

1 Article

2 **Generalizing AI-driven Assessment of Immunohistochemistry across Immunostains and**  
3 **Cancer Types: A Universal Immunohistochemistry Analyzer**

4

5 Biagio Brattoli<sup>1,+</sup>, Mohammad Mostafavi<sup>1,+</sup>, Taebum Lee<sup>1,+</sup>, Wonkyung Jung<sup>1</sup>, Jeongun Ryu<sup>1</sup>,  
6 Seonwook Park<sup>1</sup>, Jongchan Park<sup>1</sup>, Sergio Pereira<sup>1</sup>, Seunghwan Shin<sup>1</sup>, Sangjoon Choi<sup>2</sup>, Hyojin  
7 Kim<sup>3</sup>, Donggeun Yoo<sup>1</sup>, Siraj M. Ali<sup>1</sup>, Kyunghyun Paeng<sup>1</sup>, Chan-Young Ock<sup>1</sup>, Soo Ick Cho<sup>1,\*</sup>,  
8 and Seokhwi Kim<sup>4,5\*</sup>

9

10 <sup>+</sup>These authors contributed equally to this work.

11 <sup>\*</sup>Corresponding authors

12

13 <sup>1</sup>Lunit, Seoul, Republic of Korea

14 <sup>2</sup>Department of Pathology and Translational Genomics, Samsung Medical Center,  
15 Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

16 <sup>3</sup>Department of Pathology, Seoul National University Bundang Hospital, Seongnam,  
17 Republic of Korea

18 <sup>4</sup>Department of Pathology, Ajou University School of Medicine, Suwon, Republic of  
19 Korea

20 <sup>5</sup>Department of Biomedical Sciences, Ajou University Graduate School of Medicine,  
21 Suwon, Republic of Korea

22

23 Correspondence to:

24 Seokhwi Kim, M.D., Ph.D.,

25 Department of Pathology, Ajou University School of Medicine.

26 Department of Biomedical Sciences, Ajou University Graduate School of Medicine.

27 164 Worldcup-ro, Yeongtong-gu, Suwon, 16499, Republic of Korea.

28 Tel: +82-31-219-6460; Fax: 82-31-219-5934; E-mail: [seokhwikim@ajou.ac.kr](mailto:seokhwikim@ajou.ac.kr)

29 ORCID ID: 0000-0001-7646-5064

30

31 Soo Ick Cho, M.D., Ph.D.,

32 Lunit.

33 374 Gangnam-daero, Gangnam-gu, Seoul, 06241, Republic of Korea.

34 Tel: +82-2-2138-0827; Fax: 82-2-6919-2702; E-mail: [sooickcho@lunit.io](mailto:sooickcho@lunit.io)

35 ORCID ID: 0000-0003-3414-9869

36 **ABSTRACT**

37 Despite advancements in methodologies, immunohistochemistry (IHC) remains the most  
38 utilized ancillary test for histopathologic and companion diagnostics in targeted therapies.  
39 However, objective IHC assessment poses challenges. Artificial intelligence (AI) has emerged  
40 as a potential solution, yet its development requires extensive training for each cancer and IHC  
41 type, limiting versatility. We developed a Universal IHC (UIHC) analyzer, an AI model for  
42 interpreting IHC images regardless of tumor or IHC types, using training datasets from various  
43 cancers stained for PD-L1 and/or HER2. This multi-cohort trained model outperforms  
44 conventional single-cohort models in interpreting unseen IHCs (Kappa score 0.578 vs. up to  
45 0.509) and consistently shows superior performance across different positive staining cutoff  
46 values. Qualitative analysis reveals that UIHC effectively clusters patches based on expression  
47 levels. The UIHC model also quantitatively assesses c-MET expression with MET mutations,  
48 representing a significant advancement in AI application in the era of personalized medicine  
49 and accumulating novel biomarkers.

50

## 51 **Introduction**

52 Immunohistochemistry (IHC) is an antibody-based methodology that can reveal the  
53 expression and distribution of proteins in formalin-fixed paraffin-embedded (FFPE) tissues and  
54 is well established as a decision support tool for oncology diagnosis<sup>1,2</sup>. IHC results are now  
55 increasingly used to guide decision making for systemic therapy for disseminated malignancy  
56 such as for the monoclonal antibody pembrolizumab in non-small cell lung cancer (NSCLC) as  
57 based on Programmed Death-Ligand 1 (PD-L1) expression<sup>3,4</sup>. Moreover, multiple emerging  
58 classes of therapies based on monoclonal antibodies (antibody-drug conjugates [ADC], bi-  
59 specific antibodies) directly target proteins on the tumor cell surface<sup>5,6</sup>. The efficacy of these cell  
60 surface-targeting therapeutics is consistently linked with the expression of the targeted protein.  
61 Therefore, quantifying IHC assessments of these targets may facilitate the development of  
62 predictive biomarkers that are valuable in clinical practice<sup>7</sup>.

63 Recently, artificial intelligence (AI) models have been developed to quantify IHC images  
64 by tissue segmentation, cell delineation, and quantification of all relevant cells in a whole slide  
65 image (WSI)<sup>8,9</sup>. However, the development of these AI models is heavily constrained by their  
66 reliance on single training cohorts that typically contain at least several hundred or often more  
67 WSI cases of a cancer type and immunostains matched to the desired indication. Moreover, these  
68 training sets are manually labeled on a cellular/subcellular basis by pathologists with each slide  
69 taking several hours for annotation depending on complexity<sup>10,11</sup>.

70 Importantly, there is an additional limitation of ‘domain-shift’, where current deep-  
71 learning models for IHC cannot recognize elements - either immunostain for cancer type - that  
72 are not present in the training set. This limitation indicates for each immunostain-cancer type  
73 combination, an IHC training set must be created and annotated with accompanying significant

74 time and resource cost, which is particularly relevant when evaluating new antibodies for  
75 development<sup>12,13</sup>. Both the requirement for expert annotated training sets specific to each desired  
76 permutation of immunostain and cancer type and the domain shift problem intertwine to create  
77 an imperative for a universally applicable AI model that is proficient in interpreting IHC results  
78 without antecedent manually annotated matching training sets<sup>14</sup>.

79 Here, we developed a Universal IHC (UIHC) analyzer, which can assess IHC images,  
80 irrespective of the specific immunostain or cancer type. Eight models trained on WSI patches  
81 from three cancer types, immunostained for PD-L1 or human epidermal growth factor receptor 2  
82 (HER2), were defined by exposure to varying single or multiple cohorts for training. Models  
83 trained on single cohorts served as the benchmark, whereas models trained with multiple cohorts  
84 were an innovation<sup>10,15,16</sup>. All models were evaluated using a diverse test set including eight  
85 'novel' IHC stained cohorts covering twenty additional cancer types, along with two 'training'  
86 IHC (PD-L1 and HER2) stained cohorts to identify the best model for further development.

87

## 88 **Results**

### 89 **Patch-level tumor cell detection and IHC-positivity classification**

90 We trained both single-cohort-derived models (SC-models) with one dataset and  
91 multiple-cohort-derived models (MC-models) with multiple datasets based on NSCLC,  
92 urothelial carcinoma, and breast cancer datasets stained with PD-L1 22C3 and breast cancer  
93 datasets stained with HER2 (Fig. 1). Fig. 2a shows the combination of different datasets to  
94 develop the eight AI models utilized in this study. SC-models exhibit favorable performance  
95 within test sets matched for the immunostain and cancer type used for training as evidenced by

96 the cell detection (negatively stained Tumor Cell [TC-] or positively stained Tumor Cell [TC+])  
97 performance (median F1-score [min, max]) of P-L (PD-L1 22C3 of lung) on the PD-L1 22C3  
98 Lung test set (0.693 [0.686, 0.705], [Fig. 2b](#)), P-BI (PD-L1 22C3 of bladder) on the PD-L1 22C3  
99 Bladder test set (0.725 [0.719, 0.731], [Fig. 2c](#)), P-Br (PD-L1 22C3 of breast) on the PD-L1 22C3  
100 Breast test set (0.599 [0.590, 0.607], [Fig. 2d](#)), and H-Br (HER2 of breast) on the HER2 test set  
101 (0.759 [0.753, 0.766], [Fig. 2e](#)). Notably, MC-models with broader exposure beyond the matched  
102 training set (P-LBIBr [PD-L1 22C3 for lung, bladder, and breast], PH-Br [PD-L1 22C3 and  
103 HER2 for breast], PH-LBr [PD-L1 22C3 and HER2 for lung and breast], PH-LBIBr [PD-L1  
104 22C3 and HER2 for lung, bladder, and breast]) performed as well as or better than the best  
105 performing SC-model for each test set matched to a training set, regardless of immunostain or  
106 cancer type. ([Fig. 2b-e](#)).

107 For test sets containing novel elements that were not seen in training, MC-models  
108 significantly outperformed SC-models. Notably, for the test set with an experienced  
109 immunostain but unseen cancer types, such as PD-L1 22C3 Pan-cancer set in [Fig. 2f](#), MC-  
110 models trained with more cancer types (P-LBIBr) and/or an additional stain (PH-LBr and PH-  
111 LBIBr) outperformed the SC-models P-L (P-LBIBr, 0.722 [0.716, 0.730],  $p < 0.001$ ; PH-LBr,  
112 0.745 [0.735, 0.753],  $p < 0.001$ ; PH-LBIBr, 0.743 [0.735, 0.752],  $p < 0.001$ ), which were the best  
113 performing SC-models.

114 In the other novel cohorts with unseen immunostains such as PD-L1 SP142, Claudin  
115 18.2, Delta-like 3 (DLL3), fibroblast growth factor receptor 2 (FGFR2), human epidermal  
116 growth factor receptor 3 (HER3), mesenchymal-epithelial transition factor (MET), MUC16, and  
117 trophoblast cell-surface antigen 2 (TROP2), MC-models generally performed better than SC-  
118 models [Fig. 2g-n](#)). Most representatively identified in MET Pan-cancer, all the MC-models

119 outperformed the single best performing SC-model H-Br (P-LBIBr, 0.795 [0.773, 0.810],  
120  $p < 0.001$ ; PH-Br, 0.762 [0.725, 0.776],  $p < 0.001$ ; PH-LBr, 0.783 [0.767, 0.799],  $p < 0.001$ ; PH-  
121 LBIBr, 0.792 [0.776, 0.815],  $p < 0.001$ ) (Fig. 2i). This tendency for MC-models to outperform  
122 SC-models was also observed when the data was categorized by cancer type (lung, breast,  
123 bladder, pan-ovary, esophagus, colorectum, and stomach) rather than IHC type (Supplementary  
124 Fig. 1).

125

## 126 WSI-level IHC quantification of MC- and SC-models

127 The performances of the AI models at the WSI level were subsequently assessed using  
128 the test sets outlined in Supplementary Table 1. The ground truth images were categorized and  
129 annotated based on the tumor proportion score (TPS), and the models' performance was  
130 evaluated by accurately assigning the WSIs to the corresponding ground truth group (TPS <1%;  
131 1-49%;  $\geq 50\%$ ). Among the eight models, the PH-LBIBr model was the top performer for this set  
132 of test WSI cohorts, achieving a Cohen's kappa score of 0.578 and an accuracy of 0.751 (Fig. 3a,  
133 Supplementary Fig. 2a). The best SC-model overall was H-Br, but it still had significantly lower  
134 performance, with a Cohen's kappa score of 0.509 and an accuracy of 0.703. In assessing  
135 performance on the PD-L1 22C3 Lung WSI test set, PH-LBIBr was the only model to  
136 outperform the SC-model P-L, with a Cohen's kappa score of 0.652 compared to 0.638 for P-L,  
137 and an accuracy of 0.793 compared to 0.785 for P-L (Fig. 3b, Supplementary Fig. 2b). For the  
138 PD-L1 22C3 Pan-cancer WSI test set and the PD-L1 SP142 Lung WSI test set, P-LBIBr also  
139 outperformed all SC-models, including P-L (Fig. 3c-d, Supplementary Fig. 2c-d). Notably, in the  
140 multi-stain pan-cancer test set, the PH-LBIBr model consistently outperformed all SC-models

141 (P-L, P-BI, P-Br, H-Br) and MC-models with less diversity in training cohorts (PH-Br, P-LBIBr,  
142 and PH-LBr), achieving a Cohen's kappa score of 0.610 and an accuracy of 0.757 (Fig. 3e,  
143 [Supplementary Fig. 2e](#)). Confusion matrices indicated that the PH-LBIBr model performed  
144 evenly across different TPS levels, with the highest number of concordance cases and  
145 mispredictions predominantly falling into adjacent categories (e.g., fewer mispredictions of  
146  $TPS < 1\%$  as  $TPS \geq 50\%$ ) (Fig. 4a-b). Due to its consistently high performance across test sets, PH-  
147 LBIBr is designated as the UIHC model.

148

#### 149 **Performance analysis of UIHC on novel immunostains for different cutoffs**

150 For certain immunostains commonly utilized in clinical practice, such as MET, TROP2,  
151 and MUC6, the absence of consensus scoring systems poses a challenge. To evaluate the false  
152 and true positive rates for these immunostains in our analysis, we initially established a cutoff at  
153 1% to maintain a standardized ground truth (GT)-TPS, similar to the approach used for PD-L1  
154 staining, while varying the AI model-predicted TPS cutoff. In this binary classification  
155 framework, the area under the receiver operating characteristics (AUROC) curve demonstrates  
156 that the selected UIHC model (92.1%) outperforms SC-Models (Fig. 5a). Additionally, we  
157 compared our AI models across a range of cutoffs from 1% to a second value within the range of  
158 [2%, 75%], illustrating a three-way classification accuracy of 78.7% (Fig. 5b). In both analyses,  
159 the UIHC model consistently exhibits superior performance, irrespective of the specific cutoff  
160 applied for novel stain types.

161



## 162 **Histopathologic validation of inference examples of UIHC model**

163           Representative discrepancy cases between the UIHC model and SC-models were  
164 subjected to WSI-level histopathological validation by pathologists (T.L., W.J., S.C., and S.K)  
165 to assess the accuracy of the models in detecting IHC-positive cells. In a case involving MET-  
166 stained NSCLC, the SC model P-L incorrectly classified the majority of tumor cells as negative  
167 (TPS 36%) (Fig. 6a). Conversely, the UIHC model accurately identified tumor cells based on  
168 positivity (TPS 61%), yielding results similar to the average TPS assessment of 75% by  
169 pathologists. In another instance concerning FGFR2-stained gastric cancer (Fig. 6b), the P-L  
170 model encountered difficulties, often failing to recognize numerous tumor cells and distinguish  
171 between FGFR2-positive and negative cells. In contrast, the UIHC model demonstrated an  
172 ability to discern IHC positivity even amidst this intricate staining pattern.

173

## 174 **Interpreting the representations of UMAP learned by the UIHC model**

175           To ensure the absence of inadvertent biases acquired during training, we evaluated the  
176 learned representations of the UIHC model using standard UMAP (uniform manifold and  
177 projection) for visualization. Two-dimensional internal representations of various AI models  
178 were presented in two formats: the 2D projection across training and novel cohorts (Fig. 7a), and  
179 a mosaic of image patches organized based on their respective projections (Fig. 7b).

180           In Fig. 7a, ground-truth TPS values were color-coded, transitioning from blue (0%) to  
181 brown (100%). For comparison, scatter plots were presented for three different sources: raw  
182 pixels as the baseline (Fig. 7a, left), features from a self-supervised learning (SSL) model trained  
183 with the same UIHC details but on larger, unannotated datasets (Fig. 7a, center), and the UIHC

184 model (Fig. 7a, right). The pixel model (Fig. 7a, left) exhibited weak clustering signal, with high  
185 TPS patches clustered towards the bottom-right. In the SSL model (Fig. 7a, center), clustering  
186 based on TPS was not observed, but rather clustering based on cohort. The 2D projection  
187 depicted in Fig. 7a, right illustrated that the UIHC model effectively separated and clustered  
188 patches based on TPS expression level. Our visual inspections of UMAP mitigated the Clever  
189 Hans effect (skewing of results by external biases) commonly observed in machine learning<sup>17</sup>.  
190 This analysis effectively demonstrates that our approach facilitated the development of an AI-  
191 powered analyzer capable of generalizing to novel immunostains and cancer types, even in IHCs  
192 with cytoplasmic staining not included in the training data, indicating superior feature extraction  
193 through exposure to multiple cohorts.

194 Fig. 7b presents a mosaic of original image patches arranged according to their internal  
195 representation as observed in Fig. 7a. In contrast to raw pixels, the features of the UIHC model  
196 were centered around tumor cell detection and classification rather than visual attributes derived  
197 from varying source characteristics such as color contrast or brightness. Thus, the pixel  
198 representation prioritizes sorting by color, while the UIHC model remains unbiased by  
199 appearance, focusing instead on tumor type. Cohort similarity results further indicate that only  
200 the UIHC model exhibits reduced sensitivity to cohort-specific traits, indicating its lack of bias  
201 towards IHC type and emphasis on the primary task of detecting and classifying tumor cells (Fig.  
202 7c).

203

## 204 **Performance analysis of UIHC on the real-world dataset**

205 To evaluate the UIHC model's applicability as a real-world assessment tool, we employed  
206 it to quantitatively assess the expression of c-MET, a novel immunostain for the model, in three

207 cohorts of NSCLC cases known to harbor oncogenic driver alterations - MET exon 14 skipping  
208 mutations, MET amplifications, and epidermal growth factor receptor (EGFR) exon 20 insertions  
209 <sup>18</sup>. The UIHC model assigned higher tumor proportion scores (TPS) to the MET amplification  
210 group compared to the other groups (Table 1). The UIHC model yielded a MET TPS of  $94.5 \pm 2.0$   
211 for the three MET amplification cases,  $77.1 \pm 17.7$  for the six exon 14 skipping mutation cases,  
212 and  $75.7 \pm 23.2$  for the seven EGFR exon 20 insertion EGFR cases.

213

## 214 Discussion

215 In this study, we demonstrated that UIHC trained with multiple cancer types and IHC,  
216 the MC-model, is not only superior in the domain used for training SC-model trained with a  
217 single cancer type and IHC in its domain but also exhibited the capability to analyze never-  
218 before-seen immunostains and cancer types.

219 Emerging therapeutic agents, meticulously designed to target surface proteins on tumor  
220 cells, have exerted a profound influence on the landscape of oncology care. These therapeutics  
221 can be broadly categorized into targeting tumor-associated antigens (TAA, such as TROP2) and  
222 targeting immune checkpoints (IC, such as PD-L1)<sup>19-21</sup>. Specific examples include trastuzumab  
223 deruxtecan, an ADC targeting HER2, and tarlatamab, a bispecific molecule targeting DLL3 and  
224 CD3<sup>22-24</sup>.

225 IHC stands as an essential component in cancer diagnosis, and thus far, the pathologist's  
226 reading remains the gold standard for determining the expression level of a target protein<sup>4,25-29</sup>.  
227 Nonetheless, discrepancies between pathologists and poor reproducibility can hinder precise  
228 evaluation<sup>10,15,16,30-33</sup>. Efforts have been made to standardize IHC assays to maintain its role as a

229 predictive biomarker, requiring evaluations as quantitative as possible. Recently introduced deep  
230 learning models have exhibited notable advantages over traditional computational methods,  
231 primarily due to their capacity to discern intricate patterns within IHC images where the latter  
232 requires the pathologists to understand the tissue structure and morphology directly per case  
233 before analysis<sup>34,35</sup>. These models can analyze PD-L1 and HER2 expression but require training  
234 on a large, manually annotated training cohort<sup>10,11,15,16,36-38</sup>. Moreover, such deep-learning  
235 models have domain shift issues that are effective within the cancer type and immunostain  
236 defined by the training cohort, but not for indications that contain cancer types and  
237 immunostains not within the training set<sup>12,13,39</sup>.

238 In the present study, AI models underwent training using either a single cohort (SC) or  
239 multiple cohorts (MC). The MC-models, particularly those exposed to the most diverse range of  
240 cases, demonstrated superior performance compared to the SC-models. This was evident across  
241 test sets similar to the training cohorts, as well as test cohorts composed of previously unexposed  
242 (novel) immunostains and cancer types. The enhanced performance of MC-models in training  
243 cohorts can be attributed to the augmented training data. Compared to the H-Br model, the PH-  
244 Br model showed better performance on PD-L1 22C3 breast and HER2 breast, indicating the  
245 impact of increasing the volume of training data. However, the superiority of PH-Br over PH-  
246 LBr in PD-L1 22C3 bladder, which was trained with a larger cohort than PH-Br, suggests that  
247 the influence of expanding the training data volume is not straightforward. Irrespective of the  
248 volume of training data, training models using cohorts from various cancer types or  
249 immunostains together contributed to improve model performance. This phenomenon is  
250 exemplified in PD-L1 22C3 Pan-cancer, where PH-LBr, encompassing variations in both cancer  
251 type and immunostain, outperforms P-LBIBr, which only varies in cancer type, or PH-Br, which

252 only varies in immunostain. The impact of variations in cancer type or immunostain within the  
253 training data is underscored by the superior performance of MC-models compared to SC-  
254 models, particularly evident for novel cohorts. Conversely, in the case of novel cohorts such as  
255 FGFR2 IHCs, where both membrane and cytoplasmic intensity can be observed, AI models  
256 trained solely on membranous staining IHCs (e.g., PD-L1 22C3 and HER2) may experience  
257 significant performance degradation<sup>40,41</sup>. Indeed, among the novel cohorts, both SC- and MC-  
258 models exhibited the poorest performance on FGFR. The UIHC model, however, demonstrated  
259 superior performance compared to other models, particularly in detecting cytoplasmic stained  
260 TC+, whereas most SC-models struggled to identify cytoplasmic stained TC+.

261 Recent AI-related research disciplines can be divided into the two main branches of  
262 model-centric and data-centric AI<sup>42</sup>. The model-centric AI focuses on designing and optimizing  
263 the best AI models with a fixed dataset, while data-centric AI systematically and algorithmically  
264 focuses on providing the best dataset for a fixed AI model. Our study underscores the promising  
265 efficacy of training the AI model with diverse IHC and cancer type data. Notably, this is  
266 clinically meaningful because it was done without additional data work, mostly annotation in a  
267 novel cohort, so it can be applied directly to new targets. Recent trends tend to call approaches  
268 with large training set from different domains ‘foundational models’, therefore, in this sense, our  
269 UIHC could be considered one<sup>43-45</sup>. However, we reserve this name for a multi-modal system  
270 that goes beyond histopathology and combines multiple medical disciplines<sup>46</sup>.

271 To demonstrate the possible clinical utility of the current analyzer, we assessed c-MET  
272 expression in NSCLC to address the long-standing question of targeting c-MET. MET  
273 amplification is strongly believed to be correlated with increased expression of c-MET, however,  
274 so are exon 14 splicing mutations in c-MET (METex14m)<sup>47,48</sup>. Specifically, these mutations

275 lead to the omission of exon 14 and the Cbl sites which are thought to be recognized by an E3  
276 ubiquitin ligase, and thus thought to increase the amount of c-MET expressed by the tumor cell<sup>49</sup>.  
277 As theorized, c-MET amplifications lead to high expression of c-MET as seen in previous  
278 studies<sup>47,50</sup>. In contrast, tumors with METex14m had similar expression to exon 20 insertion  
279 NSCLC driven tumors. These unripe findings should be replicated in a larger cohort, but are  
280 very relevant to the development and clinical use of large molecular therapeutics targeting c-  
281 MET such as amivantanab<sup>22</sup>.

282         There are some limitations in this work. The current scope of IHC expression detection is  
283 confined to tumor cells, but not other cell types, i.e. lymphocytes and macrophages. However,  
284 the UIHC model is able to learn to assess these other cell types if given the correct training sets  
285 as consistent with a data-centric approach. Furthermore, our IHC evaluation was limited to a  
286 binary categorization of positive or negative, but will encompass multi-level protein expression  
287 assessments such as the American Society of Clinical Oncology (ASCO) / College of American  
288 Pathologists (CAP) guidelines for HER2 in the future<sup>51</sup>. In addition, the model's performance  
289 demonstrated some variability across different staining techniques and cancer types within this  
290 study. This concern could potentially be addressed through the inclusion of additional IHC stain  
291 types within the model's training dataset, in other words exposing the model to more multiple  
292 cohorts in training.

293         In conclusion, we have successfully developed a UIHC model capable of autonomously  
294 analyzing novel stains across diverse cancer types. In contrast to prevailing literature and  
295 existing image analysis products that often focus on specialized cohorts, our model's versatility  
296 and agility significantly enhance its potential in expediting research related to new IHC  
297 antibodies<sup>34</sup>. This innovative approach not only facilitates a broad spectrum of novel biomarker

298 investigations but also holds the potential to assist in the development of pioneering

299 therapeutics.

300

## 301 **Methods**

### 302 **Dataset preparation for AI model development**

#### 303 Histopathology dataset for annotation

304 The dataset used to develop the model consists of a total of 3,046 WSIs including lung  
305 (NSCLC), urothelial carcinoma, and breast cancer cases stained for PD-L1 22C3 pharmDx IHC  
306 (Agilent Technologies, Santa Clara, CA) and breast cancer WSIs stained for anti-HER2/neu  
307 (4B5) (Ventana Medical Systems, Tucson, AZ), as reported in previous studies ([Fig. 1](#),  
308 [Supplementary Table S2](#))<sup>10,15,16</sup>. All data for this study were obtained from commercially  
309 available sources from Cureline Inc. (Brisbane, CA, US), Aurora Diagnostics (Greensboro, NC,  
310 US), Neogenomics (Fort Myers, FL, US), Superbiochips (Seoul, Republic of Korea) or were  
311 available by the permission of Institutional Review Board (IRB) from Samsung Medical Center  
312 (IRB no. 2018-06-103), Seoul National University Bundang Hospital (IRB no. B-2101/660-30),  
313 and Ajou University Medical Center (IRB no. AJOU-IRB-KS-2023-425). All slide images and  
314 clinical information were de-identified and pseudonymized.

315 The WSIs were divided into training, tuning (also called validation), and test sets. Since  
316 WSIs are too large for computation, a section of size  $0.04\text{mm}^2$  (patch, i.e. tile) is extracted.

317 To evaluate and compare the models, we collected patch-level test sets from ten different  
318 stain types: PD-L1 22C3 (lung, bladder, breast, liver, prostate, colorectum, stomach, biliary tract,  
319 and pancreas), HER2 (breast), PD-L1 SP142 (lung), various immunostain types including  
320 Claudin 18.2, DLL3, FGFR2, HER3, MET, MUC16, and TROP2 (pan-cancer). The test sets of  
321 PD-L1 22C3 lung, bladder, and breast originated from the same cohort of training and tuning  
322 sets mentioned above (internal test set in [Supplementary Table 2](#)), which could be referred to as



323 training domain. Patches of PD-L1 22C3 other than lung, bladder, and breast were from WSIs of  
324 colorectum (n = 19), liver (n = 20), stomach (n = 18), prostate (n = 18), pancreas (n = 19), and  
325 biliary tract (n = 20). For the novel domain test set, which is never shown to the AI model during  
326 training, we collect patches from novel cancer types and novel immunostain types. Additionally,  
327 patches of various immunostain types were from pan-cancer (more than 25 cancer types) tissue  
328 microarray (TMA) cores (Superbiochips, Seoul, Republic of Korea)<sup>52-55</sup>. Detailed information on  
329 antibodies for various immunostain types is provided in [Supplementary Table 3](#). All slides were  
330 scanned by P1000 scanner (3DHistech, Budapest, Hungary) or Aperio AT2 scanner (Leica  
331 Biosystems Imaging, Buffalo Grove, IL, US). Within a WSI, up to three patches are selected and  
332 then resized to 1024x1024 pixels, at a normalized Microns-Per Pixel (MPP) of 0.19  $\mu\text{m}$ . Such  
333 MPP normalization is required to unify the resolution of the patches since WSIs scanned from  
334 different scanners can have different MPP values. The patches are extracted manually to avoid  
335 uninteresting areas, such as the white background. No patches of the same WSI can be found in  
336 different sets, to prevent information leakage between the training and test sets.

337

### 338 Patch-level annotation for AI model development

339 We define two general cell classes for IHC by TC- or TC+ ([Fig. 8a](#)). In most of the IHC  
340 staining, except HER2, the expression was described as either positive or negative. Patches  
341 stained with HER2 are traditionally annotated with four levels of IHC quantification as follows;  
342 H0 (negative), H1+ (faint/barely perceptible and incomplete membrane staining), H2+ (weak to  
343 moderate complete membrane staining), and H3+ (complete, intense circumferential membrane  
344 staining)<sup>56</sup>. To unify the categories across stains, we remapped H0 to negatively stained Tumor  
345 Cell (TC-) and the remaining H1~H3 to positively stained Tumor Cell (TC+).

346 All annotations were performed by board-certified pathologists. The interpretation of  
347 tumor cell positivity by pathologists was determined by following the guidelines for PD-L1 or  
348 HER2<sup>51,57</sup>. The training set was composed of 574,620 TC+ and 1,415,033 TC-, while the tuning  
349 set contained 138,429 TC+ and 316,808 TC- ([Supplementary Table 4, Fig. 8b](#)). The tuning set  
350 was used to select the best checkpoint during the model training phase. The total TC+ and TC-  
351 annotated from the patches of the test set are described in [Supplementary Table 5](#).

352

### 353 WSI-level test sets for AI model performance validation

354 Given that a single patch is a tiny fraction (<1%) of a WSI, performance of any model on  
355 the WSI-level can significantly deviate from patch-level assessment<sup>58</sup>. Therefore, a  
356 comprehensive comparison of our model performance on WSI was conducted with the key  
357 output of WSI-level TPS<sup>59,60</sup>.

358 We collected four WSI-level test sets: PD-L1 22C3 lung (n = 479), PD-L1 22C3 pan-  
359 cancer (n = 135), PD-L1 SP142 lung (n = 178) and a novel, multi-stain test set (n = 140) as  
360 presented in [Fig. 8c](#) and [Supplementary Table 1](#). The test set containing PD-L1 22C3 lung  
361 cancer was used in previous publications.<sup>(10)</sup> The PD-L1 22C3 Pan-cancer contains cancer  
362 types of biliary tract (n = 23), colorectum (n = 23), liver (n = 23), stomach (n = 23), prostate (n =  
363 22), and pancreas (n = 21). The test set containing PD-L1 SP142 lung (n = 178) was derived  
364 from the same cohort of PD-L1 22C lung cancer. IHC in the multi-stain test set included MET,  
365 MUC16, HER3, TROP2, DLL3, FGFR2, Claudin 18.2, SP142, and E-Cadherin across ten  
366 cancer types. Except for PD-L1 22C3 lung cancer, they all corresponded to novel domains.  
367 Representative image samples from both training and novel groups are illustrated in

368 [Supplementary Fig. 4 and 5.](#)

369 The multi-stain test set contains following stains: Claudin18.2 (n = 18), DLL3 (n = 16),  
370 E-Cadherin (n = 10), FGFR2 (n = 18), HER3 (n = 15), MET (n = 25), MUC16 (n = 16), PD-L1  
371 SP142 (n = 10), and TROP2 (n = 12) across ten cancer types (lung, breast, bladder, cervix,  
372 colorectum, esophagus, liver, lung, melanoma, stomach). Within the multi-stain test set, except  
373 for PD-L1 SP142 which is applied only on lung cancer, other staining antibodies (n = 130) are  
374 used for: stomach (n = 39), bladder (n = 28), breast (n = 23), lung (n = 19), cervix (n = 5),  
375 esophagus (n = 5), melanoma (n = 4), colorectum (n = 3), head and neck (n = 3), liver (n = 1).

376 TPS evaluation for all datasets was performed by three independent board-certified  
377 pathologists (S.C., H.K., and S.K. for PD-L1 22C3 lung, S.C., W.J., and S.K. for PD-L1 SP142  
378 lung, and T.L., S.C., and S.K. for PD-L1 22C3 pan-cancer and multi-stain set).

379

## 380 **AI Model development process**

### 381 Development of Universal IHC algorithm

382 Our approach's inference pipeline consists of training dataset preparation, AI model  
383 development, and performance validation with diverse cohorts ([Fig. 1](#)). Specifically, after  
384 extracting patches and annotating cells from designated training cohorts, several AI models are  
385 trained with single-cohort (standard approach) or multiple-cohort data (innovation). Each  
386 model's parameters have been tuned using their domain-specific tuning (validation) set. Using  
387 combinations of the above cohorts, we produce eight models as described in [Fig. 2a](#). While SC-  
388 models (H-Br, P-L, P-Bl, and P-Br) are trained on a single cohort<sup>10,15,16</sup>. MC-models such as P-  
389 LBlBr, PH-Br, PH-LBr, and PH-LBlBr are trained on multiple cohorts. Among these candidate

390 models, we aim to identify the model that exhibits the highest degree of generalization for  
391 designation as a UIHC model.

392 Models are then tested on patches or WSIs exclusively held out from the training dataset.  
393 Our testing encompassed multiple cohorts, including ‘training cohorts’ and ‘novel cohorts’.  
394 Most patches posed greater challenges as the staining proteins or cancer types were not part of  
395 the training data for any AI models presented in this study.

396

### 397 Label pre-processing

398 Inspired by the previous work that trains the cell detection model with point annotations,  
399 we define cell detection as a segmentation task<sup>13,61</sup>. At training time, we provide the cell labels  
400 as a segmentation map by drawing a disk centered on each cell point annotation. We use a fixed  
401 radius of  $\sim 1.3\mu m$ , corresponding to 7 pixels at a resolution of 0.19 MPP. Finally, we assign the  
402 value of pixels within each disk based on the class of a cell, ‘1’ for TC-, ‘2’ for TC+. ‘0’ is  
403 assigned for the remaining pixels.

404

### 405 Inference post-processing

406 Given that we treat cell detection as a segmentation task, a post-processing phase is  
407 needed to extract 2D coordinates and classes of predicted cells from the probability map output  
408 by the network. We apply *skim-age.feature.peak\_local\_max* on the model’s output, which finds  
409 the locations of local maximums of the probability map to get the set of predicted cell points<sup>62</sup>.  
410 Lastly, we obtain each cell’s class and probability value in the cell segmentation map through  
411 *argmax*. This probability is used as the confidence score.

412

413 Network architecture and training details

414 For all of our models, we use DeepLabV3+ as our base architecture with a ResNet34  
415 encoder, which is a popular architecture specifically designed for the segmentation task<sup>63,64</sup>.  
416 During training, we augment the patches with a set of standard data augmentation methods for  
417 computer vision. In particular, we utilize center crop, horizontal and vertical flip, rotation,  
418 gaussian noise, color jittering, and gray scaling. Random values are sampled for each  
419 augmentation every time an image is loaded. Network parameters are initialized by Kaiming  
420 initialization<sup>65</sup>. The model is optimized using the Adam optimizer<sup>66</sup>. Dice loss is used to train the  
421 model<sup>67</sup>. The initial learning rate is set to  $1e-4$ , adapted using the cosine learning rate  
422 scheduler<sup>68</sup>. All the models have been trained for 150 epochs and evaluated at every 10 epochs to  
423 choose the best checkpoint on a hold-out tuning set. An epoch that shows the highest mF1 score  
424 on the tuning set is chosen as the best epoch and used for all evaluation purposes. All of the  
425 models are trained and evaluated with the same machine specifications as follows: 4 NVIDIA  
426 Tesla T4 GPUs each with 16GB of GPU memory and 216GB of RAM.

427

428 Inference details on whole slide images (WSIs)

429 For WSI inference we use the full WSI for tumor proportion score (TPS) calculation,  
430 excluding white background and in-house control tissue regions. The WSI is divided into  
431  $1024 \times 1024$  pixels of non-overlapping patches with an MPP of normalized 0.19 (following the  
432 training data), which are fed to our network, producing a prediction map with the same size as  
433 the input. All outputs are then combined to obtain a prediction map for the full WSI.

434

## 435 **AI Model evaluation**

### 436 Metrics for AI model performance

437           The performance evaluation of the AI model was analyzed at the patch-level and WSI-  
438 level. At the patch-level, performance was measured by F1 score, which compares the results of  
439 pathologists' annotation of each cell with the results of the AI model. At the slide-level, TPS by  
440 pathologists or UIHC was divided into categories based on a given cutoff threshold (1%/50% [3  
441 classes]). Then performance was evaluated by comparing the TPS categories from pathologists  
442 to the AI model using Cohen's Kappa. The details of F1 score and TPS are described in the  
443 Supplementary methods.

444

### 445 Model interpretation by visualization of data distribution

446           To gain deeper insights into the learned patterns of the UIHC model, we delved into its  
447 inference process by extracting internal representations of the network for each image patch in  
448 the test set. We visualized these representations in 2D using UMAP, a widely used method for  
449 dimensionality reduction method for visualization<sup>69</sup>. We utilized a 2D projection where each  
450 point is a patch and the Euclidean distance between two points indicates the similarity within the  
451 network's internal representation. For this experiment, we developed two baselines to provide  
452 context for our UIHC:

- 453           1. Raw pixel representation, by simply downsizing the image from  $1024 \times 1024 \times 3$  to  
454            $32 \times 32 \times 3$  and flattening the pixels, producing a  $1 \times 3072$  vector. RGB-channel is kept since  
455           the color is important for IHC quantification. For the same reason this is a valid baseline, in

456 fact, simply looking at the intensity of the brown color is a good indicator for TC-/+.

457 2. We developed a second network for comparing two deep learning models. Since patches  
458 with a manual annotation are a small fraction of all slides, we trained a ResNet34 using a  
459 state-of-the-art SSL method called Barlow Twins instead<sup>70</sup>. This allowed us to train the  
460 model on a large number of histopathology patches from different types of stains (PD-L1  
461 22C3, and HER2) and cancer types without the need for any manual annotations.

462 3. The best UIHC model is used as the representative UIHC model for the qualitative analysis.

463 To extract the internal representation from the deep learning models (UIHC and SSL),  
464 each patch runs through the ResNet34 encoder producing a  $16 \times 16 \times 512$  tensor of shape  
465 Height  $\times$  Width  $\times$  Channels. The output tensor is averaged over spatial dimensions, thus producing  
466 a  $1 \times 512$  vector. After producing a vector for each of the  $N$  patches in our test set, we obtain a  
467 matrix of  $N \times 512$  ( $N \times 3072$  for *Pixel*). Finally, we can project the  $2858 \times 512$  matrix to  $N \times 2$  by  
468 using UMAP, a popular non-linear dimensionality reduction algorithm<sup>69</sup>. This 2-dimension  
469 matrix can be easily plotted as a scatter plot using *matplotlib* and *seaborn*. To calculate cohort  
470 similarities, we compute the Wilcoxon test between all cohort pairs, producing a similarity  
471 matrix of size [# cohorts  $\times$  # cohorts] containing p-values. Then we average the upper-triangular  
472 matrix shown in the bar plot. In addition, the mosaic of image patch is drawn by discretizing the  
473 latent representations and replacing each point with the corresponding original patch image<sup>71</sup>.  
474 For each discretized point in space, the median patch is selected as the representative of that  
475 cluster.

476

## 477 Analysis of a genomically defined MET NSCLC dataset with the UIHC model

478 To further validate the performance of the UIHC model, we ran AI model inference on  
479 MET-stained NSCLC WSIs (n = 15) with gene mutation/amplification profiles. The cases were  
480 all diagnosed with NSCLC at Ajou University Medical Center and confirmed by next-generation  
481 sequencing to have either EGFR exon20ins, MET exon skipping, or MET amplification  
482 alterations.

483

484

### 485 **Reporting summary**

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

488

### 489 **Data availability**

490 The processed data can be provided by the corresponding authors after formal requests  
491 and assurances of confidentiality are provided.

492

### 493 **Code availability**

494 Deep-learning-related code was implemented using *pytorch* version 1.12, Python version  
495 3.9 and publicly available neural network architectures, like ResNet (open-source available  
496 online, e.g. <https://github.com/pytorch/vision/blob/main/torchvision/models/resnet.py>) and  
497 DeepLabV3 (open-source available online, e.g. [https://github.com/VainF/DeepLabV3Plus-  
498 Pytorch](https://github.com/VainF/DeepLabV3Plus-Pytorch)). For UMAP we utilize the official, open-source implementation (available at



499 <https://umap-learn.readthedocs.io/en/latest/clustering.html#using-umap-for-clustering>). All plots  
500 were generated with publicly available libraries, matplotlib version 3.5.2 (available at  
501 <https://github.com/matplotlib/matplotlib/tree/v3.5.2>) and seaborn version 0.12.2 (available at  
502 <https://seaborn.pydata.org/whatsnew/v0.12.2.html>), using Google Colab (available at  
503 <https://colab.google/>).  
504

505 **References**

- 506 1. Stack, E. C., Wang, C., Roman, K. A. & Hoyt, C. C. Multiplexed immunohistochemistry,  
507 imaging, and quantitation: a review, with an assessment of Tyramide signal amplification,  
508 multispectral imaging and multiplex analysis. *Methods* **70**, 46–58 (2014).
- 509 2. Cregger, M., Berger, A. J. & Rimm, D. L. Immunohistochemistry and quantitative analysis  
510 of protein expression. *Arch. Pathol. Lab. Med.* **130**, 1026–1030 (2006).
- 511 3. Slamon, D. J. et al. Use of Chemotherapy plus a Monoclonal Antibody against HER2 for  
512 Metastatic Breast Cancer That Overexpresses HER2. *N. Engl. J. Med.* **344**, 783–792  
513 (2001).
- 514 4. Garon, E. B. et al. Pembrolizumab for the Treatment of Non–Small-Cell Lung Cancer. *N.*  
515 *Engl. J. Med.* **372**, 2018–2028 (2015).
- 516 5. Fuentes-Antrás, J., Genta S., Vijenthira, A. & Siu, L. L. Antibody–drug conjugates: In  
517 search of partners of choice. *Trends Cancer* (2023).
- 518 6. Qian, L. et al. The Dawn of a New Era: Targeting the “Undruggables” with Antibody-  
519 Based Therapeutics. *Chem. Rev.* **123**, 7782–7853 (2023).
- 520 7. Patel, S. P. & Kurzrock, R. PD-L1 expression as a predictive biomarker in cancer  
521 immunotherapy. *Mol. Cancer Ther.* **14**, 847–856 (2015).
- 522 8. Baxi, V., Edwards, R., Montalto, M. & Saha, S. Digital pathology and artificial intelligence  
523 in translational medicine and clinical practice. *Mod. Pathol.* **35**, 23–32 (2022).
- 524 9. Ibrahim, A. et al. Artificial intelligence in digital breast pathology: techniques and  
525 applications. *The Breast* **49**, 267–273 (2020).
- 526 10. Choi, S. et al. Artificial intelligence–powered programmed death ligand 1 analyser reduces  
527 interobserver variation in tumour proportion score for non–small cell lung cancer with

- 528 better prediction of immunotherapy response. *Eur. J. Cancer* **170**, 17–26 (2022).
- 529 11. Wu, S. et al. The role of artificial intelligence in accurate interpretation of HER2  
530 immunohistochemical scores 0 and 1+ in breast cancer. *Mod. Pathol.* **36**, 100054 (2023).
- 531 12. Wang, Z. et al. Global and local attentional feature alignment for domain adaptive nuclei  
532 detection in histopathology images. *Artif. Intell. Med.* **132**, 102341 (2022).
- 533 13. Swiderska-Chadaj, Z. et al. Learning to detect lymphocytes in immunohistochemistry with  
534 deep learning. *Med. Image Anal.* **58**, 101547 (2019).
- 535 14. Zhou, K., Liu, Z., Qiao, Y., Xiang, T. & Loy, C. C. Domain generalization: A survey. *IEEE*  
536 *Trans. Pattern Anal. Mach. Intell.* (2022).
- 537 15. Jung, M. et al. Augmented interpretation of HER2, ER, and PR in breast cancer by artificial  
538 intelligence analyzer: enhancing interobserver agreement through a reader study of 201  
539 cases. *Breast Cancer Res.* **26**, 31 (2024).
- 540 16. Lee, K. S. et al. An artificial intelligence-powered PD-L1 combined positive score  
541 (CPS) analyser in urothelial carcinoma alleviating interobserver and intersite variability.  
542 *Histopathology*, his.15176 (2024).
- 543 17. Lapuschkin, S. et al. Unmasking Clever Hans predictors and assessing what machines  
544 really learn. *Nat. Commun.* **10**, 1096 (2019).
- 545 18. Li, M. M. et al. Standards and guidelines for the interpretation and reporting of sequence  
546 variants in cancer: a joint consensus recommendation of the Association for Molecular  
547 Pathology, American Society of Clinical Oncology, and College of American Pathologists.  
548 *J. Mol. Diagn.* **19**, 4–23 (2017).
- 549 19. Darvin, P., Toor, S. M., Sasidharan Nair V. & Elkord, E. Immune checkpoint inhibitors:  
550 recent progress and potential biomarkers. *Exp. Mol. Med.* **50**, 1–11 (2018).

- 551 20. Liu, C.-C., Yang, H., Zhang, R., Zhao, J.-J. & Hao, D.-J. Tumour-associated antigens and  
552 their anti-cancer applications. *Eur. J. Cancer Care (Engl.)* **26**, e12446 (2017).
- 553 21. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev.*  
554 *Cancer* **12**, 252–264 (2012).
- 555 22. Cho, B. C. et al. Amivantamab, an epidermal growth factor receptor (EGFR) and  
556 mesenchymal-epithelial transition factor (MET) bispecific antibody, designed to enable  
557 multiple mechanisms of action and broad clinical applications. *Clin. Lung Cancer* **24**, 89–  
558 97 (2023).
- 559 23. Ahn, M.-J. et al. Tarlatamab for Patients with Previously Treated Small-Cell Lung Cancer.  
560 *N. Engl. J. Med.* **389**, 2063–2075 (2023).
- 561 24. Cortés, J. et al. Trastuzumab Deruxtecan versus Trastuzumab Emtansine for Breast Cancer.  
562 *N. Engl. J. Med.* **386**, 1143–1154 (2022).
- 563 25. Aeffner, F. et al. The gold standard paradox in digital image analysis: manual versus  
564 automated scoring as ground truth. *Arch. Pathol. Lab. Med.* **141**, 1267–1275 (2017).
- 565 26. Sharma, P. & Allison, J. P. The future of immune checkpoint therapy. *Science* **348**, 56–61  
566 (2015).
- 567 27. Vu, T. & Claret, F. X. Trastuzumab: updated mechanisms of action and resistance in breast  
568 cancer. *Front. Oncol.* **2**, 62 (2012).
- 569 28. Sorensen, S. F. et al. PD-L1 expression and survival among patients with advanced non–  
570 small cell lung cancer treated with chemotherapy. *Transl. Oncol.* **9**, 64–69 (2016).
- 571 29. Xu, Y. et al. The association of PD-L1 expression with the efficacy of anti-PD-1/PD-L1  
572 immunotherapy and survival of non-small cell lung cancer patients: a meta-analysis of  
573 randomized controlled trials. *Transl. Lung Cancer Res.* **8**, 413 (2019).

- 574 30. Brunnström, H. et al. PD-L1 immunohistochemistry in clinical diagnostics of lung cancer:  
575 inter-pathologist variability is higher than assay variability. *Mod. Pathol.* **30**, 1411–1421  
576 (2017).
- 577 31. Robert, M. E. et al. High interobserver variability among pathologists using combined  
578 positive score to evaluate PD-L1 expression in gastric, gastroesophageal junction, and  
579 esophageal adenocarcinoma. *Mod. Pathol.* **36**, 100154 (2023).
- 580 32. Jonmarker Jaraj, S. et al. Intra- and interobserver reproducibility of interpretation of  
581 immunohistochemical stains of prostate cancer. *Virchows Arch.* **455**, 375–381 (2009).
- 582 33. Hoda, R. S. et al. Interobserver variation of PD-L1 SP142 immunohistochemistry  
583 interpretation in breast carcinoma: a study of 79 cases using whole slide imaging. *Arch.*  
584 *Pathol. Lab. Med.* **145**, 1132–1137 (2021).
- 585 34. Van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the  
586 clinic. *Nat. Med.* **27**, 775–784 (2021).
- 587 35. Rakha, E. A., Vougas, K. & Tan, P. H. Digital technology in diagnostic breast pathology  
588 and immunohistochemistry. *Pathobiology* **89**, 334–342 (2022).
- 589 36. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. Revisiting unreasonable effectiveness of  
590 data in deep learning era. *Proceedings of the IEEE International Conference on Computer*  
591 *Vision* 843–852 (2017).
- 592 37. Mahajan, D. et al. Exploring the limits of weakly supervised pretraining. *Proceedings of the*  
593 *European Conference on Computer Vision (ECCV)* 181–196 (2018)
- 594 38. Kann, B. H., Hosny, A. & Aerts, H. J. Artificial intelligence for clinical oncology. *Cancer*  
595 *Cell* **39**, 916–927 (2021).
- 596 39. Tizhoosh, H. R. & Pantanowitz, L. Artificial intelligence and digital pathology: challenges

- 597 and opportunities. *J. Pathol. Inform.* **9**, 38 (2018).
- 598 40. Park, Y. S. et al. FGFR2 assessment in gastric cancer using quantitative real-time  
599 polymerase chain reaction, fluorescent in situ hybridization, and immunohistochemistry.  
600 *Am. J. Clin. Pathol.* **143**, 865–872 (2015).
- 601 41. Schrupf, T., Behrens, H.-M., Haag, J., Krüger, S. & Röcken, C. FGFR2 overexpression  
602 and compromised survival in diffuse-type gastric cancer in a large central European cohort.  
603 *PLoS One* **17**, e0264011 (2022).
- 604 42. Zha, D. et al. Data-centric Artificial Intelligence: A Survey. *arXiv* preprint  
605 arXiv:2303.10158 (2023).
- 606 43. Chen, R. J. et al. Towards a general-purpose foundation model for computational  
607 pathology. *Nat. Med.* **30**, 850-862 (2024).
- 608 44. Gerardin, Y. et al. Foundation AI models predict molecular measurements of tumor purity.  
609 *Cancer Res.* **84**, 7402–7402 (2024).
- 610 45. Campanella, G., Vanderbilt, C. & Fuchs, T. Computational Pathology at Health System  
611 Scale–Self-Supervised Foundation Models from Billions of Images *AAAI 2024 Spring*  
612 *Symposium on Clinical Foundation Models* (2024).
- 613 46. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**,  
614 259–265 (2023).
- 615 47. Awad, M. M. et al. MET exon 14 mutations in non–small-cell lung cancer are associated  
616 with advanced age and stage-dependent MET genomic amplification and c-Met  
617 overexpression. *J. Clin. Oncol.* **34**, 721-730 (2016).
- 618 48. Tong, J. H. et al. MET amplification and exon 14 splice site mutation define unique  
619 molecular subgroups of non–small cell lung carcinoma with poor prognosis. *Clin. Cancer*

- 620        *Res.* **22**, 3048–3056 (2016).
- 621    49. Davies, K. D., Ritterhouse, L. L., Snow, A. N. & Sidiropoulos, N. MET exon 14 skipping  
622        mutations: essential considerations for current management of non–small-cell lung cancer.  
623        *J. Mol. Diagn.* **24**, 841–843 (2022).
- 624    50. Ha, S. Y. et al. MET overexpression assessed by new interpretation method predicts gene  
625        amplification and poor survival in advanced gastric carcinomas. *Mod. Pathol.* **26**, 1632–  
626        1641 (2013).
- 627    51. Ivanova, M. et al. Standardized pathology report for HER2 testing in compliance with 2023  
628        ASCO/CAP updates and 2023 ESMO consensus statements on HER2-low breast cancer.  
629        *Virchows Arch.* **484**, 3–14 (2024).
- 630    52. Ke, H.-L. et al. High Ubiquitin-Specific Protease 2a Expression Level Predicts Poor  
631        Prognosis in Upper Tract Urothelial Carcinoma. *Appl. Immunohistochem. Mol. Morphol.*  
632        **30**, 304–310 (2022).
- 633    53. Chu, P.-Y., Tzeng, Y.-D. T., Tsui, K.-H., Chu, C.-Y. & Li, C.-J. Downregulation of ATP  
634        binding cassette subfamily a member 10 acts as a prognostic factor associated with immune  
635        infiltration in breast cancer. *Aging* **14**, 2252 (2022).
- 636    54. Choi, K. M. et al. The interferon-inducible protein viperin controls cancer metabolic  
637        reprogramming to enhance cancer progression. *J. Clin. Invest.* **132**, e157302 (2022).
- 638    55. Tzeng, Y. T. et al. Integrated analysis of pivotal biomarker of LSM1 , immune cell  
639        infiltration and therapeutic drugs in breast cancer. *J. Cell. Mol. Med.* **26**, 4007–4020 (2022).
- 640    56. Wolff, A. C. et al. Human epidermal growth factor receptor 2 testing in breast cancer:  
641        ASCO–College of American Pathologists Guideline Update. *J. Clin. Oncol.* **41**, 3867–3872  
642        (2023).

- 643 57. Paver, E. C. et al. Programmed death ligand-1 (PD-L1) as a predictive marker for  
644 immunotherapy in solid tumours: a guide to immunohistochemistry implementation and  
645 interpretation. *Pathology (Phila.)* **53**, 141–156 (2021).
- 646 58. Ciga, O. et al. Overcoming the limitations of patch-based learning to detect cancer in whole  
647 slide images. *Sci. Rep.* **11**, 8894 (2021).
- 648 59. Saito, Y. et al. Inter-tumor heterogeneity of PD-L1 expression in non-small cell lung  
649 cancer. *J. Thorac. Dis.* **11**, 4982 (2019).
- 650 60. Reck, M. et al. Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell  
651 Lung Cancer. *N. Engl. J. Med.* **375**, 1823–1833 (2016).
- 652 61. Ryu, J. et al. OCELOT: Overlapped Cell on Tissue Dataset for Histopathology.  
653 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*  
654 23902–23912 (2023).
- 655 62. Van der Walt, S. et al. scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
- 656 63. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous  
657 separable convolution for semantic image segmentation. *Proceedings of the European*  
658 *Conference on Computer Vision (ECCV)* 801–818 (2018).
- 659 64. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition.  
660 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–  
661 778 (2016).
- 662 65. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level  
663 performance on imagenet classification *Proceedings of the IEEE International Conference*  
664 *on Computer Vision* 1026–1034 (2015).
- 665 66. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv*



- 666 arXiv:1412.6980 (2017).
- 667 67. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Jorge Cardoso, M. Generalised Dice  
668 Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. *Deep*  
669 *Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision*  
670 *Support* 240–248 (2017).
- 671 68. Loshchilov, I. & Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv*  
672 arXiv:1608.03983 (2017).
- 673 69. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and  
674 Projection for Dimension Reduction. *arXiv* arXiv:1802.03426 (2020).
- 675 70. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: Self-supervised  
676 learning via redundancy reduction. *International Conference on Machine Learning* (PMLR)  
677 12310–12320 (2021).
- 678 71. Dolezal, J. M. et al. Slideflow: Deep Learning for Digital Histopathology with Real-Time  
679 Whole-Slide Visualization. *arXiv* arXiv:2304.04142 (2023).
- 680
- 681

682 **Acknowledgments**

683 We appreciate Heeyeon Kay for the language editing. This work was supported by the National  
684 Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT)  
685 (2022R1C1C1007289), Republic of Korea, new faculty research fund of Ajou University School  
686 of Medicine, and Lunit.

687

688 **Author contributions**

689 B.B., M.M., S.P., S.S., D.Y., S.M.A., K.P, C-Y.O., S.I.C., and S.K. conceptualized the research.  
690 T.L., S.S., S.C., H.K., C-Y.O., S.I.C., and S.K. contributed to data acquisition. B.B. and M.M.  
691 conceived the experiments, B.B., M.M., J.R., and J.P. conducted the experiments. B.B., M.M.,  
692 J.R., S.P., J.P., S.P, D.Y., and K.P. developed algorithms. B.B., M.M, T.L, J.R., S.P., S.P., S.C.,  
693 H.K., S.I.C., and S.K. interpreted and validated the algorithms. B.B., M.M., S.I.C., S.M.A., and  
694 S.K. prepared an initial draft of the manuscript. All authors reviewed and approved the final  
695 version of the manuscript.

696

697 **Competing interests**

698 Biagio Brattoli, Mohammad Mostafavi, Taebum Lee, Jeongun Ryu, Seonwook Park, Jongchan  
699 Park, Sergio Pereira, Seunghwan Shin, Donggeun Yoo, Siraj M. Ali, Kyunghyun Paeng, Chan-  
700 Young Ock, and Soo Ick Cho are employees of Lunit and/or have stock/stock options in Lunit.

701

702

703 **Figure legend**

704 **Fig. 1. Overview of the universal immunohistochemistry (UIHC) artificial intelligence (AI)**  
705 **model development.** Single-cohort-derived models (SC-models) were trained using one dataset,  
706 while multiple-cohort-derived models (MC-models) were trained using multiple datasets,  
707 including lung, urothelial carcinoma, and breast cancer samples stained with Programmed  
708 Death-Ligand 1 (PD-L1) 22C3, as well as breast cancer samples stained with human epidermal  
709 growth factor receptor 2 (HER2). The AI models' performance was validated on both the  
710 training cohorts and novel cohorts that were not included in the training phase. These novel  
711 cohorts consisted of samples stained for human epidermal growth factor receptor 3 (HER3),  
712 MUC16, mesenchymal-epithelial transition factor (MET), trophoblast cell-surface antigen 2  
713 (TROP2), and fibroblast growth factor receptor 2 (FGFR2).

714  
715 **Fig. 2. Patch-level quantitative analysis of the artificial intelligence (AI) models. a** List of  
716 eight AI models trained on different cohort combinations. H-Br, HER2 of breast; P-L, PD-L1  
717 22C3 of lung; P-Br, PD-L1 22C3 of breast; P-LBIBr, PD-L1 22C3 of lung, bladder, and breast;  
718 PH-B, PD-L1 22C3 and HER2 of breast; PH-LBr, PD-L1 22C3 and HER2 of lung and breast;  
719 PH-LBIBr, PD-L1 22C3 and HER2 of lung, bladder, and breast. The different stain  
720 combinations (e.g. PD-L1 or HER2 is utilized or not), are visualized by color. **b-e** Performance  
721 of the eight models in training cohorts, where the stain type may be utilized during training – **b**  
722 PD-L1 22C3 in lung cancer, **c** PD-L1 22C3 in bladder cancer, **d** PD-L1 22C3 in breast cancer, **e**  
723 HER2 in breast cancer, **f** PD-L1 22C3 in pan-cancer. **g-n** Performance of the eight models in  
724 novel cohorts - **g** PD-L1 SP142, **h** Claudin 18.2, **i** DLL3, **j** FGFR2, **k** HER3, **l** MET, **m** MUC16,  
725 **n** TROP2 - where none of the test immunostain types has ever been utilized during the training

726 phase by any of the models. PD-L1, programmed death-ligand 1; HER2, human epidermal  
727 growth factor receptor 2; DLL3, delta-like 3; FGFR2, fibroblast growth factor receptor 2; HER3,  
728 human epidermal growth factor receptor 3; MET, mesenchymal-epithelial transition factor;  
729 TROP, trophoblast cell-surface antigen; mF1, mean F1 score.

730

731 **Fig. 3. Whole slide image (WSI)-level quantitative analysis of the artificial intelligence (AI)**  
732 **models.** The quantitative analysis is based on comparing the tumor proportion score (TPS) score  
733 in different training settings. The reported Cohen's Kappa scores are computed using the  
734 pathologists' labeled category as ground truth. **a** Macro-averaged Cohen's Kappa scores of the  
735 eight AI models over all the stains. **b** Cohen's Kappa scores of the AI models in PD-L1 22C3  
736 Lung dataset. **c** Cohen's Kappa scores of the AI models in PD-L1 22C3 Pan-cancer dataset. **d**  
737 Cohen's Kappa scores of the AI models in PD-L1 SP142 Lung dataset. **e** Cohen's Kappa scores  
738 of the AI models in multi-stain Pan-cancer dataset. The X-axis presents the summation of  
739 utilized stain types and the organ types of each cohort when training (e.g. PH-Br [PD-L1 22C3  
740 and HER2 of breast] is 3 as it has 2 stains and 1 cancer type). H-Br, HER2 of breast; P-L, PD-  
741 L1 22C3 of lung; P-Br, PD-L1 22C3 of breast; P-LBIBr, PD-L1 22C3 of lung, bladder, and  
742 breast; PH-B, PD-L1 22C3 and HER2 of breast; PH-LBr, PD-L1 22C3 and HER2 of lung and  
743 breast; PH-LBIBr, PD-L1 22C3 and HER2 of lung, bladder, and breast.

744

745 **Fig. 4. Performance analysis of the artificial intelligence (AI) models on whole slide image**  
746 **(WSI) categorized by tumor proportion score (TPS).** **a** Confusion matrices of multiple-  
747 cohort-derived models (P-LBIBr [PD-L1 22C3 of lung, bladder, and breast], PH-Br [PD-L1  
748 22C3 and HER2 of breast], PH-LBr [PD-L1 22C3 and HER2 of lung and breast], PH-LBIBr

749 [PD-L1 22C3 and HER2 of lung, bladder, and breast]). **b** Confusion matrices of single-cohort-  
750 derived models (H-Br [HER2 of breast], P-Br [PD-L1 22C3 of breast], P-L [PD-L1 22C3 of  
751 lung], and P-BI [PD-L1 22C3 of bladder]). 1% and 50% were utilized as TPS cutoffs.

752

753 **Fig. 5. Performance analysis of the artificial intelligence (AI) models on novel**

754 **immunostains with varying interpretation cutoffs. a** The receiver operating characteristic  
755 (ROC) curve by changing the cutoff over the predicted TPS and measuring false and true

756 positive rates. In this experiment, we fixed the ground truth TPS cutoff to 1% since it is the most

757 common and intuitive. **b** Comparing UIHC and single-cohort models across a range of 1% and

758 the second cutoff value within the [2%, 75%] range, illustrating the 3-way classification

759 accuracy. UIHC, universal immunohistochemistry model; H-Br, HER2 of breast; -Br, PD-L1

760 22C3 of breast; P-L, PD-L1 22C3 of lung; P-BI, PD-L1 22C3 of bladder. AVG, average.

761

762 **Fig. 6. Histopathologic validation of the universal immunohistochemistry (UIHC) model. a**

763 Lung cancer whole slide image (WSI) is stained with mesenchymal-epithelial transition factor

764 (MET). The UIHC model predicts more accurate classes unlike the P-L model which confuses

765 positively stained Tumor Cell (TC+) with negatively stained Tumor Cell (TC-). **b** Gastric cancer

766 WSI is stained with fibroblast growth factor receptor 2 (FGFR2). P-L, PD-L1 22C3 of lung; TC,

767 tumor cell.

768

769 **Fig. 7. Qualitative analysis of the artificial intelligence (AI)-learned representation. a** Two-

770 dimensional (2D) projection of internal representation colored by tumor proportion score (TPS).

771 Each patch is encoded to a 2D plot using three representations: raw pixels, self-supervised

772 learning model (SSL), and the universal immunohistochemistry (UIHC) model. Each dot  
773 represents one image patch from either an observed cohort available during training or from a  
774 novel cohort never seen by the UIHC model. The color represents the TPS within the patch. **b** A  
775 mosaic of image patches sorted by the internal representation. Using the same 2D representation  
776 as **a**, actual patches are displayed. **c** The assessment of cohort similarity through  $p$ -values. A  
777 higher  $p$ -value in UIHC signifies an inability to differentiate cohorts by UIHC, thus  
778 demonstrating the independence of UIHC from cohort effects.

779  
780 **Fig. 8. Data pipeline for Universal Immunohistochemistry (UIHC) artificial intelligence (AI)**  
781 **model. a** Example of annotation process; patches extracted from whole slide images (WSIs),  
782 then cells are manually annotated by expert pathologists. WSIs are split into  $0.04 \text{ mm}^2$  patches  
783 (resized to  $1024 \times 1024$  pixels at  $0.19$  microns-per pixel). **b** Patch-level annotation count by its  
784 positivity (negatively stained Tumor Cell [TC-] or positively stained Tumor Cell [TC+]). **c** The  
785 number of WSI in the WSI-level dataset only for testing.

786 **Table 1. Validation of universal immunohistochemistry (UIHC) model on cases with next-**  
 787 **generation sequencing results.**

Case no.	Organ	Group	Mutation/Amplification detail	UIHC TPS	Average UIHC TPS according to the group
1	Lung	EGFR exon20ins	p.Ala763_Tyr764insPheGlnGluAla	68.6	75.7±23.2
2	Lung	EGFR exon20ins	p.Ala767_Val769dup	85.7	
3	Lung	EGFR exon20ins	p.Asp770_Asn771insGly	88.4	
4	Lung	EGFR exon20ins	p.Ser768_Asp770dup	27.1	
5	Lung	EGFR exon20ins	p.His773_Val774insThrHis	80.0	
6	Lung	EGFR exon20ins	p.Pro772_His773insProAsnPro	98.0	
7	Lung	EGFR exon20ins	p.P772_H773dup	82.2	
8	Lung	MET exon 14 skipping	c.3082+2T>G	74.1	77.1±17.7
9	Lung	MET exon 14 skipping	c.2942-28_2944del	88.9	
10	Lung	MET exon 14 skipping	c.3025C>T	89.9	
11	Lung	MET exon 14 skipping	c.3082+1G>C	53.1	
12	Lung	MET exon 14 skipping	c.3082+2T>C	60.0	
13	Lung	MET exon 14 skipping	c.3082G>T	96.7	
14	Lymph node	MET amplification	8 copies	94.6	94.5±2.0
15	Lung	MET amplification	4 copies	92.5	

16	Lung	MET amplification	5 copies	96.5	
----	------	-------------------	----------	------	--

788 EGFR, epidermal growth factor receptor; MET, mesenchymal-epithelial transition

789 factor; TPS, tumor proportion score.

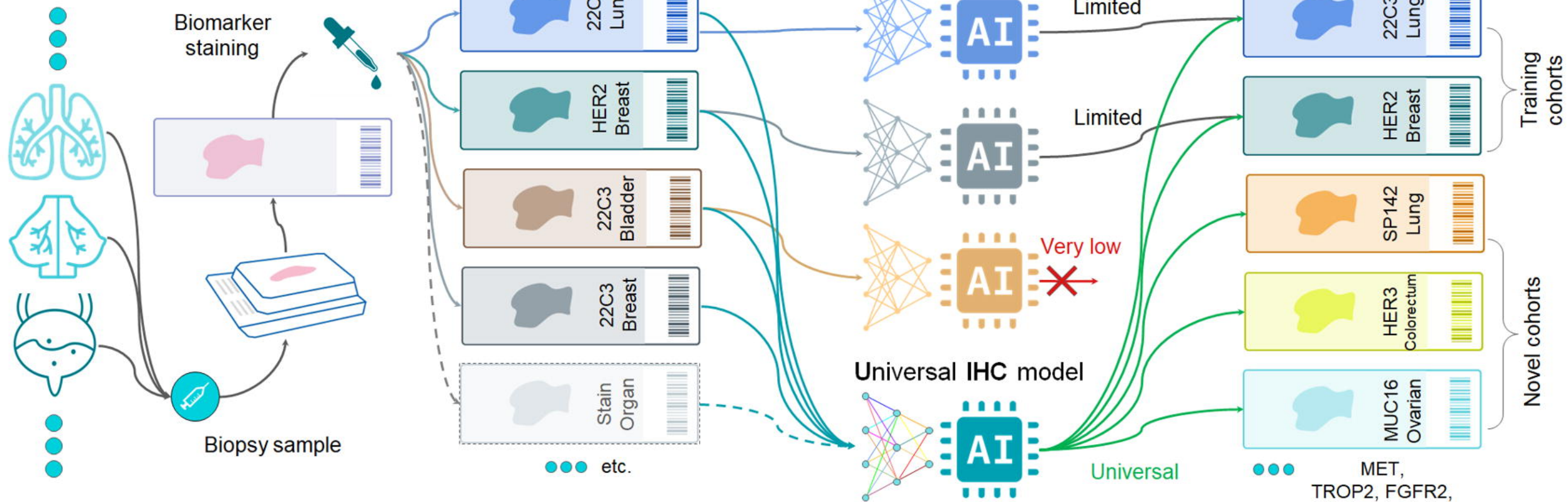


## Cancer types

## Stained WSI from various cohorts

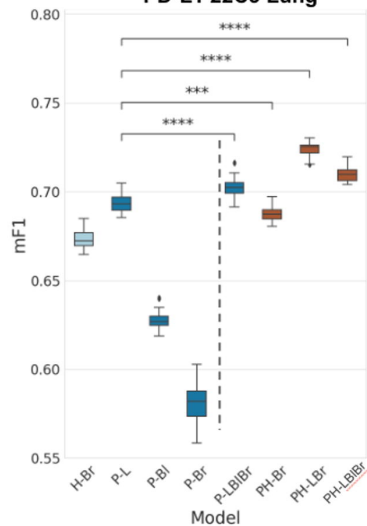
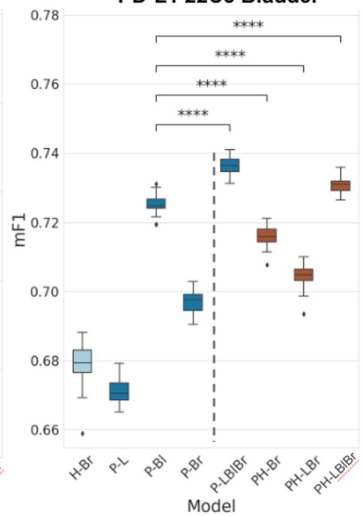
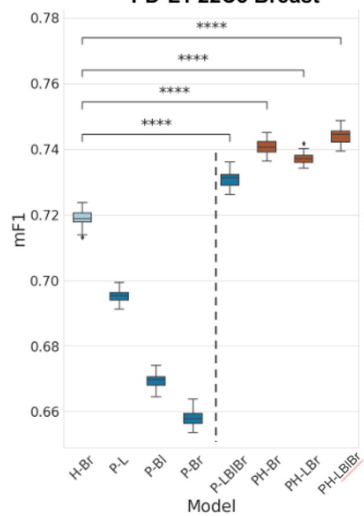
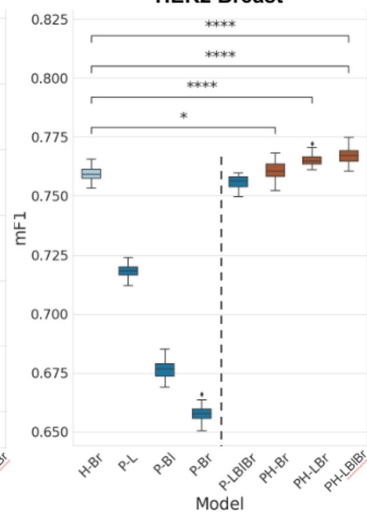
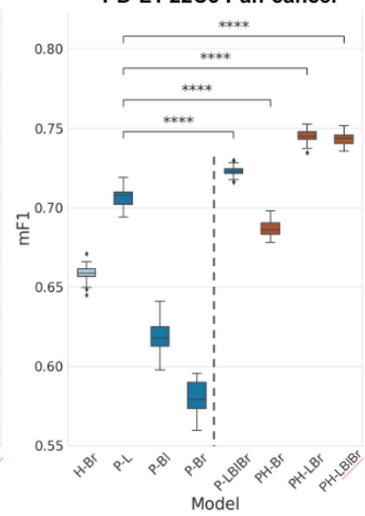
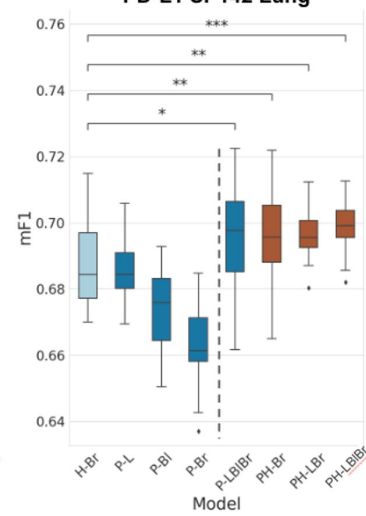
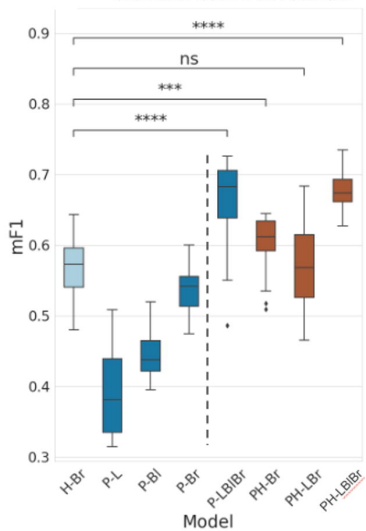
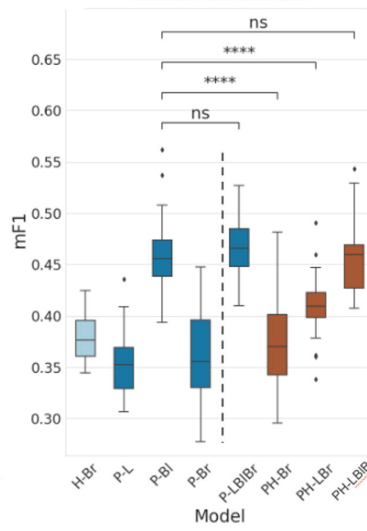
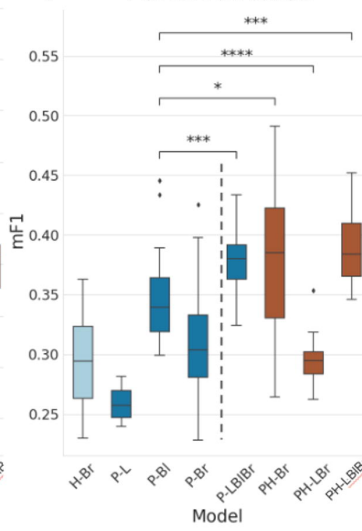
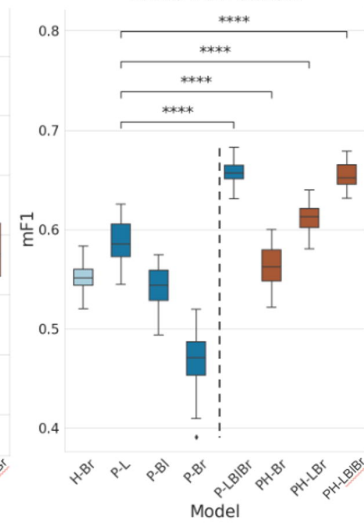
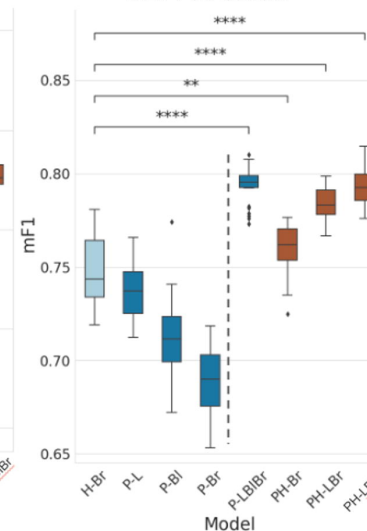
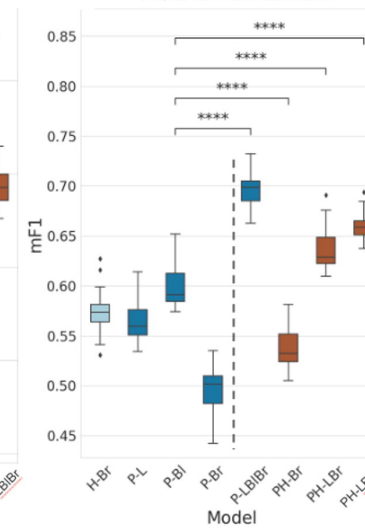
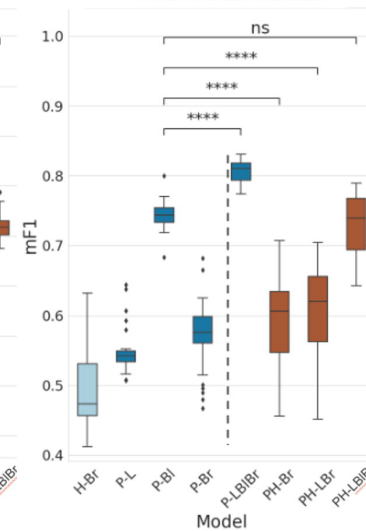
## Cohort specific AI models

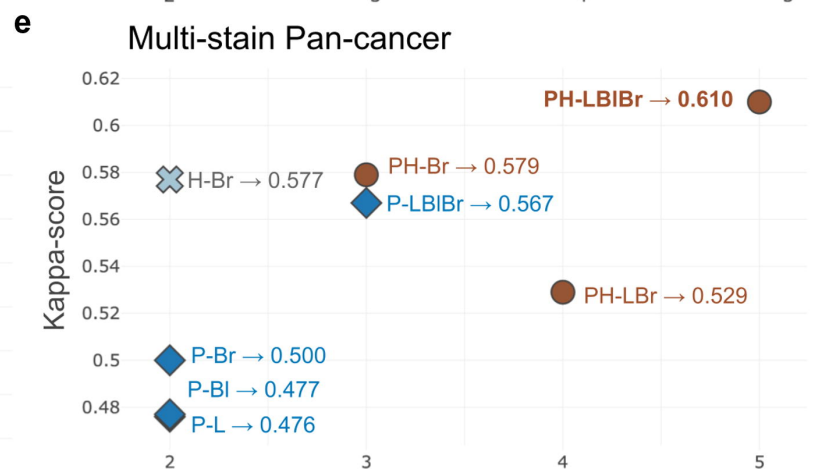
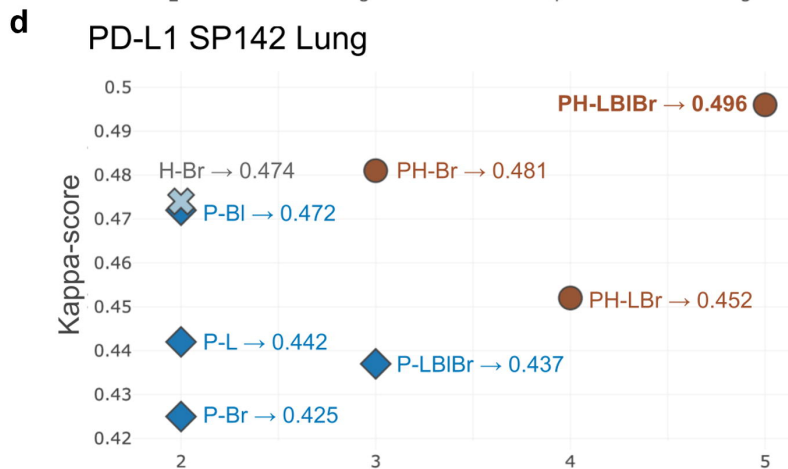
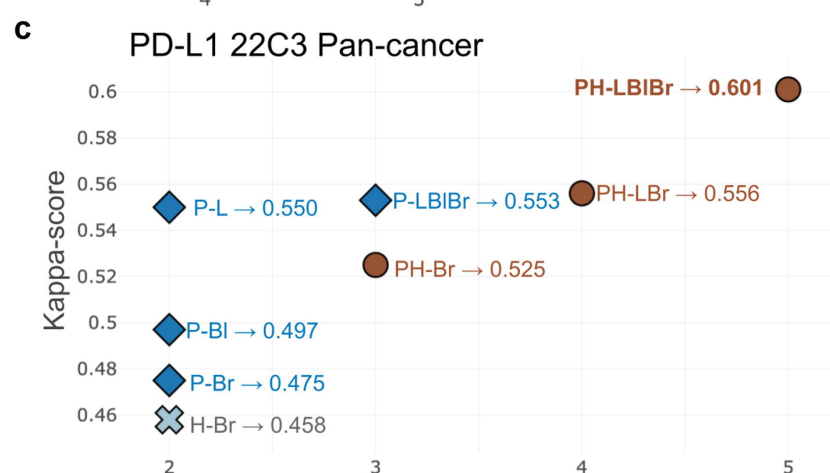
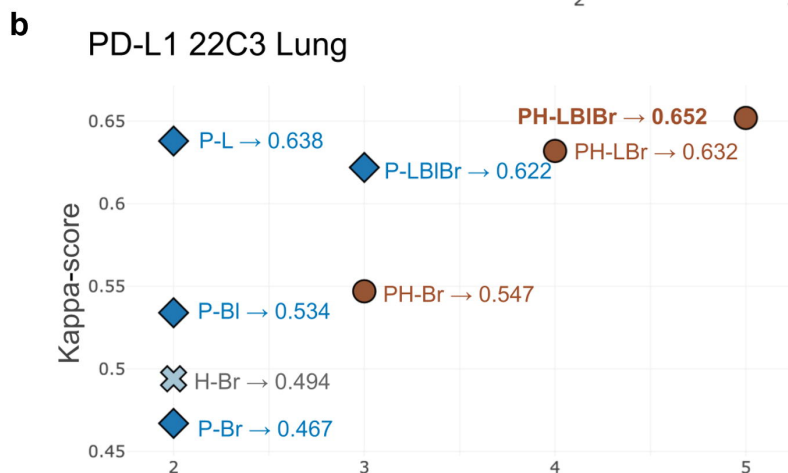
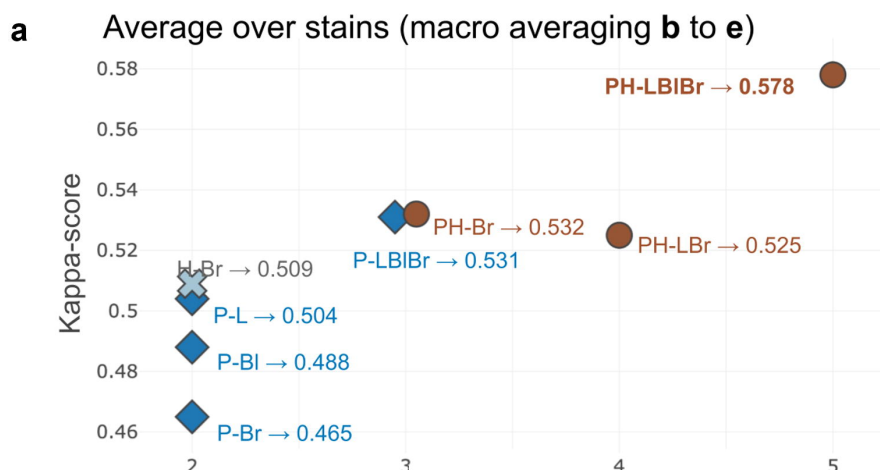
## Training and novel cohorts

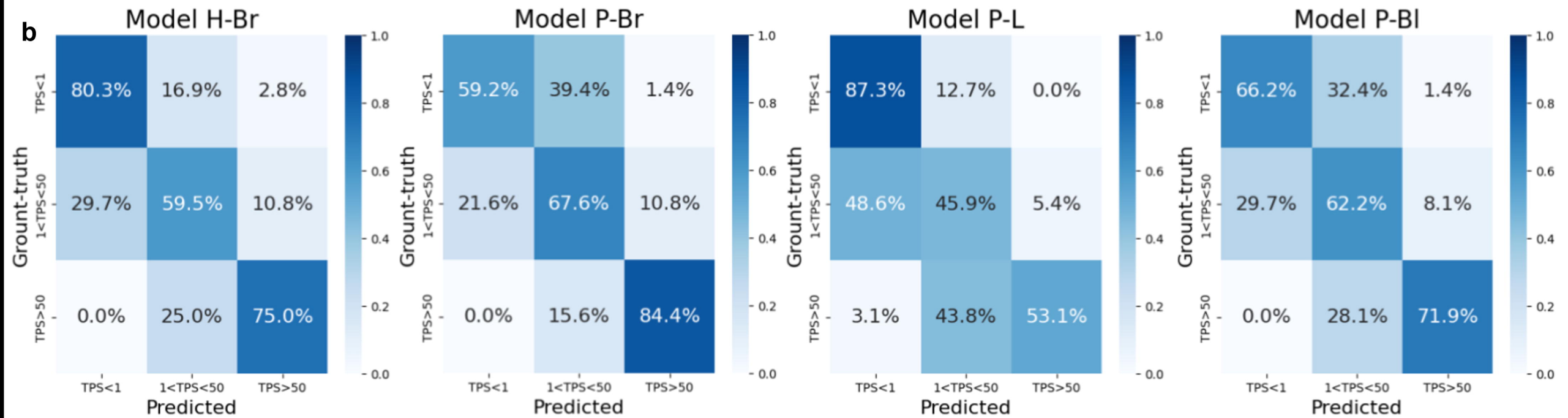
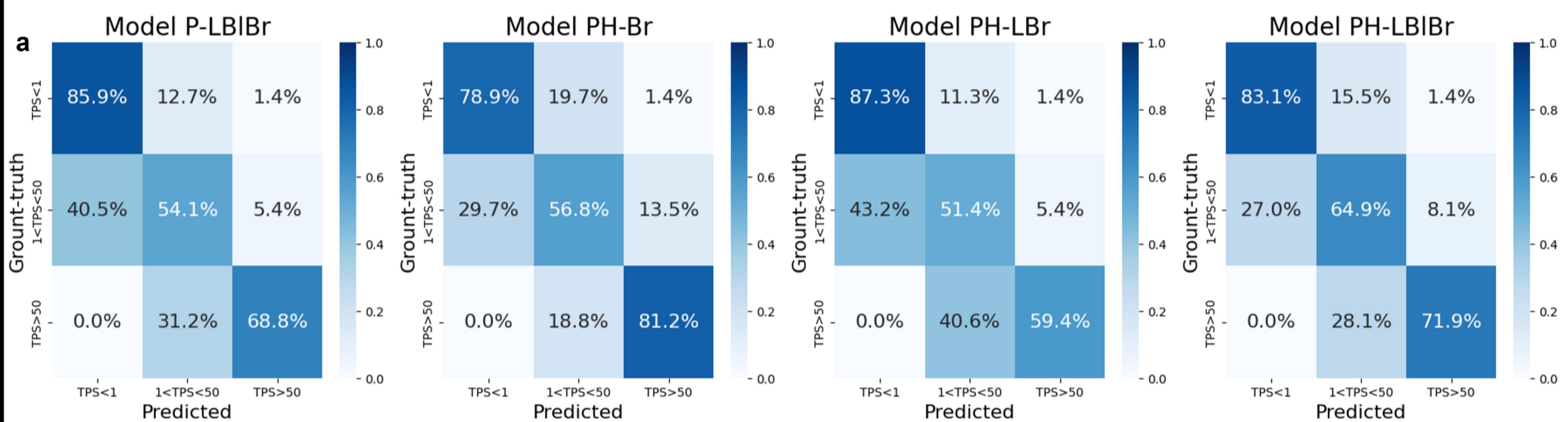


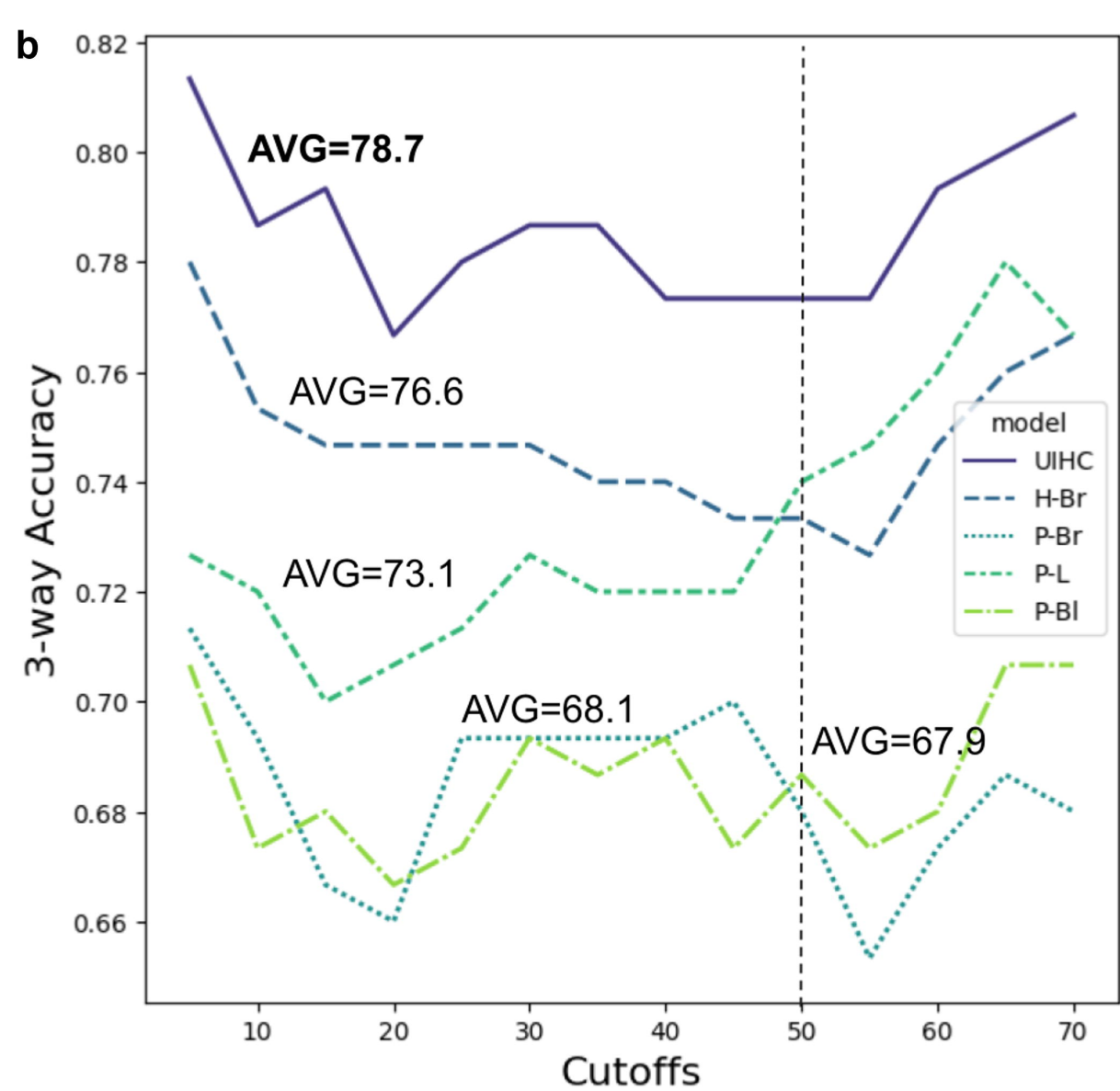
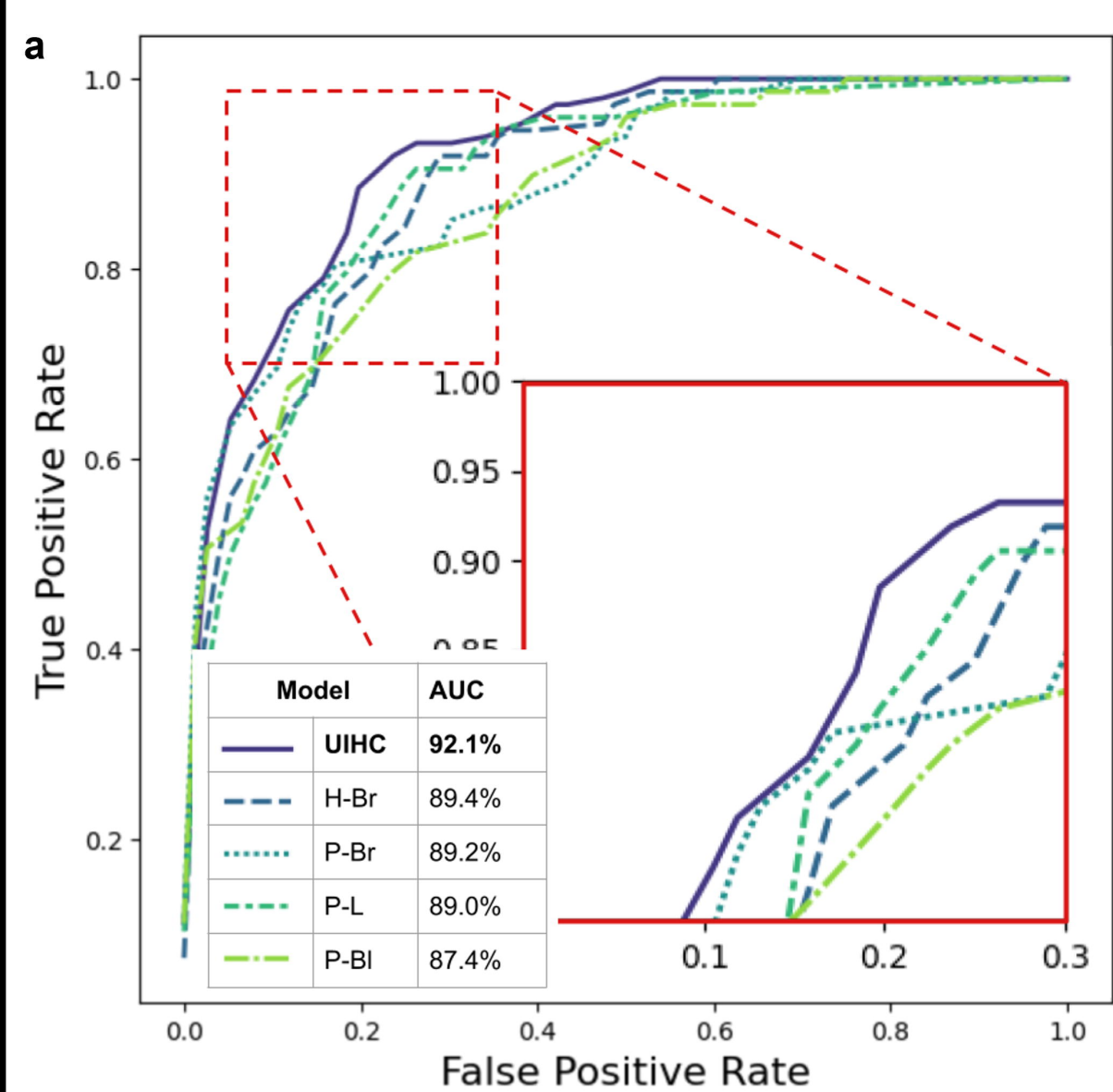
**a****Dataset combinations**

Model name	22C3 Lung	22C3 Bladder	22C3 Breast	HER2 Breast
H-Br				✓
P-L	✓			
P-BI		✓		
P-Br			✓	
P-L B Br	✓	✓	✓	
PH-Br			✓	✓
PH-LBr	✓		✓	✓
PH-L B Br	✓	✓	✓	✓

**b****PD-L1 22C3 Lung****c****PD-L1 22C3 Bladder****d****PD-L1 22C3 Breast****e****HER2 Breast****f****PD-L1 22C3 Pan-cancer****g****PD-L1 SP142 Lung****h****Claudin 18.2 Pan-cancer****i****DLL3 Pan-cancer****j****FGFR2 Pan-cancer****k****HER3 Pan-cancer****l****MET Pan-cancer****m****MUC16 Pan-cancer****n****TROP2 Pan-cancer**

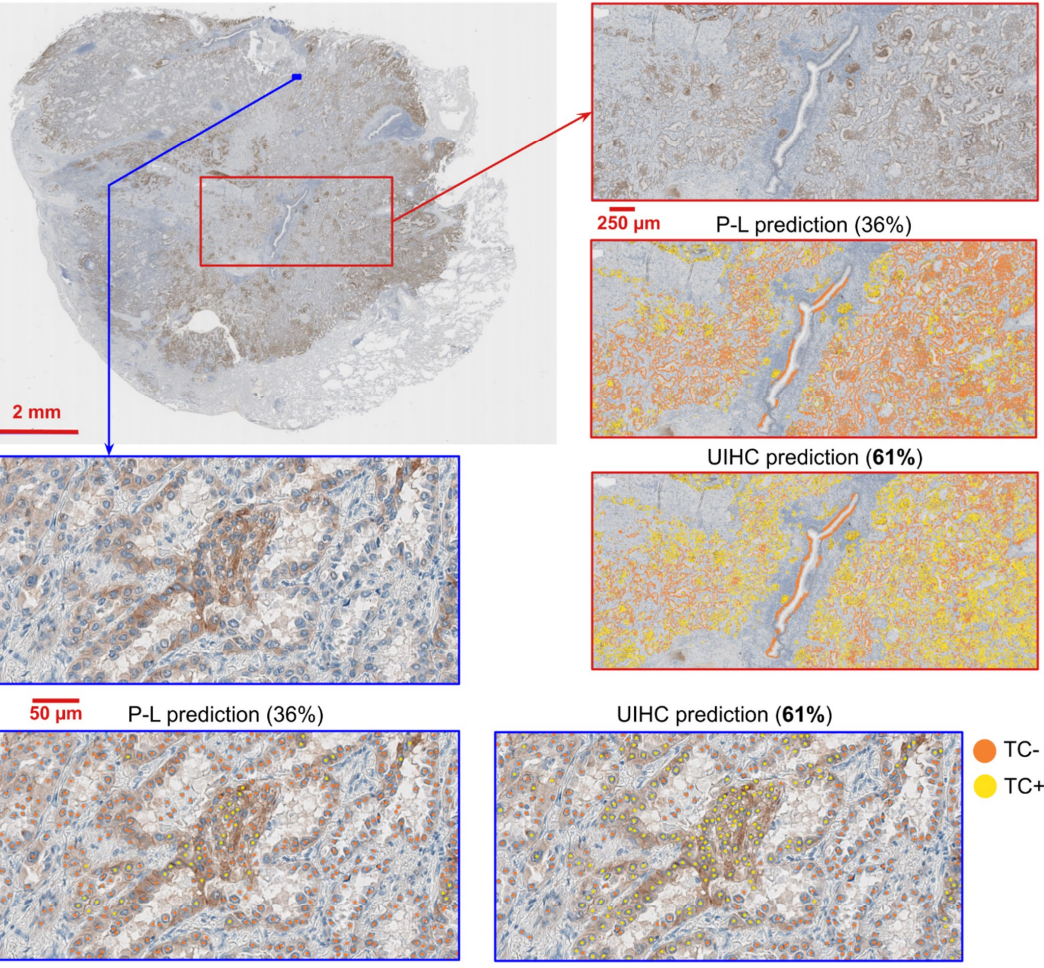






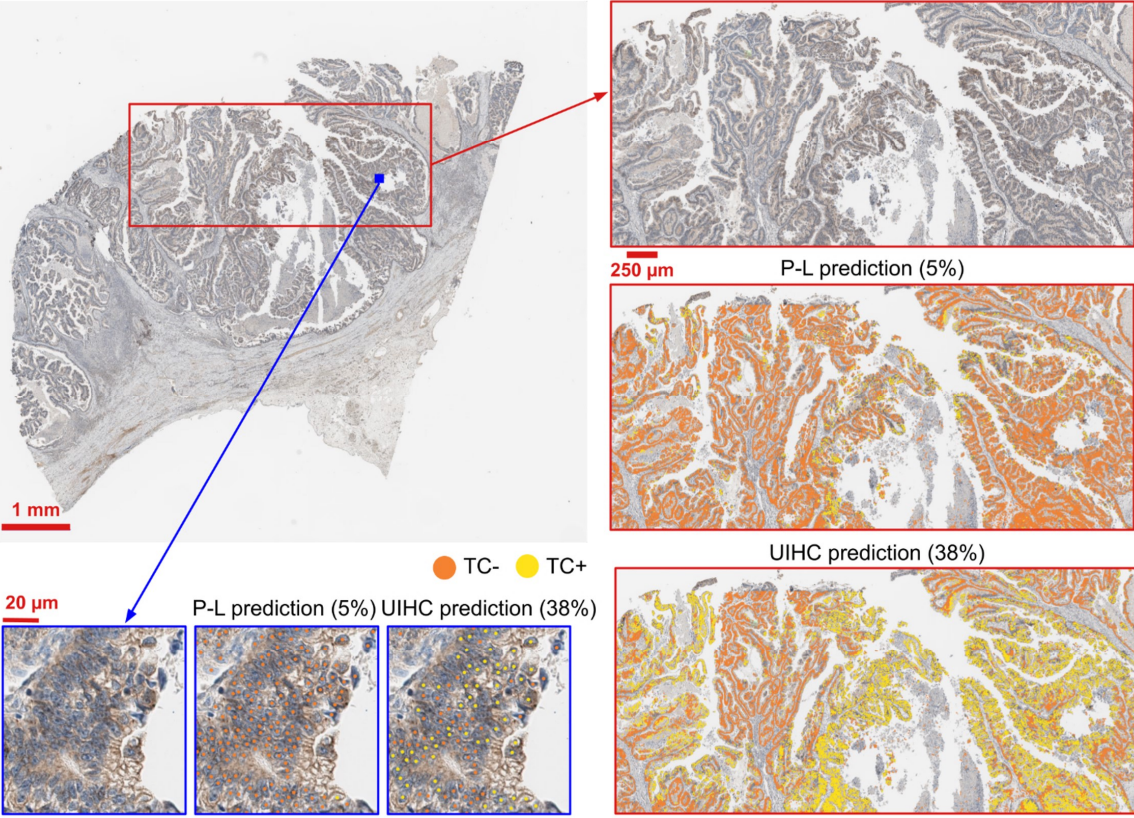
**a**

MET Lung



**b**

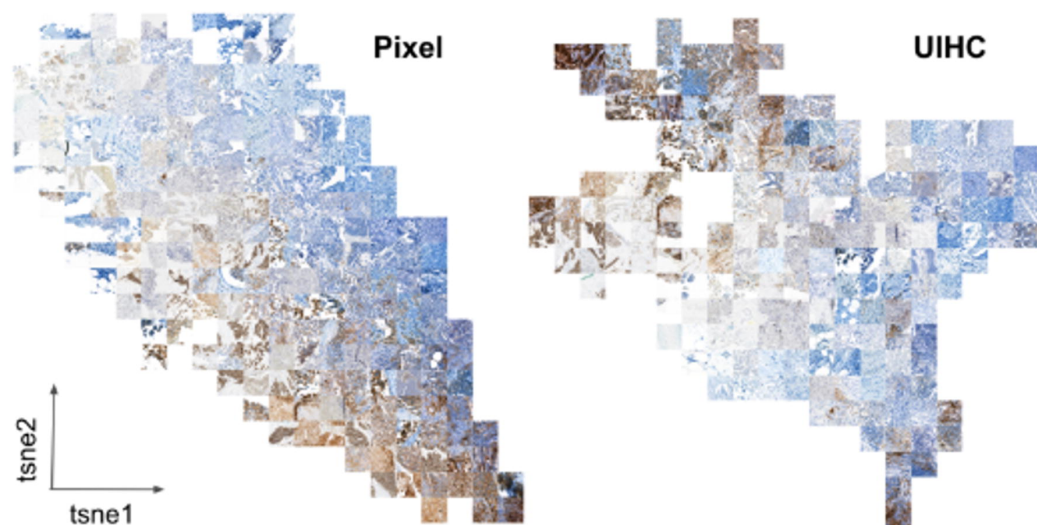
FGFR2 Stomach



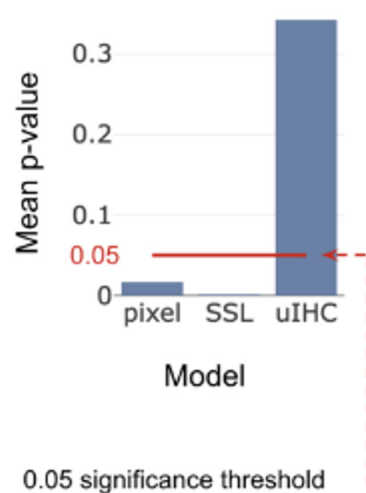
**a** 2D projection of internal representation grouped by TPS

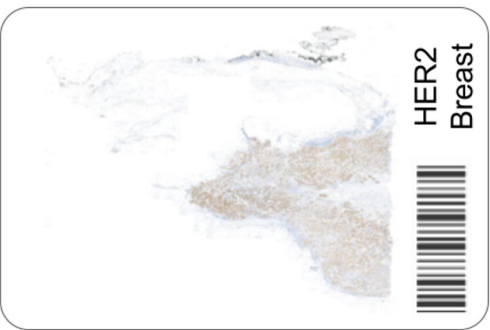
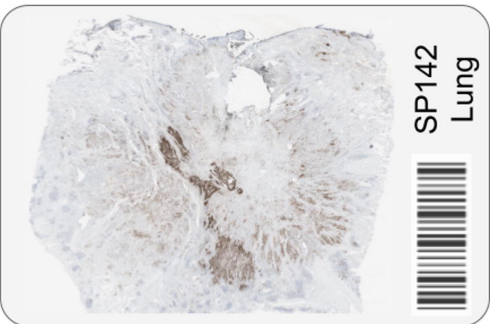
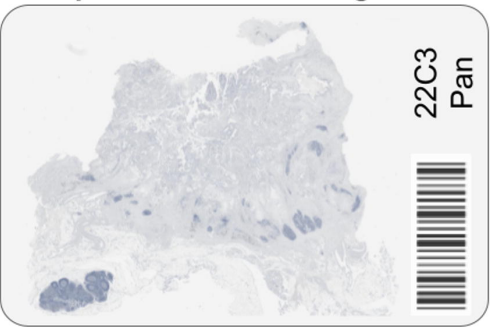
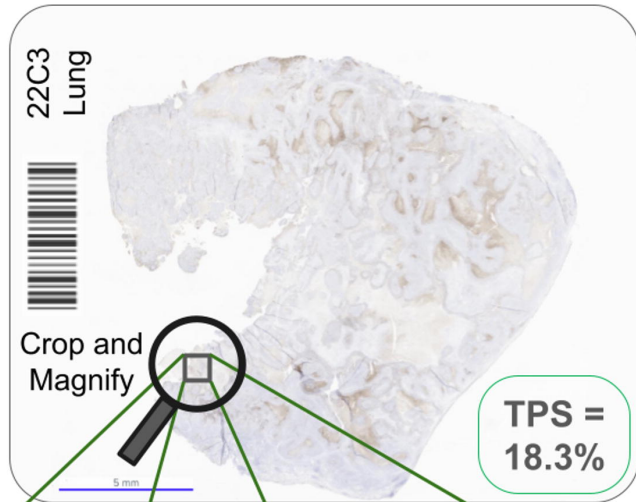
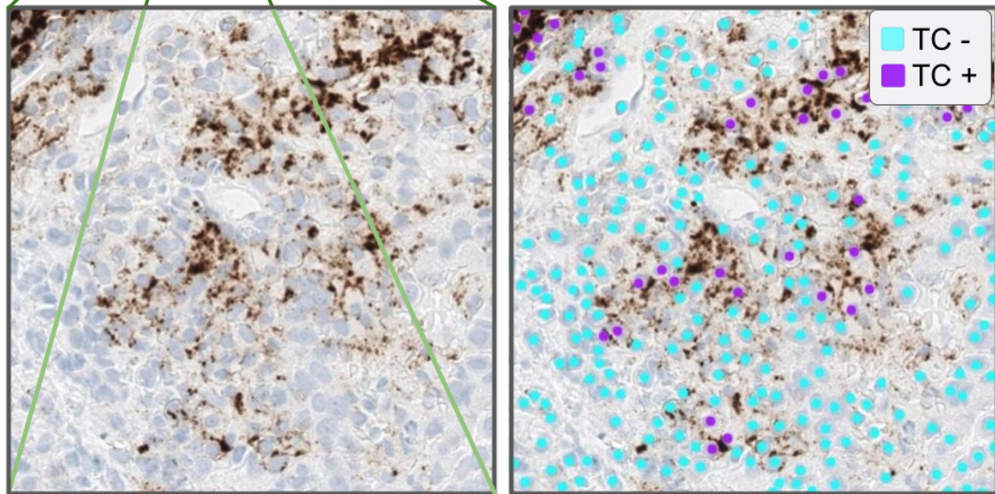
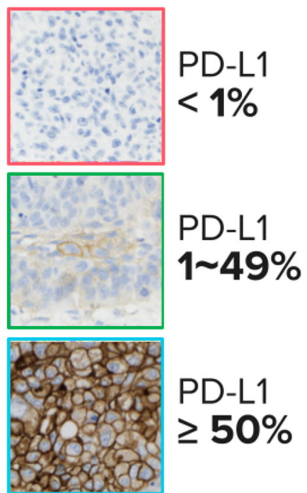
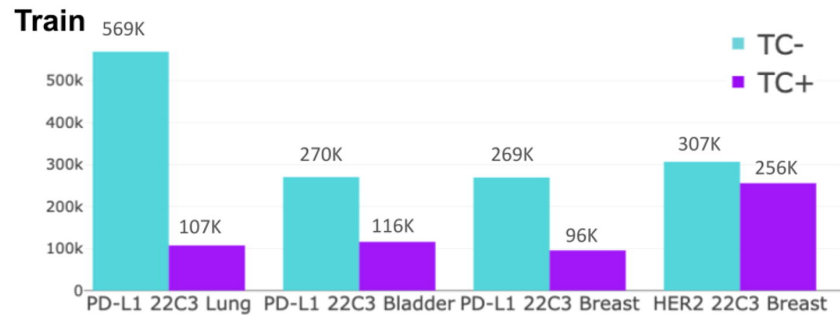
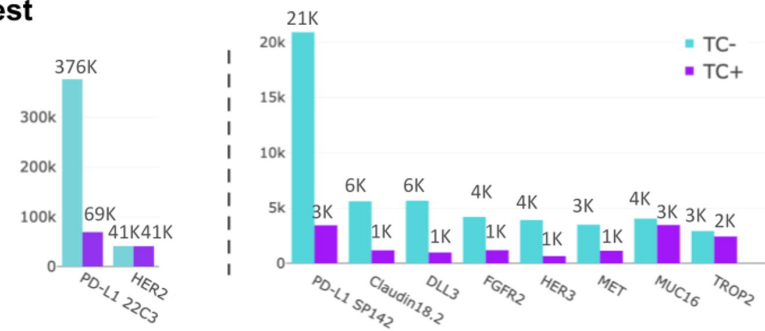
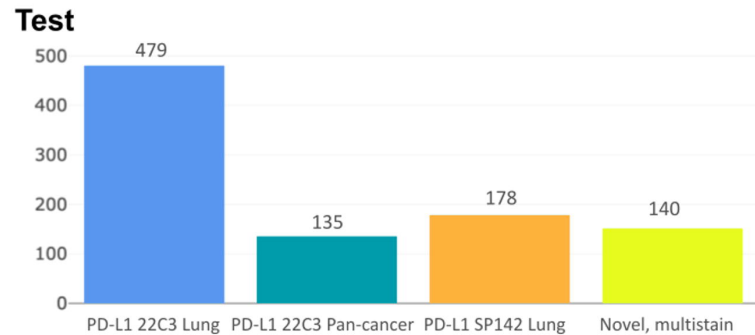


**b** Mosaic of image patches sorted by the internal representation

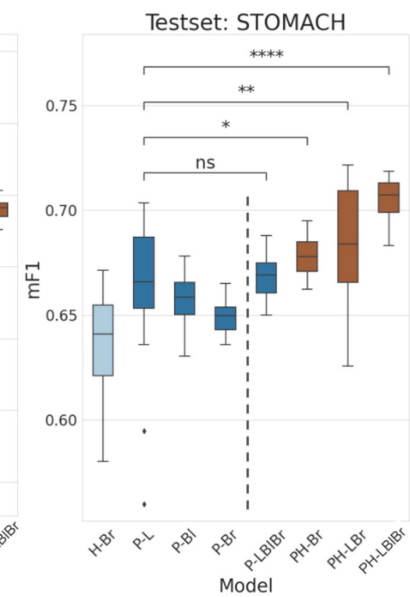
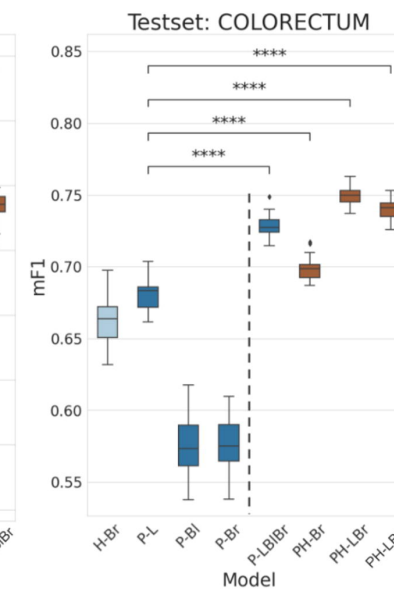
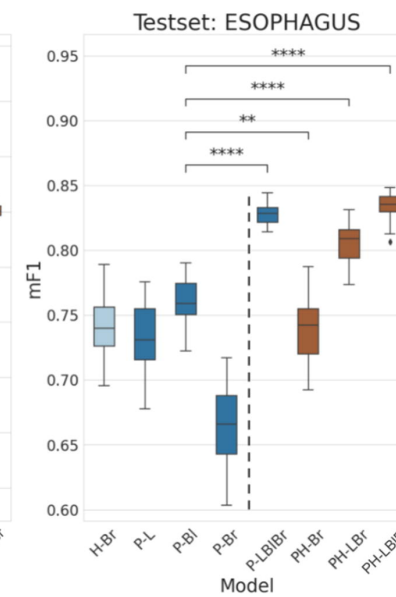
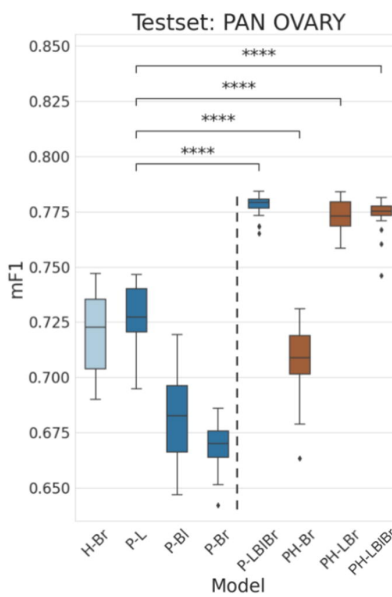
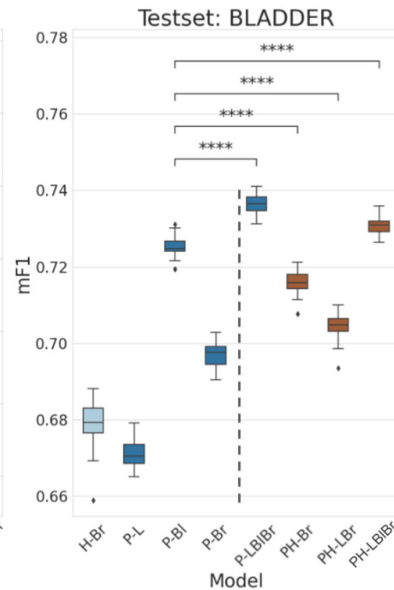
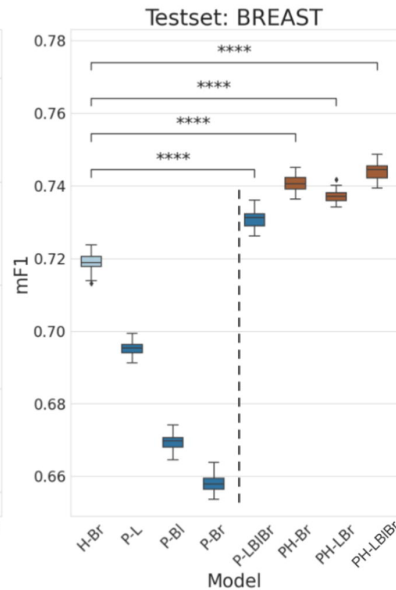
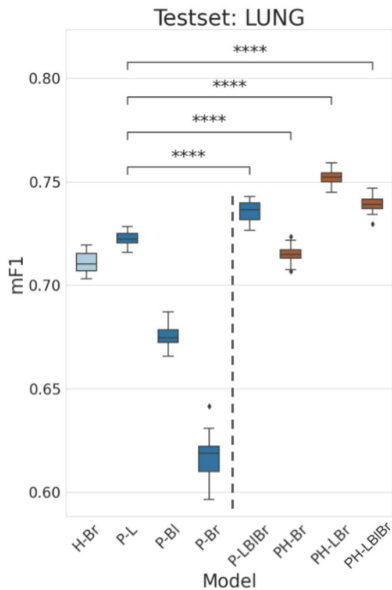


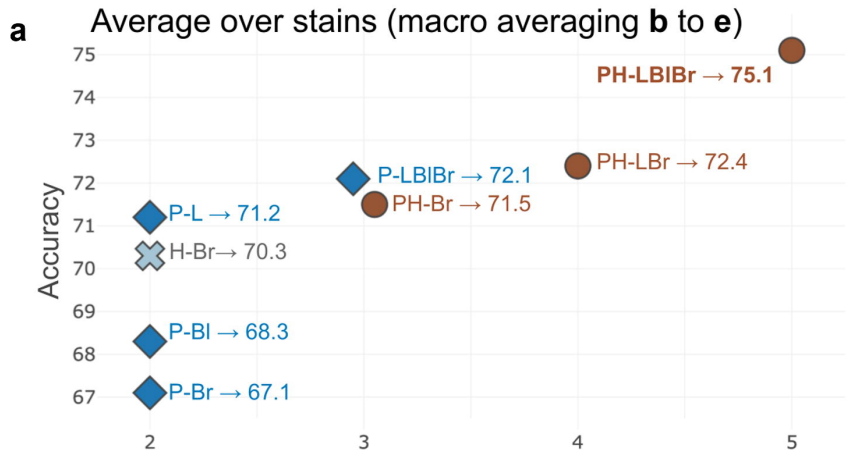
**c** Cohort similarity



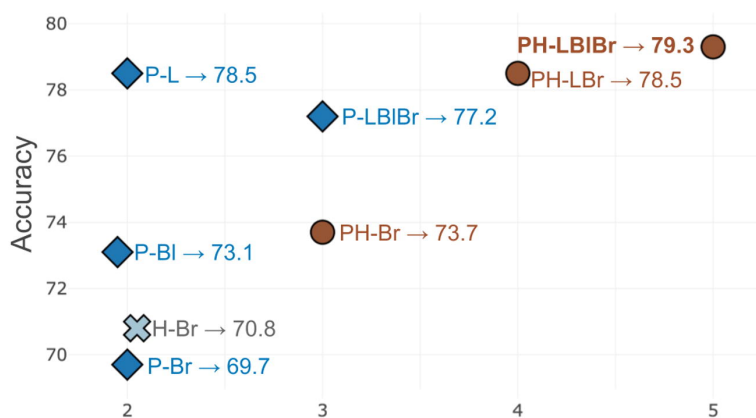
**a****Sample testset UIHC targets****WSI level annotation****Annotation guide****Patch-level cell annotations (1K×1K resolution)****b****Test****c WSI-level dataset only for test**



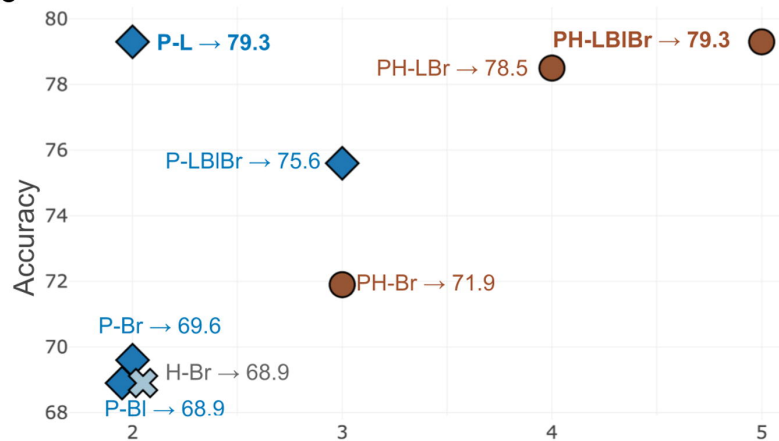




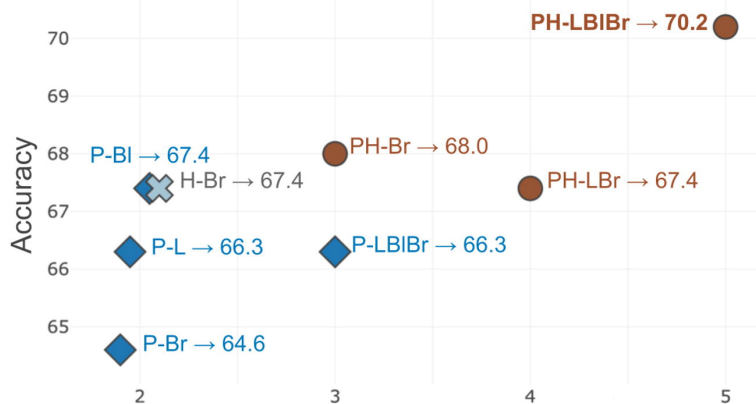
**b** PD-L1 22C3 Lung



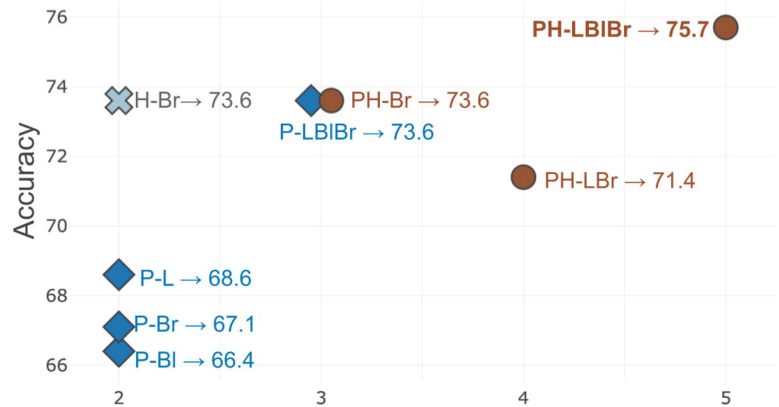
**c** PD-L1 22C3 Pan-cancer



**d** PD-L1 SP142 Lung

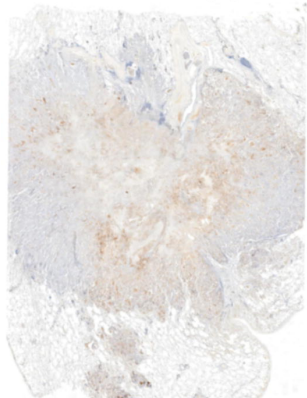


**e** Multi-stain Pan-cancer

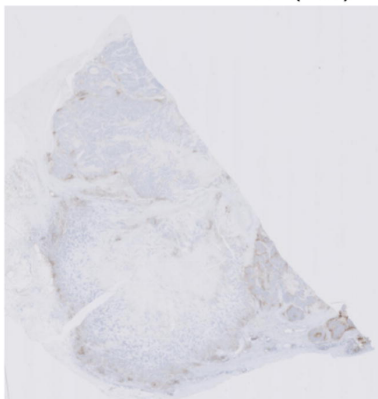


**a** IHCs samples from training cohorts (TPS %)

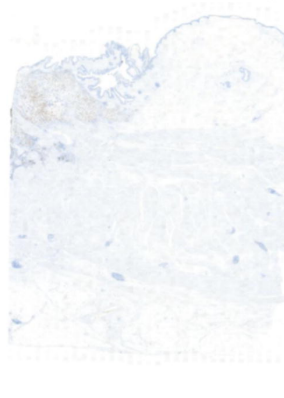
PD-L1 22C3 lung (40%)



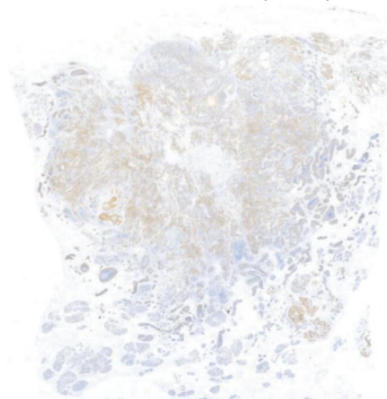
PD-L1 22C3 breast (0%)



PD-L1 22C3 bladder (0%)

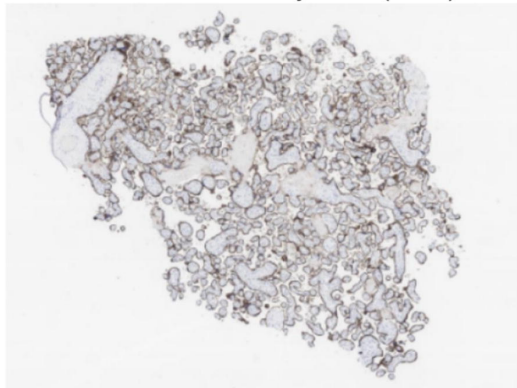


HER2 breast (90%)

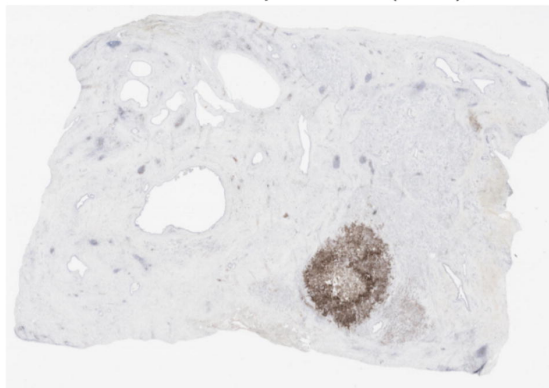


**b** IHCs samples from novel cohorts (TPS %)

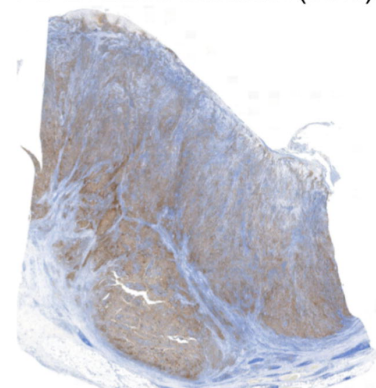
PD-L1 22C3 biliary tract (85%)



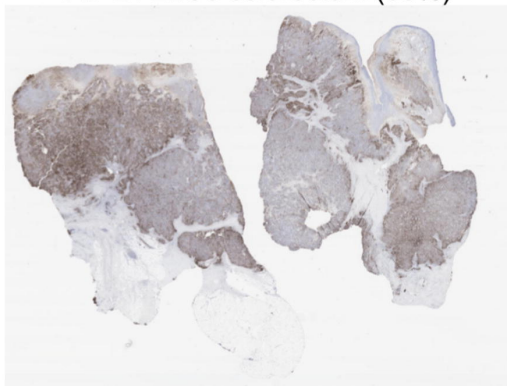
PD-L1 22C3 pancreas (70%)



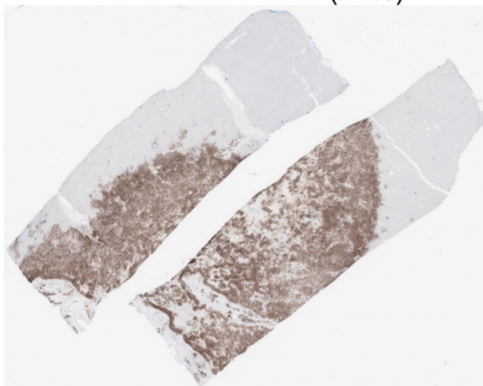
PD-L1 22C3 stomach (93%)



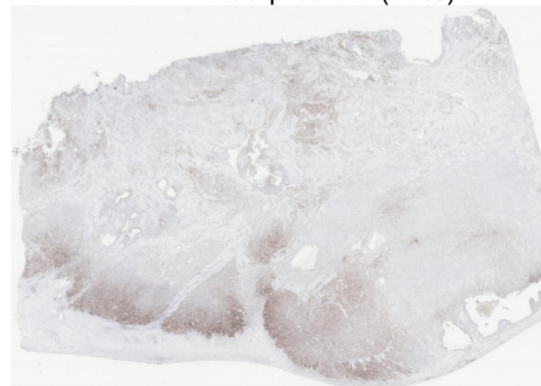
PD-L1 22C3 colorectum (68%)



PD-L1 22C3 Liver (93%)

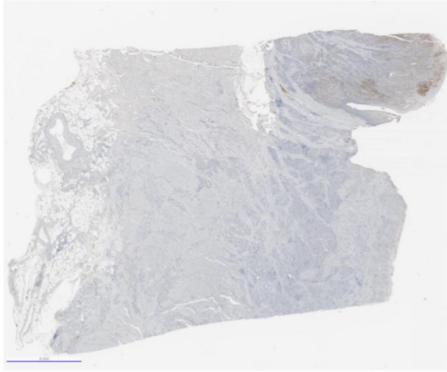


PD-L1 22C3 prostate (50%)

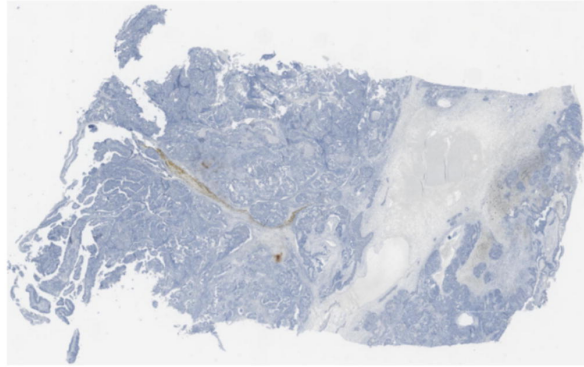


# Novel IHCs samples (TPS)

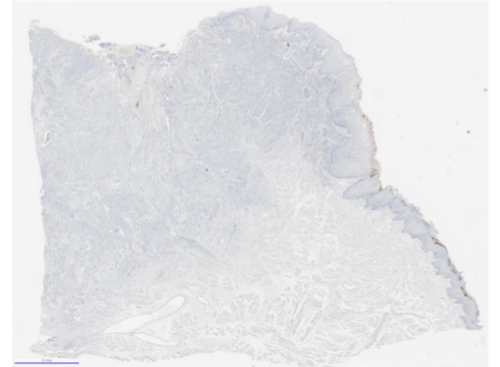
Claudin 18.2 stomach (13%)



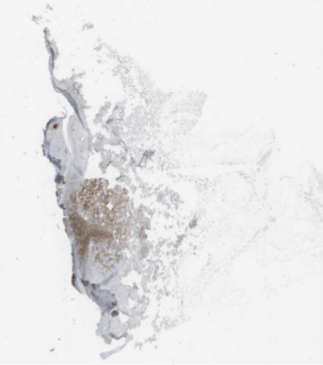
MUC16 cervix (0%)



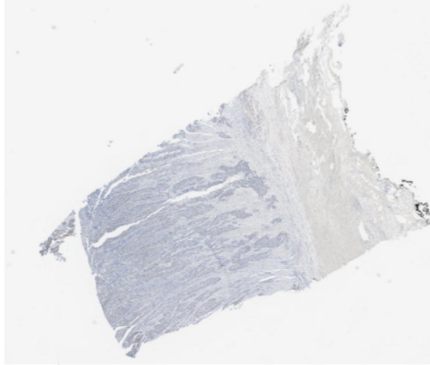
HER3 head & neck (0%)



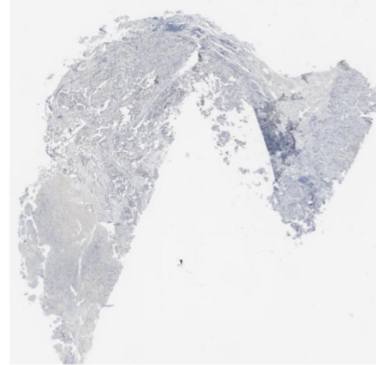
DLL3 melanoma (75%)



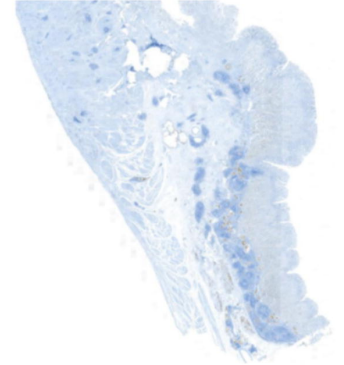
MET esophagus (4%)



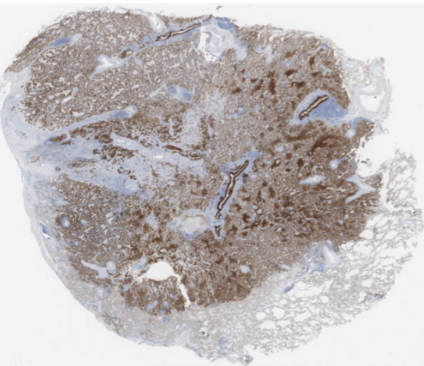
FGFR2 bladder (13%)



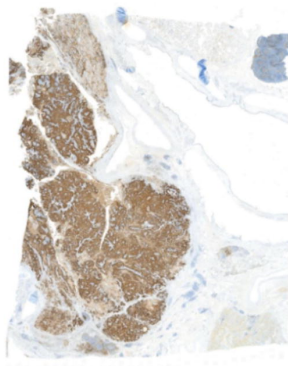
E-Cadherin colorectum (100%)



TROP2 lung (96%)



SP142 lung (100%)



SP263 lung (43%)

