Supplementary Materials for: Predicting Homologous Recombination Deficiency and Treatment Responses using a Computed Tomography-based Foundation Model: A Preclinical Study

Sheng Kuang[a*], Lesley Schuitmaker[a], Min Wu[b], Zohaib Salahuddin[a], Alexander van der Wiel[a], Jella van de Laak[a], Natasja Lieuwes[a], Rianne Biemans[a], Jennifer Jung[a], Ala Yaromina[a], Ludwig J. Dubois[a], Henry C. Woodruff[a], Philippe Lambin[a, c*]

a. Department of Precision Medicine, GROW – Research Institute for Oncology and Reproduction, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, Netherlands
b. Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom
c. Department of Radiology and Nuclear Medicine, GROW - Research Institute for Oncology and Reproduction, Maastricht University Medical Center+, Maastricht, Netherlands

*Corresponding author: Sheng Kuang (sheng.kuang@maastrichtuniversity.nl)

Supplementary Results

Effect of using pretraining weights of MedicalNet on sDL performance

To assess whether relatively large datasets pretraining affects performance, we also fine-tuned a ResNet-50 initialized with MedicalNet weights using identical data splits and training settings. MedicalNet is a pretrained ResNet-50 on around 1.6 K CT and MRI scans. Specifically, we initialized ResNet-50 with MedicalNet weights and fine-tuned it using the same data splits and training settings as the from-scratch model. The best learning rate was determined under the grid search.

On the training set, the MedicalNet model achieved AUCs ranging from 0.70 to 0.76, while our sDL baseline ranged from 0.71 to 0.78, with largely overlapping confidence intervals. On the test set, MedicalNet achieved AUCs of 0.80-0.82, and the sDL baseline achieved AUCs of 0.77-0.85. Overall, the pretrained MedicalNet did not provide a consistent improvement over the sDL baseline, and both remained clearly below the FM. We attribute these results to two factors. First, the MedicalNet pretraining dataset is smaller and less diverse than that of the FM, limiting the incremental benefit. Second, the sDL baseline was trained with a pipeline deliberately tailored to maximize data utilization and improve generalizability, resulting in competitive performance.

Supplementary Tables

Table S1. Detailed information of isogenic xenograft models.

Cell line	Cancer Type	DNA repair pathway	Provider	Mouse strain	Sex	Age
LNCaP AR parental	Prostate	Parental	MSKCC	NOD.Cg-Prkdc^{SCID} Il2rg^{tm1Wjl}/$_{SzJ}$	male	8-10 weeks
LNCaP AR FANCA$^{-/-}$	Prostate	FA	MSKCC			
LNCaP AR FANCD2$^{-/-}$	Prostate	FA	MSKCC			
HCT116 parental	Colorectal	Parental	Horizon Discovery	NU-Foxn1 nu/nu	female	8-10 weeks
HCT116 BRCA2$^{-/-}$	Colorectal	HR	Ximbio			
HCT116 DNA-PKcs$^{-/-}$	Colorectal	NHEJ	Horizon Discovery			
DLD-1 parental	Colorectal	Parental	Horizon Discovery	BALB/c nu/nu	male	8-10 weeks
DLD-1 BRCA2$^{-/-}$	Colorectal	HR	Horizon Discovery			

FA: Fanconi anemia pathway; HR: homologous recombination; NHEJ: non-homologous end joining; FANCA: FA complementation group A; FANCD2: FA complementation group D2; BRCA2: breast cancer type 2 susceptibility protein; DNA-PKcs: DNA-dependent protein kinase

Table S2. Calibration performance for the three HRD classification models (HCR, sDL, FM) across 40 kVp, 80 kVp and combined CT. Brier score and expected calibration error are shown pre- and post-Platt calibration with 95% CIs.

Model	Modality	Brier		ECE	
		Pre	Post	Pre	Post
HCR	40 kVp	0.236 [0.226, 0.247]	0.232 [0.223, 0.240]	0.070 [0.055, 0.101]	0.054 [0.037, 0.083]
	80 kVp	0.251 [0.239, 0.262]	0.243 [0.234, 0.253]	0.113 [0.093, 0.143]	0.086 [0.065, 0.114]
	Combined	0.249 [0.237, 0.260]	0.241 [0.232, 0.250]	0.102 [0.081, 0.130]	0.067 [0.052, 0.098]
sDL	40 kVp	0.238 [0.225, 0.251]	0.227 [0.217, 0.237]	0.109 [0.086, 0.136]	0.055 [0.038, 0.085]
	80 kVp	0.225 [0.215, 0.235]	0.224 [0.217, 0.231]	0.031 [0.028, 0.065]	0.035 [0.022, 0.065]
	Combined	0.223 [0.211, 0.235]	0.217 [0.209, 0.225]	0.066 [0.053, 0.095]	0.027 [0.019, 0.059]
FM	40 kVp	0.245 [0.229, 0.262]	0.237 [0.222, 0.252]	0.167 [0.145, 0.194]	0.136 [0.113, 0.162]
	80 kVp	0.210 [0.195, 0.224]	0.208 [0.195, 0.222]	0.108 [0.089, 0.134]	0.090 [0.070, 0.116]
	Combined	0.215 [0.199, 0.229]	0.212 [0.197, 0.225]	0.123 [0.105, 0.149]	0.100 [0.081, 0.127]

Table S3. Comparison of prediction performance of HRD between sDL-MedicalNet and sDL-Scratch on training set.

Model	CT Type	AUC	Accuracy	Sensitivity	Specificity
sDL-MedicalNet	40 kVp	0.76 [0.71, 0.80]	73% [71%, 75%]	75% [60%, 90%]	71% [55%, 86%]
	80 kVp	0.70 [0.67, 0.73]	68% [65%, 72%]	63% [45%, 82%]	72% [57%, 87%]
	Combined	0.74 [0.69, 0.79]	69% [64%, 75%]	67% [44%, 89%]	74% [56%, 92%]
sDL-Scratch	40 kVp	0.78 [0.71, 0.85]	73% [69%, 77%]	72% [55%, 90%]	74% [60%, 88%]
	80 kVp	0.71 [0.65, 0.77]	68% [64%, 72%]	71% [54%, 89%]	64% [45%, 84%]
	Combined	0.74 [0.62, 0.86]	71% [62%, 79%]	76% [49%, 100%]	66% [36%, 95%]

Table S4. Comparison of prediction performance of HRD between sDL-MedicalNet and sDL-Scratch on test set.

Model	CT Type	AUC	Accuracy	Sensitivity	Specificity
sDL-MedicalNet	40 kVp	0.82 [0.72, 0.87]	72% [62%, 78%]	73% [62%, 82%]	71% [58%, 79%]
	80 kVp	0.80 [0.71, 0.87]	72% [62%, 78%]	77% [66%, 86%]	67% [52%, 76%]
	Combined	0.82 [0.72, 0.88]	70% [63%, 76%]	75% [66%, 86%]	65% [53%, 74%]
sDL-Scratch	40 kVp	0.85 [0.78, 0.91]	78% [72%, 84%]	73% [62%, 82%]	83% [76%, 91%]
	80 kVp	0.77 [0.68, 0.84]	69% [62%, 76%]	73% [68%, 86%]	63% [53%, 72%]
	Combined	0.82 [0.73, 0.88]	80% [72%, 85%]	77% [68%, 86%]	83% [73%, 90%]

Table S5. Number of radiomic features retained after each step of the handcrafted radiomics (HCR) feature selection pipeline.

Steps	40 kVp CT	80 kVp CT	Combined CT
Overall	372	372	–
Constant & volume-correlated feature removal	331	329	–
Highly inter-correlated features removal	166	156	–
Lasso regression	12 ($\alpha = 0.0196$)	14 ($\alpha = 0.0113$)	26

Table S6. Select features of handcrafted radiomics (HCR). Twelve and fourteen features were selected in the 40 and 80 kVp CT, respectively.

40 kVp CT	80 kVp CT
LoG-1-FirstOrder_10Perc	LoG-1-FirstOrder_10Perc
LoG-1-FirstOrder_90Perc	LoG-1-FirstOrder_90Perc
LoG-1-FirstOrder_Min	LoG-1-GLSZM_GrayLvlNonUni
LoG-1-GLSZM_GrayLvlNonUni	LoG-2-FirstOrder_10Perc
LoG-1-GLSZM_SmallAreaHighGrayLvlEmp	LoG-2-GLCM_ClusterShade
LoG-2-GLRLM_GrayLvlVar	LoG-2-GLRLM_GrayLvlVar
LoG-2-GLSZM_GrayLvlNonUni	LoG-2-GLSZM_GrayLvlVar
LoG-3-FirstOrder_10Perc	LoG-2-GLSZM_HighGrayLvlZoneEmp
LoG-3-GLCM_ClusterShade	LoG-2-NGTDM_Busyness
LoG-3-GLSZM_LargeAreaLowGrayLvlEmp	LoG-3-FirstOrder_Median
LoG-3-GLSZM_SizeZoneNonUni	LoG-3-GLCM_ClusterShade
LoG-3-GLSZM_SmallAreaHighGrayLvlEmp	LoG-3-GLSZM_GrayLvlNonUni
	LoG-3-GLSZM_LargeAreaLowGrayLvlEmp
	LoG-3-GLSZM_SmallAreaHighGrayLvlEmp

Table S7. Prediction performance of HRD across different HCR models with five-fold cross-validation on training set.

Model	CT Type	AUC	Accuracy	Sensitivity	Specificity
SVM	40 kVp	0.79 [0.74, 0.85]	73% [68%, 78%]	69% [54%, 84%]	77% [60%, 94%]
	80 kVp	0.77 [0.68, 0.87]	71% [63%, 80%]	68% [44%, 91%]	74% [51%, 97%]
	Combined	0.79 [0.72, 0.85]	74% [68%, 80%]	76% [59%, 92%]	71% [54%, 87%]
RF	40 kVp	0.73 [0.65, 0.81]	68% [61%, 76%]	71% [61%, 80%]	67% [47%, 87%]
	80 kVp	0.65 [0.57, 0.73]	66% [59%, 73%]	53% [22%, 85%]	77% [57%, 97%]
	Combined	0.69 [0.56, 0.83]	66% [57%, 75%]	53% [33%, 73%]	80% [51%, 100%]
LR	40 kVp	0.73 [0.68, 0.77]	69% [64%, 74%]	67% [59%, 75%]	70% [57%, 84%]
	80 kVp	0.75 [0.64, 0.85]	71% [62%, 79%]	56% [34%, 79%]	86% [78%, 94%]
	Combined	0.71 [0.60, 0.82]	67% [59%, 76%]	60% [29%, 91%]	76% [52%, 100%]

Table S8. Prediction performance of HRD across different HCR models on test set.

Model	CT Type	AUC	Accuracy	Sensitivity	Specificity
SVM	40 kVp	0.76 [0.69, 0.84]	69% [62%, 77%]	66% [60%, 80%]	71% [57%, 77%]
	80 kVp	0.70 [0.61, 0.77]	65% [55%, 68%]	56% [50%, 70%]	75% [54%, 74%]
	Combined	0.71 [0.64, 0.79]	65% [57%, 72%]	66% [53%, 74%]	65% [53%, 75%]
RF	40 kVp	0.69 [0.61, 0.78]	62% [56%, 71%]	64% [58%, 77%]	60% [46%, 73%]
	80 kVp	0.70 [0.63, 0.79]	61% [57%, 72%]	42% [48%, 72%]	79% [58%, 79%]
	Combined	0.74 [0.68, 0.81]	69% [55%, 70%]	50% [57%, 78%]	89% [48%, 70%]
LR	40 kVp	0.63 [0.57, 0.72]	57% [54%, 68%]	53% [53%, 74%]	62% [47%, 67%]
	80 kVp	0.72 [0.64, 0.80]	63% [59%, 73%]	41% [51%, 75%]	86% [59%, 79%]
	Combined	0.67 [0.60, 0.77]	59% [56%, 70%]	47% [50%, 73%]	71% [53%, 76%]

Supplementary Figures

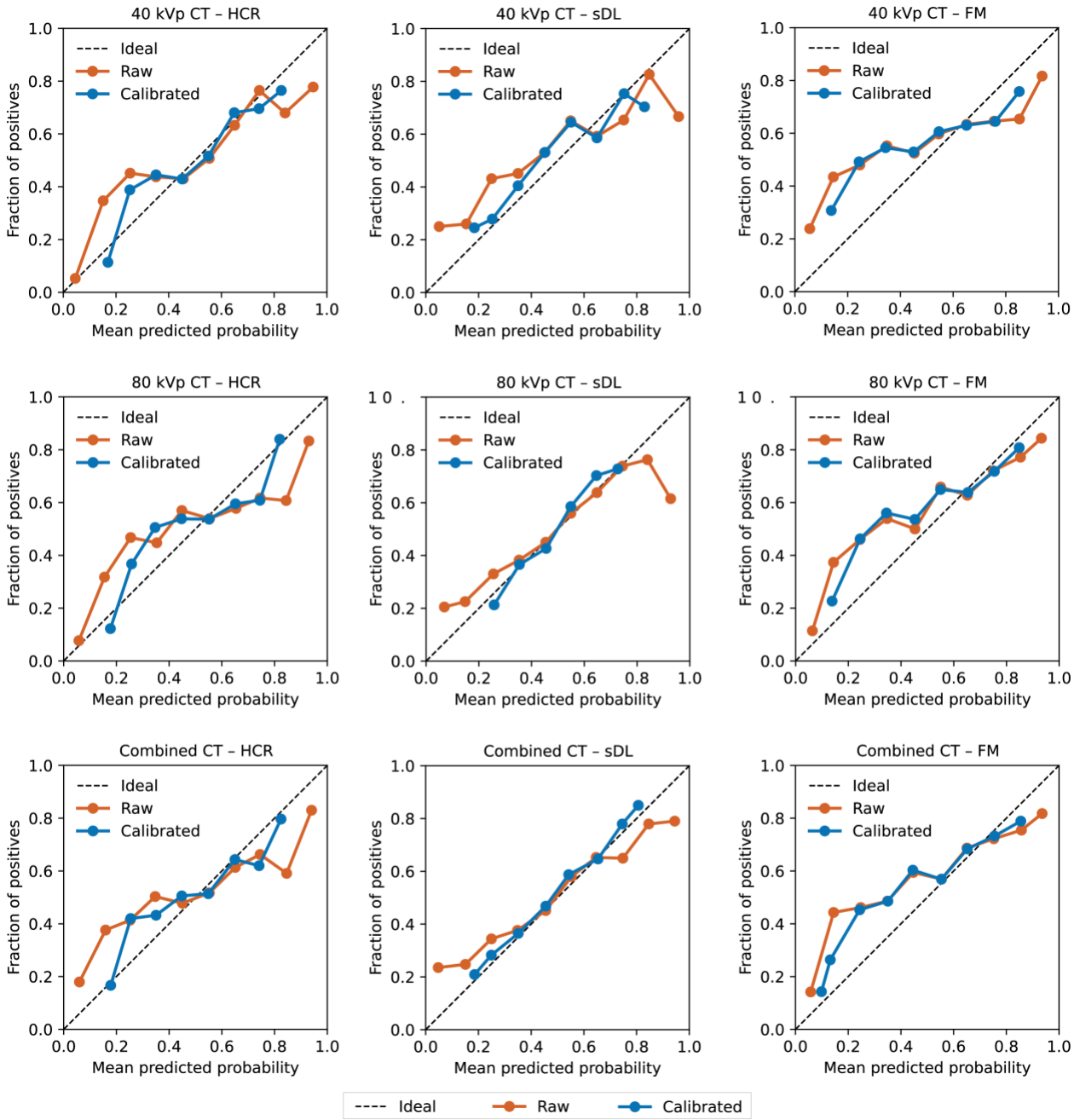

Figure S1. Reliability diagrams for the three HRD classification models (HCR, sDL, FM) across 40 kVp, 80 kVp and combined CT. Calibrators were fit on cross-validation folds and applied to the test predictions.

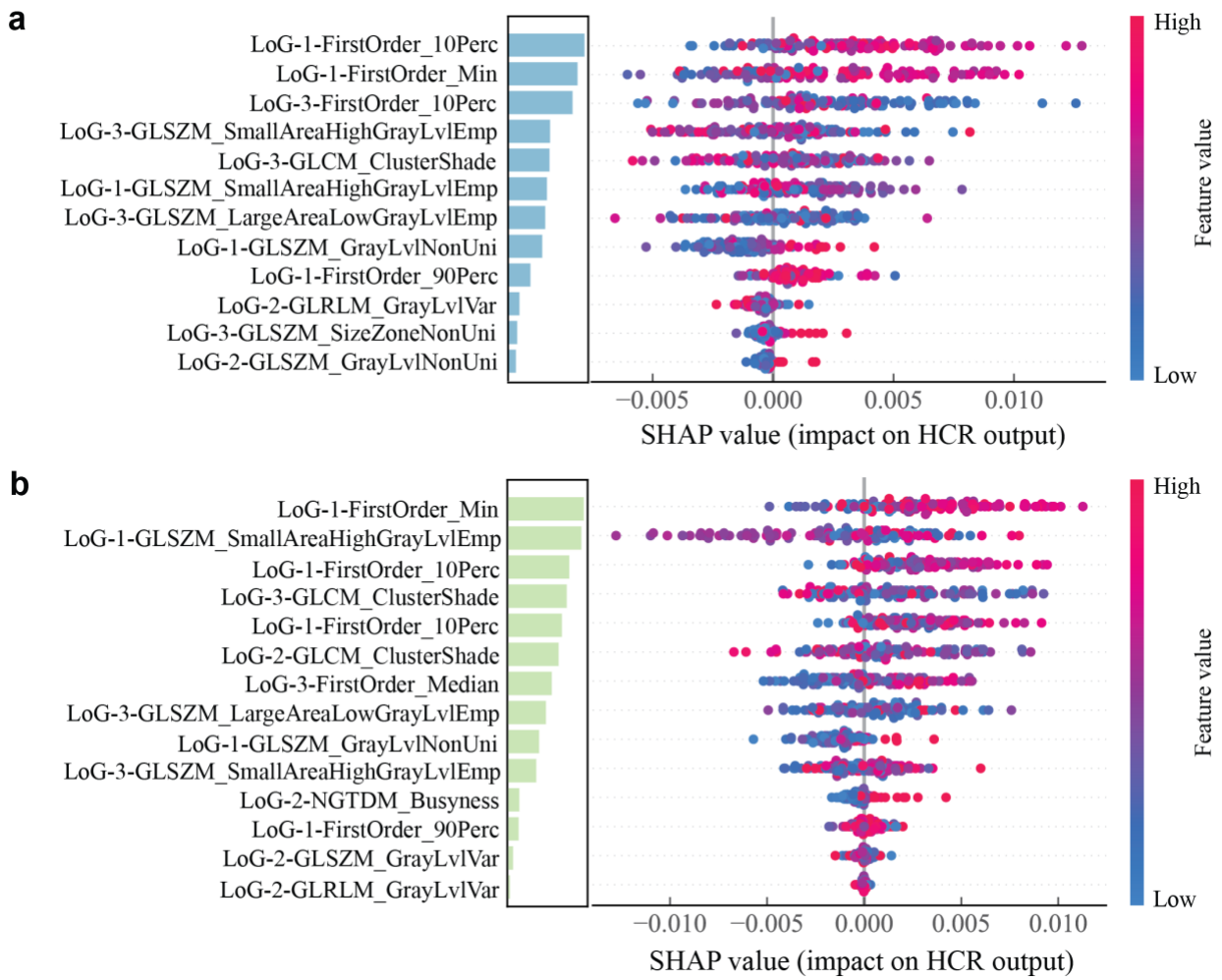

Figure S2. SHAP (SHapley Additive exPlanations) analysis of the SVM-based HCR model. (a) 40 kVp CT (b) 80 kVp CT. The left bar plot indicates feature impact ranked by mean absolute SHAP value, while the right bee swarm plot shows the SHAP distribution. Each point represents a single test subject with color encoding the original feature value (red: high, blue: low). Points to the right (positive SHAP) increase the predicted probability of HRD; points to the left (negative SHAP) decrease it.

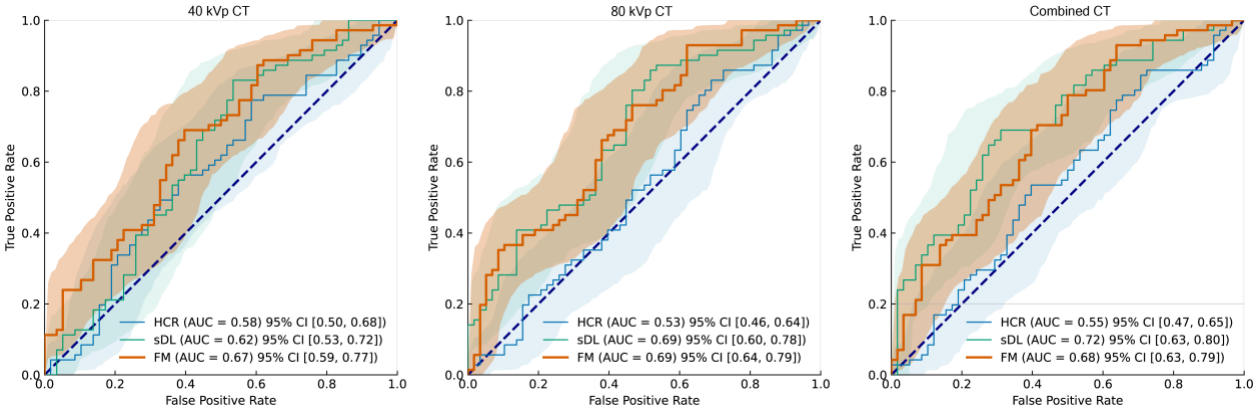

Figure S3. Receiver Operating Characteristic (ROC) curves and area under the curves (AUCs) for HRD classification on interference test set. ROC curves compare the performance of three models—handcrafted radiomics (HCR), supervised deep learning (sDL), and foundation model (FM)—for detecting HRD across different CT energy levels: 40 kVp CT, 80 kVp CT, and combined CT. The lines represent the ROC curves, while the error bars depict the 95% confidence intervals (CIs).

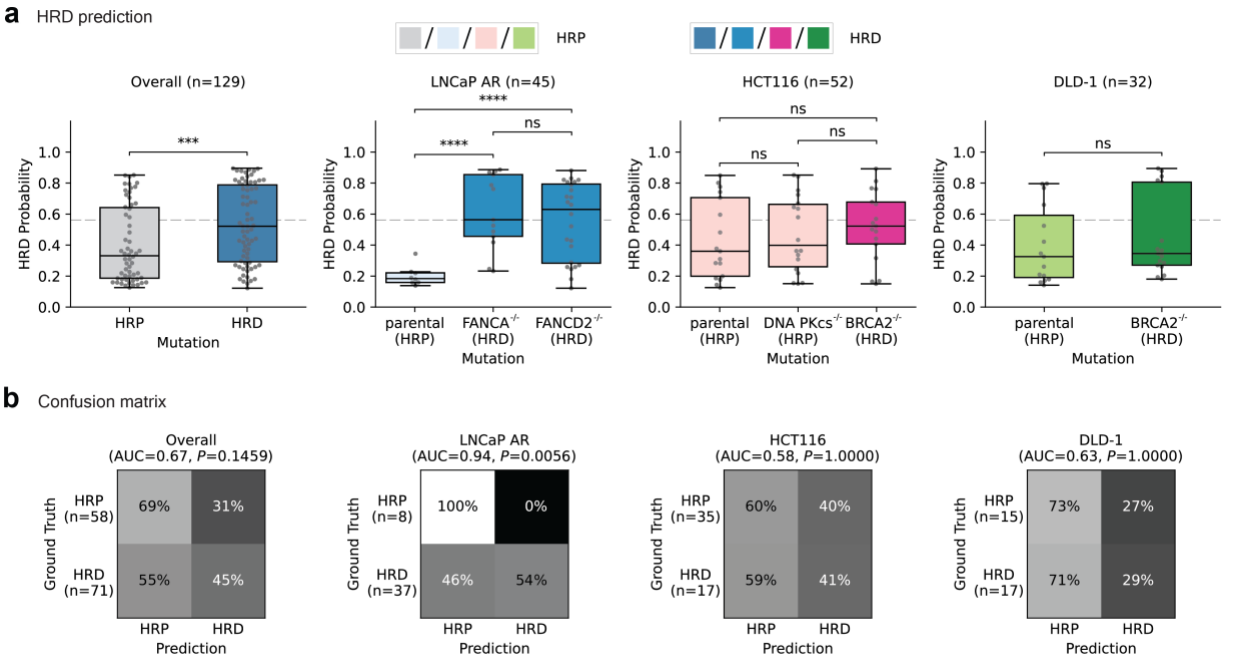

Figure S4. Performance of the foundation model in detecting HRD across the LNCaP AR, HCT116, and DLD-1 xenografts on interference test set. (a) Predicted HRD probability for both true HRP and HRD groups across the three cell lines. (b) AUC, sensitivity and specificity of the model's performance for all cell lines and individually for LNCaP AR, HCT116, and DLD-1. *$p < 0.05$, *** $p < 0.001$, **** $p < 0.0001$, ns. non-significant.

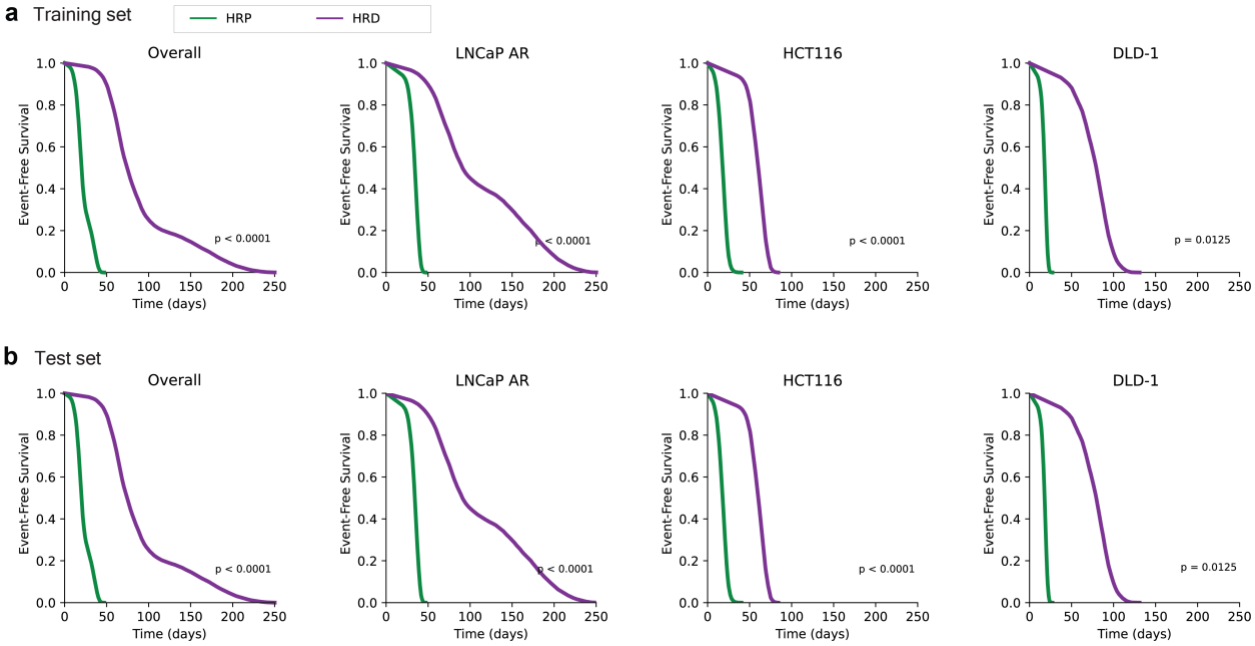

Figure S5. Kaplan-Meier survival curve of event-free survival (EFS) stratified by HRD classes in CP-506 treatment on the training and test sets.

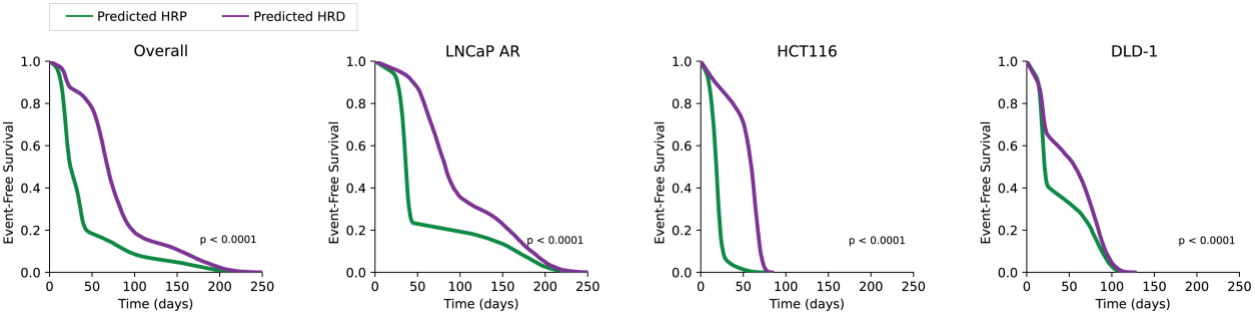

Figure S6. Kaplan-Meier survival curve of EFS stratified by predicted HRD classes in CP-506 treatment on test sets.

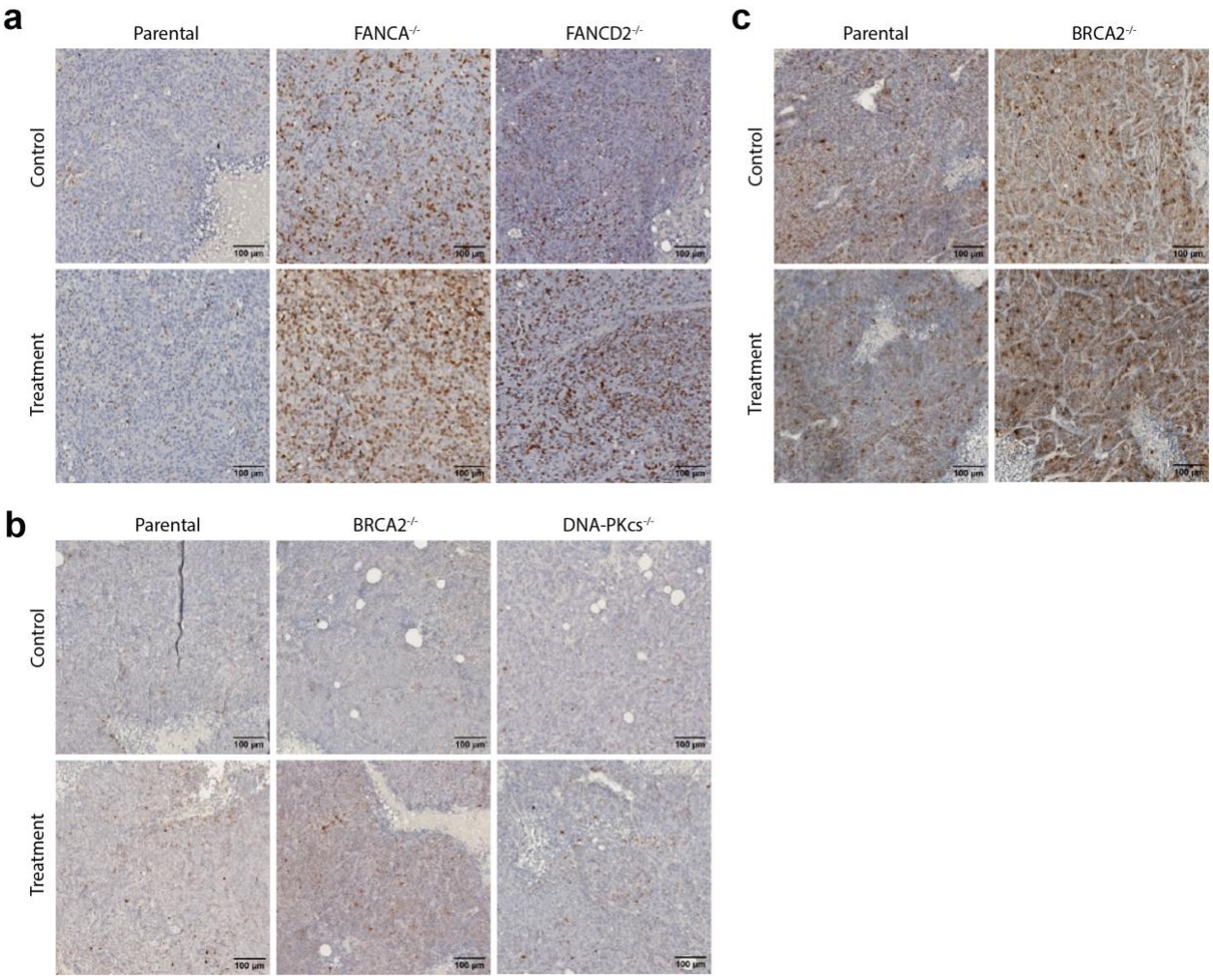

Figure S7. Representative γ-H2AX immunohistochemistry images 48 hours post treatment from (a) LNCaP AR, (b) HCT116, and (c) DLD-1 xenografts.