol Oncol.* 2021  
661 Apr;15(4):1024–39.
- 662 51. Na B, Yu X, Withers T, Gilleran J, Yao M, Foo TK, et al. Therapeutic targeting of BRCA1 and TP53  
663 mutant breast cancer through mutant p53 reactivation. *NPJ Breast Cancer.* 2019 Apr 15;5:14.
- 664 52. Lai Z, Brosnan M, Sokol ES, Xie M, Dry JR, Harrington EA, et al. Landscape of homologous  
665 recombination deficiencies in solid tumours: analyses of two independent genomic datasets. *BMC*  
666 *Cancer.* 2022 Jan 3;22(1):13.

667 **Supplementary Figures and Tables**



668

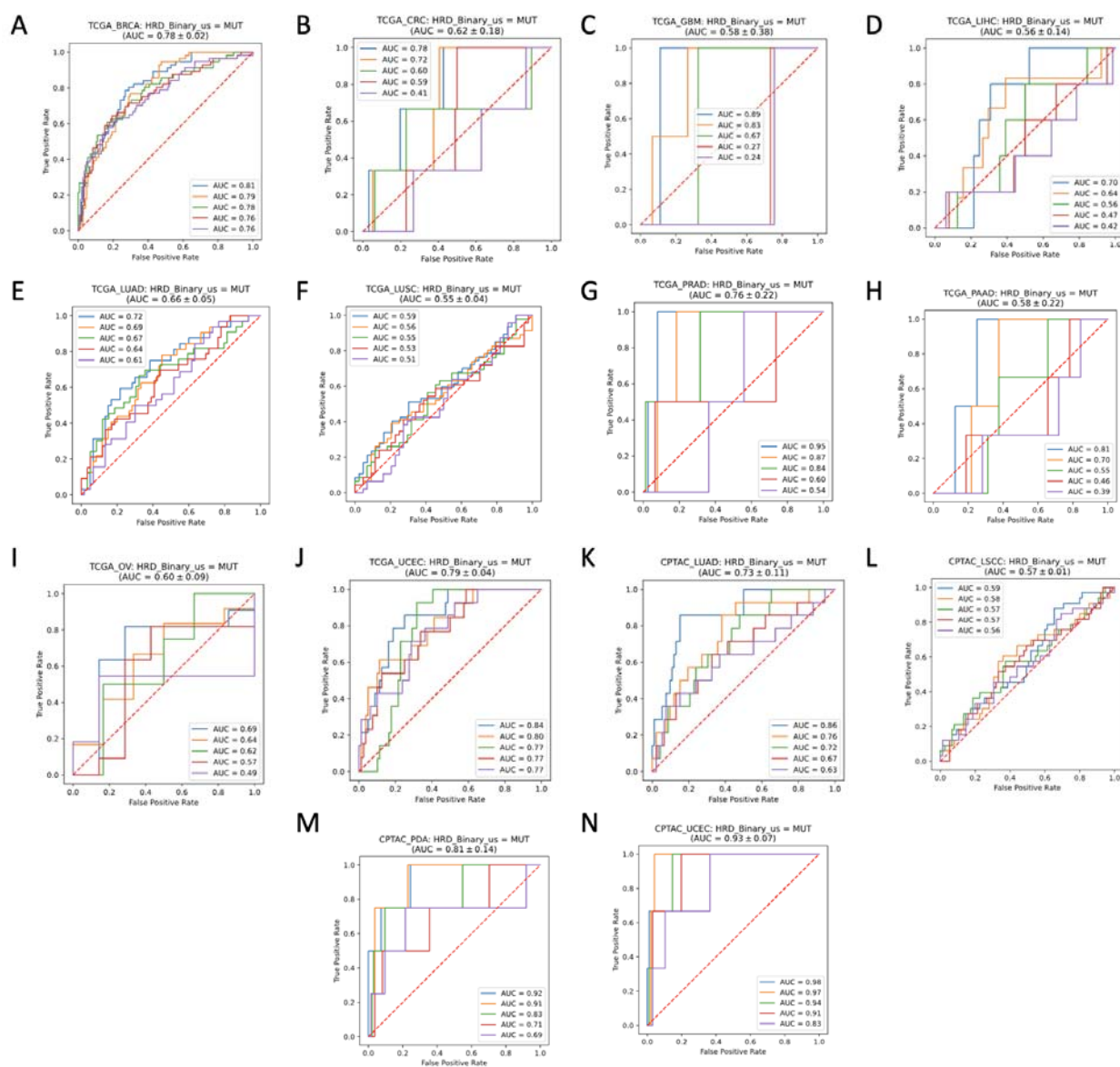
669

670 **Supplementary Figure 1: Homologous recombination deficiency prevalences across the**  
 671 **cohorts.** (A) Overview of the total patient count (n=573) in the CPTAC cohort before merging the image  
 672 data with the molecular data and afterward. (B) Overview of the total patient count (n=5,155) in the  
 673 TCGA cohort before merging the image data with the molecular data and afterward. (C) Distribution of  
 674 the homologous recombination deficiency high (HRD-H) and low (HRD-L) patient number among the  
 675 different tumor types of The Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumor Analysis



676 Consortium (CPTAC). Abbreviations: BRCA=breast invasive carcinoma; CRC=colorectal cancer;  
 677 GBM=glioblastoma; LIHC=liver cancer; LUAD=lung adenocarcinoma; LUSC/LSCC=lung squamous cell  
 678 carcinoma; OV=ovarian cancer; PAAD/PDA=pancreatic adenocarcinoma; PRAD=prostate  
 679 adenocarcinoma; UCEC=endometrial cancer  
 680

Internal 5 fold Cross Validation and tumor wise external validation ROCs

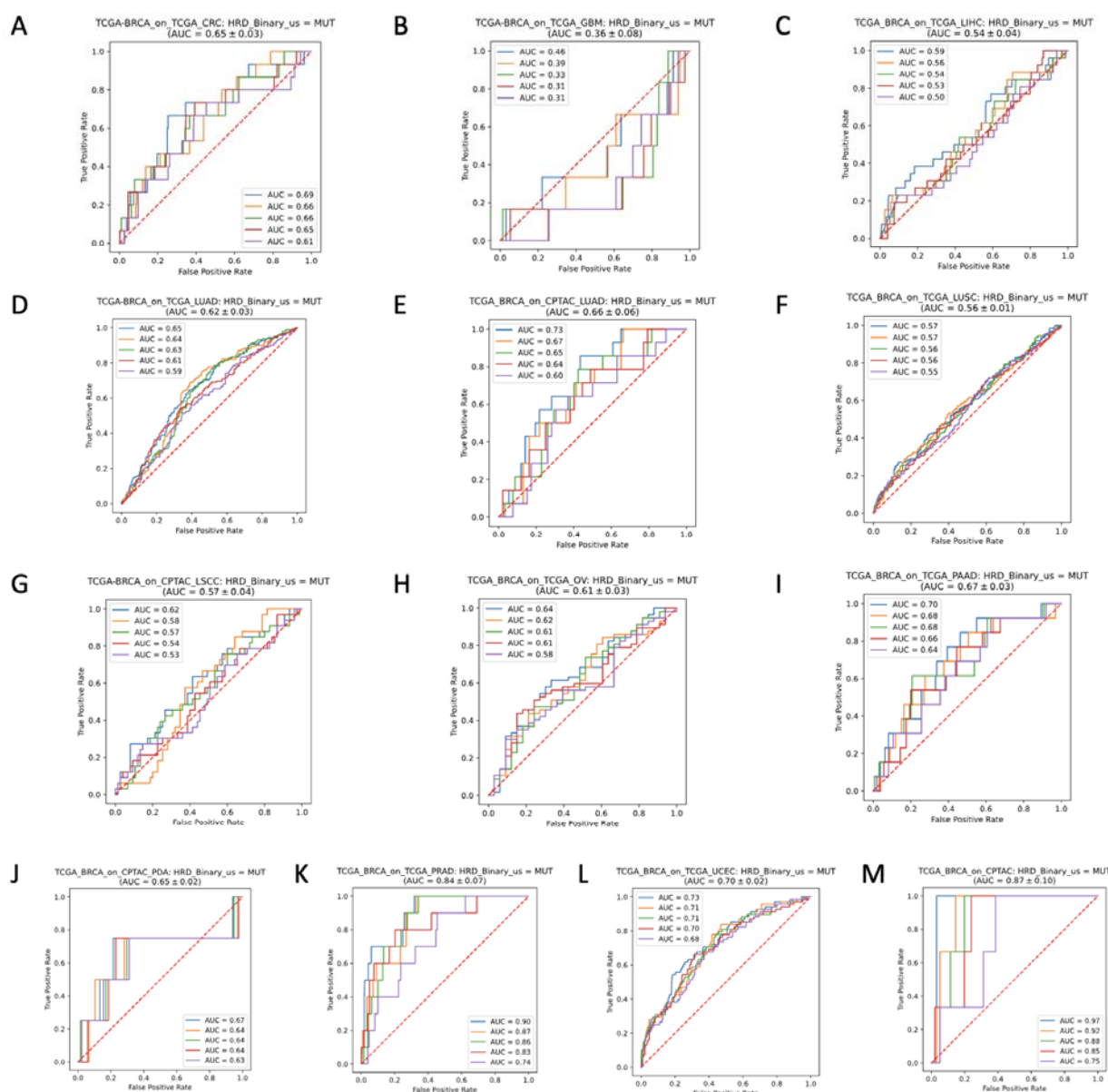


681  
 682 **Supplementary Figure 2: Receiving operating curve for the Internal Validation and tumor wise**  
 683 **external validation.** The Receiving operating curve (ROC) is shown for the five-fold internal cross-  
 684 validation experiment for each of the models in The Cancer Genome Atlas (TCGA) for the Homologous  
 685 recombination deficiency (HRD) binary score for (A) TCGA-BRCA, (B) TCGA-CRC, (C) TCGA-GBM,



686 (D) TCGA-LIHC, (E) TCGA-LUAD, (F) TCGA-LUSC, (G) TCGA-PAAD, (H) TCGA-PRAD, (I) TCGA-OV,  
 687 (J) TCGA-UCEC; Roc curves for the external validation on the Clinical Proteomic Tumor Analysis  
 688 Consortium (CPTAC) for each previously trained model for (K) CPTAC-LUAD, (L) CPTAC-LSCC, (M)  
 689 CPTAC-PDA, (N) CPTAC-UCEC. Abbreviations: BRCA=breast invasive carcinoma; CRC=colorectal  
 690 cancer; GBM=glioblastoma; LIHC=liver cancer; LUAD=lung adenocarcinoma; LUSC/LSCC=lung  
 691 squamous cell carcinoma; OV=ovarian cancer; PAAD/PDA=pancreatic adenocarcinoma;  
 692 PRAD=prostate adenocarcinoma; UCEC=endometrial cancer  
 693

### Cross cancer external validation ROCs



694

695

696 **Supplementary Figure 3: Receiving operating curve for the cross-cancer external validation.** The

697 Receiving operating curve (ROC) is shown for the cross-cancer external validation experiment for each

698 model trained on The Cancer Genome Atlas (TCGA) breast cancer (BRCA) cohort for the Homologous

699 recombination deficiency (HRD) binary score on (A) TCGA-CRC, (B) TCGA-GBM, (C) TCGA-LIHC, (D)

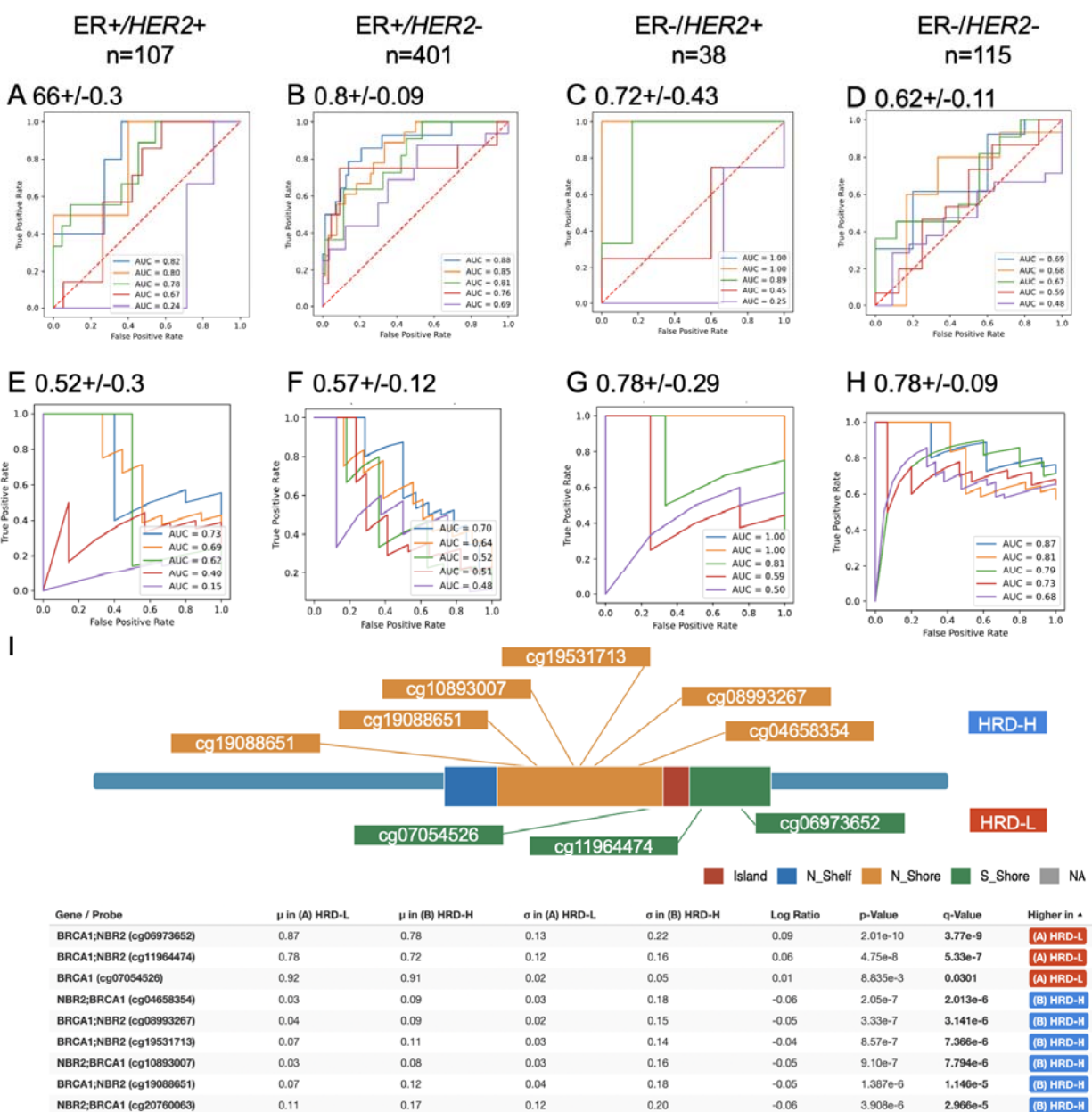
700 TCGA-LUAD, (E) CPTAC-LUAD, (F) TCGA-LUSC, (G) CPTAC-LSCC, (H) TCGA-OV, (I) TCGA-PAAD,

701 (J) CPTAC-PDA, (K) TCGA-PRAD, (L) TCGA-UCEC, (M) CPTAC-UCEC. Abbreviations: BRCA=breast

702 invasive carcinoma; CRC=colorectal cancer; GBM=glioblastoma; LIHC=liver cancer; LUAD=lung

703 adenocarcinoma; LUSC/LSCC=lung squamous cell carcinoma; OV=ovarian cancer;

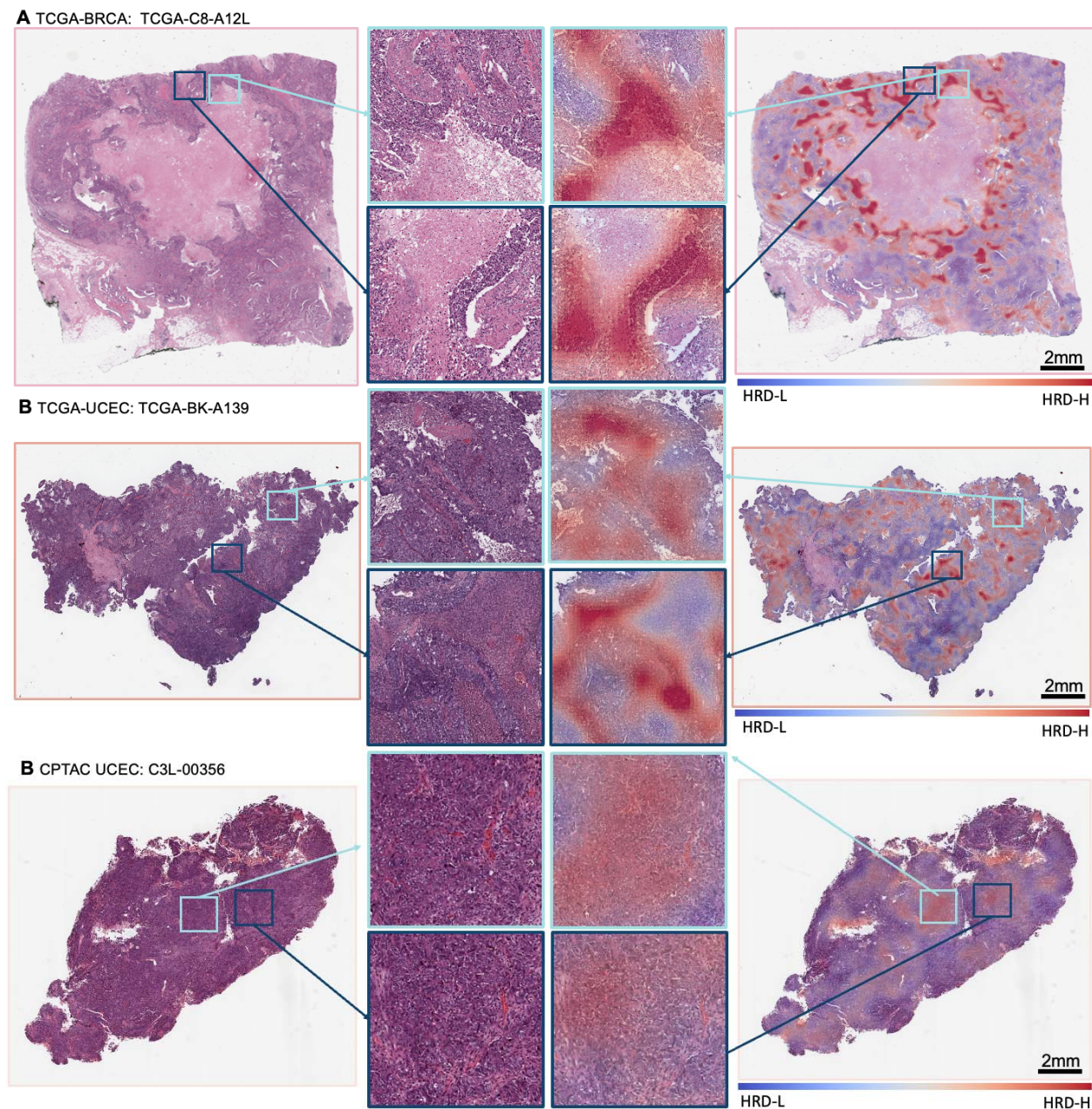
704 PAAD/PDA=pancreatic adenocarcinoma; PRAD=prostate adenocarcinoma; UCEC=endometrial cancer



705  
706 **Supplementary Figure 4: Subgroup analysis and overview the *BRCA1* promotor methylations in**  
707 **TCGA-BRCA.** The Receiving operating curve (ROC) and Precision Recall curve (PRC) are shown for  
708 the five-fold internal cross-validation experiment for each of the models in The Cancer Genome Atlas -  
709 breast cancer (TCGA-BRCA) cohort for the Homologous recombination deficiency (HRD) score. ROC  
710 curve is represented for the four different subgroups (A) estrogen receptor positive (ER+) and *HER2+*  
711 (B) ER+ and *HER2-* (C) ER negative (ER-) and *HER2+* (D) ER- and *HER2-*. The PRC curve is shown  
712 for (E) ER+/HER2+, (F) ER+/HER2-, (G) ER-/HER2+, (H) ER-/HER2-. (I) Sketched representation of



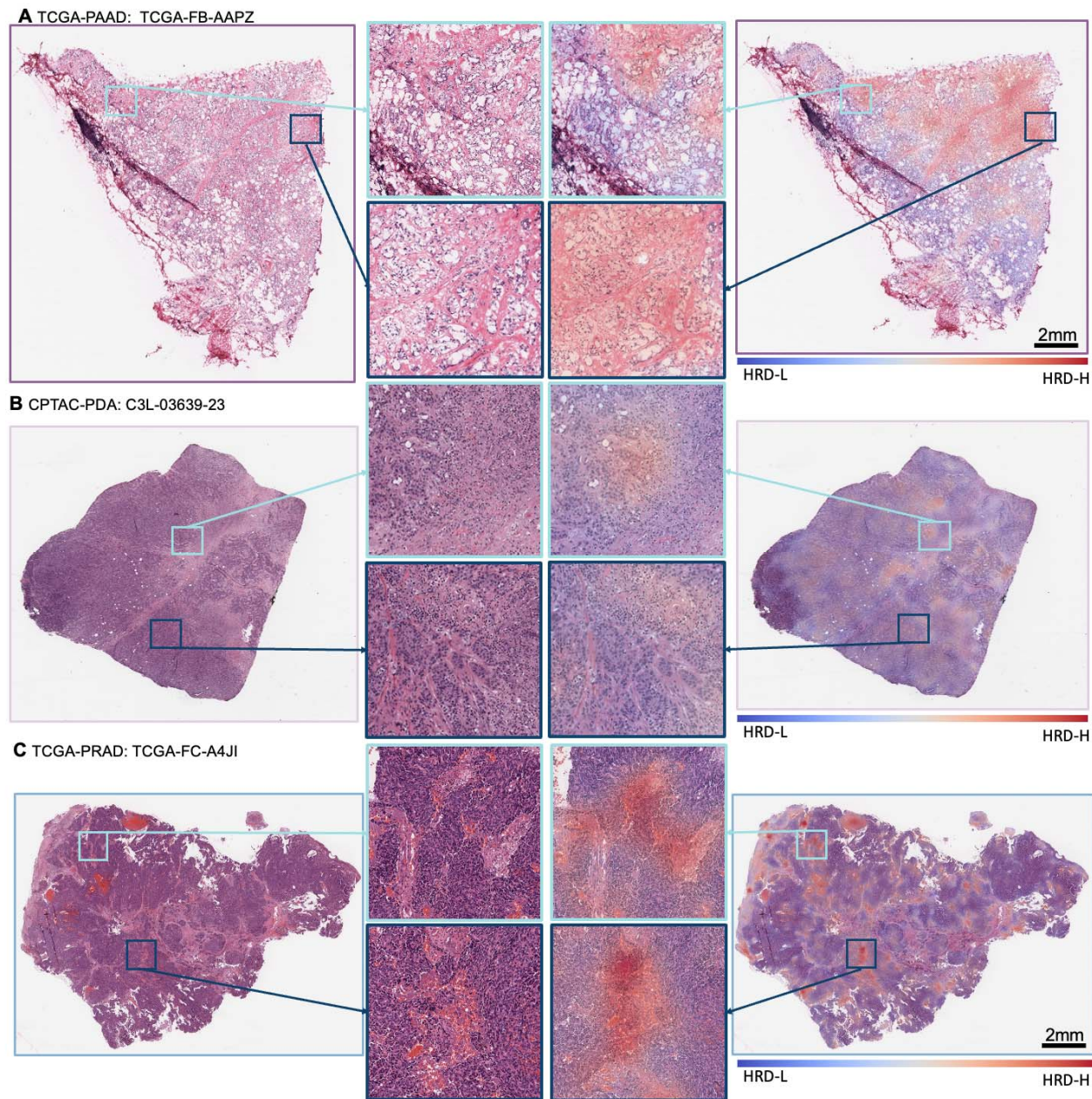
713 the occurring promotor methylations (accessed with HM27 and HM450) in the *BRCA1* gene for the  
714 ground truth Homologous recombination deficiency high (HRD-H) and low (HRD-L) subgroups.



715  
716 **Supplementary Figure 5: Morphological features of Homologous recombination deficiency in**  
717 **breast and endometrial cancer.** Whole Slide Image (WSI) and classification heatmap (ground truth:  
718 Homologous recombination deficiency high (HRD-H) and prediction: HRD-H) with magnifications of two  
719 different regions. The model was trained on The cancer genome atlas (TCGA) breast cancer (BRCA)  
720 cohort and deployed cross cancer wise. Top true positive predicted patients are shown for (A) TCGA-



721 BRCA, (B) Clinical Proteomic Tumor Analysis Consortium (CPTAC) endometrial cancer (UCEC) and  
722 (C) TCGA-UCEC.



723  
724 **Supplementary Figure 6: Morphological features of Homologous recombination deficiency in**  
725 **pancreatic and prostate adenocarcinoma.** Whole Slide Image (WSI) and classification heatmap  
726 (ground truth: Homologous recombination deficiency high (HRD-H) and prediction: HRD-H) with  
727 magnifications of two different regions. The model was trained on The cancer genome atlas (TCGA)  
728 breast cancer (BRCA) cohort and deployed cross cancer wise. Top true positive predicted patients are  
729 shown for (A) TCGA pancreatic adenocarcinoma (PAAD), (B) Clinical Proteomic Tumor Analysis

730 Consortium (CPTAC) pancreatic adenocarcinoma (PDA) and (C) TCGA prostate adenocarcinoma  
731 (PRAD).

732  
733 **Supplementary Table 1: All raw statistical results.** All raw experimental results related to Figure 2,  
734 including receiving operating curve (ROC) with 95% confidence interval (CI), Precision-Recall Curve  
735 (PRC) with 95% confidence interval (CI), p-values and Homologous recombination deficiency (HRD)  
736 high (HRD-H) and HRD-low (HRD-L) patient numbers based on the ground truth, for internal 5-fold  
737 cross-validation on The Cancer Genome Atlas (TCGA) external validation on Clinical Proteomic Tumor  
738 Analysis Consortium (CPTAC). [Supplementary\_Table\_1\_All\_statistical\_results.xlsx] in separate file

739  
740 **Supplementary Table 2: Homologous recombination deficiency score Tables.** Training data and  
741 calculated homologous recombination deficiency score (HRD) out of the three subscores loss of  
742 heterozygosity (LOH), telomeric allelic imbalance (TAI) and large-scale state transitions (LST) available  
743 as continuous (HRDsum) and binary (HRD\_Binary) target with a chosen cut off of  $HRD-L < 42$   $HRD-H \geq 42$   
744 for patients of The Cancer Genome Atlas (TCGA, Sheet1) and Clinical Proteomic Tumor  
745 Analysis Consortium (CPTAC, Sheet2).

746  
747 **Supplementary Table 3: Weblink for customized Homologous recombination deficiency (HRD)**  
748 **subgroups.** Weblink for accessing the clinical and molecular characteristics for both ground truth and  
749 prediction Homologous recombination Deficiency (HRD) subgroups at [www.cbioportal.org](http://www.cbioportal.org) for The  
750 Cancer Genome Atlas breast cancer (TCGA-BRCA) Pan Cancer Atlas 2018 study and the TCGA-  
751 BRCA Firehose Legacy cohort.