Article

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

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

[+]These authors contributed equally to this work.

[*]Corresponding authors

[1]Lunit, Seoul, Republic of Korea

[2]Department of Pathology and Translational Genomics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[3]Department of Pathology, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

[4]Department of Pathology, Ajou University School of Medicine, Suwon, Republic of Korea

[5]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea

Correspondence to:

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

Department of Pathology, Ajou University School of Medicine.

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

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

Tel: +82-31-219-6460; Fax: 82-31-219-5934; E-mail: seokhwikim@ajou.ac.kr

ORCID ID: 0000-0001-7646-5064

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

Lunit.

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

Tel: +82-2-2138-0827; Fax: 82-2-6919-2702; E-mail: sooickcho@lunit.io

ORCID ID: 0000-0003-3414-9869

ABSTRACT

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

Introduction

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

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

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

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

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

Results

Patch-level tumor cell detection and IHC-positivity classification

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

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

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

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

outperformed the single best performing SC-model H-Br (P-LBlBr, 0.795 [0.773, 0.810], 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-LBlBr, 0.792 [0.776, 0.815], p<0.001) (Fig. 2l). This tendency for MC-models to outperform SC-models was also observed when the data was categorized by cancer type (lung, breast, bladder, pan-ovary, esophagus, colorectum, and stomach) rather than IHC type (Supplementary Fig. 1).

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

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

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

Performance analysis of UIHC on novel immunostains for different cutoffs

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

Histopathologic validation of inference examples of UIHC model

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

Interpreting the representations of UMAP learned by the UIHC model

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

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

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

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

Performance analysis of UIHC on the real-world dataset

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

cohorts of NSCLC cases known to harbor oncogenic driver alterations - MET exon 14 skipping mutations, MET amplifications, and epidermal growth factor receptor (EGFR) exon 20 insertions [18]. The UIHC model assigned higher tumor proportion scores (TPS) to the MET amplification group compared to the other groups (Table 1). The UIHC model yielded a MET TPS of 94.5±2.0 for the three MET amplification cases, 77.1±17.7 for the six exon 14 skipping mutation cases, and 75.7±23.2 for the seven EGFR exon 20 insertion EGFR cases.

Discussion

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

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

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

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

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

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

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

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

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

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

In conclusion, we have successfully developed a UIHC model capable of autonomously analyzing novel stains across diverse cancer types. In contrast to prevailing literature and existing image analysis products that often focus on specialized cohorts, our model's versatility and agility significantly enhance its potential in expediting research related to new IHC antibodies[34]. 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

Methods

Dataset preparation for AI model development

Histopathology dataset for annotation

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

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

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

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

Patch-level annotation for AI model development

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

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

WSI-level test sets for AI model performance validation

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

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

Supplementary Fig. 4 and 5.

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

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

AI Model development process

Development of Universal IHC algorithm

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

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

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

Label pre-processing

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

Inference post-processing

Given that we treat cell detection as a segmentation task, a post-processing phase is needed to extract 2D coordinates and classes of predicted cells from the probability map output by the network. We apply *skim-age.feature.peak_local_max* on the model's output, which finds the locations of local maximums of the probability map to get the set of predicted cell points[62]. Lastly, we obtain each cell's class and probability value in the cell segmentation map through *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[63,64].
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[65]. The model is optimized using the Adam optimizer[66]. Dice loss is used to train the
421 model[67]. The initial learning rate is set to $1e-4$, adapted using the cosine learning rate
422 scheduler[68]. 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 1024×1024 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.

AI Model evaluation

Metrics for AI model performance

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

Model interpretation by visualization of data distribution

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

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

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

2. We developed a second network for comparing two deep learning models. Since patches with a manual annotation are a small fraction of all slides, we trained a ResNet34 using a state-of-the-art SSL method called Barlow Twins instead[70]. This allowed us to train the model on a large number of histopathology patches from different types of stains (PD-L1 22C3, and HER2) and cancer types without the need for any manual annotations.
3. The best UIHC model is used as the representative UIHC model for the qualitative analysis.

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

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

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

Reporting summary

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

Data availability

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

Code availability

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

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

References

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

better prediction of immunotherapy response. *Eur. J. Cancer* **170**, 17–26 (2022).

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

20. Liu, C.-C., Yang, H., Zhang, R., Zhao, J.-J. & Hao, D.-J. Tumour-associated antigens and their anti-cancer applications. *Eur. J. Cancer Care (Engl.)* **26**, e12446 (2017).

21. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).

22. Cho, B. C. et al. Amivantamab, an epidermal growth factor receptor (EGFR) and mesenchymal-epithelial transition factor (MET) bispecific antibody, designed to enable multiple mechanisms of action and broad clinical applications. *Clin. Lung Cancer* **24**, 89–97 (2023).

23. Ahn, M.-J. et al. Tarlatamab for Patients with Previously Treated Small-Cell Lung Cancer. *N. Engl. J. Med.* **389**, 2063–2075 (2023).

24. Cortés, J. et al. Trastuzumab Deruxtecan versus Trastuzumab Emtansine for Breast Cancer. *N. Engl. J. Med.* **386**, 1143–1154 (2022).

25. Aeffner, F. et al. The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. *Arch. Pathol. Lab. Med.* **141**, 1267–1275 (2017).

26. Sharma, P. & Allison, J. P. The future of immune checkpoint therapy. *Science* **348**, 56–61 (2015).

27. Vu, T. & Claret, F. X. Trastuzumab: updated mechanisms of action and resistance in breast cancer. *Front. Oncol.* **2**, 62 (2012).

28. Sorensen, S. F. et al. PD-L1 expression and survival among patients with advanced non–small cell lung cancer treated with chemotherapy. *Transl. Oncol.* **9**, 64–69 (2016).

29. Xu, Y. et al. The association of PD-L1 expression with the efficacy of anti-PD-1/PD-L1 immunotherapy and survival of non-small cell lung cancer patients: a meta-analysis of randomized controlled trials. *Transl. Lung Cancer Res.* **8**, 413 (2019).

30. Brunnström, H. et al. PD-L1 immunohistochemistry in clinical diagnostics of lung cancer: inter-pathologist variability is higher than assay variability. *Mod. Pathol.* **30**, 1411–1421 (2017).

31. Robert, M. E. et al. High interobserver variability among pathologists using combined positive score to evaluate PD-L1 expression in gastric, gastroesophageal junction, and esophageal adenocarcinoma. *Mod. Pathol.* **36**, 100154 (2023).

32. Jonmarker Jaraj, S. et al. Intra- and interobserver reproducibility of interpretation of immunohistochemical stains of prostate cancer. *Virchows Arch.* **455**, 375–381 (2009).

33. Hoda, R. S. et al. Interobserver variation of PD-L1 SP142 immunohistochemistry interpretation in breast carcinoma: a study of 79 cases using whole slide imaging. *Arch. Pathol. Lab. Med.* **145**, 1132–1137 (2021).

34. Van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).

35. Rakha, E. A., Vougas, K. & Tan, P. H. Digital technology in diagnostic breast pathology and immunohistochemistry. *Pathobiology* **89**, 334–342 (2022).

36. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. *Proceedings of the IEEE International Conference on Computer Vision* 843-852 (2017).

37. Mahajan, D. et al. Exploring the limits of weakly supervised pretraining. *Proceedings of the European Conference on Computer Vision (ECCV)* 181–196 (2018)

38. Kann, B. H., Hosny, A. & Aerts, H. J. Artificial intelligence for clinical oncology. *Cancer Cell* **39**, 916–927 (2021).

39. Tizhoosh, H. R. & Pantanowitz, L. Artificial intelligence and digital pathology: challenges

and opportunities. *J. Pathol. Inform.* **9**, 38 (2018).

40. Park, Y. S. et al. FGFR2 assessment in gastric cancer using quantitative real-time polymerase chain reaction, fluorescent in situ hybridization, and immunohistochemistry. *Am. J. Clin. Pathol.* **143**, 865–872 (2015).

41. Schrumpf, T., Behrens, H.-M., Haag, J., Krüger, S. & Röcken, C. FGFR2 overexpression and compromised survival in diffuse-type gastric cancer in a large central European cohort. *PLoS One* **17**, e0264011 (2022).

42. Zha, D. et al. Data-centric Artificial Intelligence: A Survey. *arXiv* preprint arXiv:2303.10158 (2023).

43. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850-862 (2024).

44. Gerardin, Y. et al. Foundation AI models predict molecular measurements of tumor purity. *Cancer Res.* **84**, 7402–7402 (2024).

45. Campanella, G., Vanderbilt, C. & Fuchs, T. Computational Pathology at Health System Scale–Self-Supervised Foundation Models from Billions of Images *AAAI 2024 Spring Symposium on Clinical Foundation Models* (2024).

46. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).

47. Awad, M. M. et al. MET exon 14 mutations in non–small-cell lung cancer are associated with advanced age and stage-dependent MET genomic amplification and c-Met overexpression. *J. Clin. Oncol.* **34**, 721-730 (2016).

48. Tong, J. H. et al. MET amplification and exon 14 splice site mutation define unique molecular subgroups of non–small cell lung carcinoma with poor prognosis. *Clin. Cancer*

Res. **22**, 3048–3056 (2016).

49. Davies, K. D., Ritterhouse, L. L., Snow, A. N. & Sidiropoulos, N. MET exon 14 skipping mutations: essential considerations for current management of non–small-cell lung cancer. *J. Mol. Diagn.* **24**, 841–843 (2022).

50. Ha, S. Y. et al. MET overexpression assessed by new interpretation method predicts gene amplification and poor survival in advanced gastric carcinomas. *Mod. Pathol.* **26**, 1632–1641 (2013).

51. Ivanova, M. et al. Standardized pathology report for HER2 testing in compliance with 2023 ASCO/CAP updates and 2023 ESMO consensus statements on HER2-low breast cancer. *Virchows Arch.* **484**, 3–14 (2024).

52. Ke, H.-L. et al. High Ubiquitin-Specific Protease 2a Expression Level Predicts Poor Prognosis in Upper Tract Urothelial Carcinoma. *Appl. Immunohistochem. Mol. Morphol.* **30**, 304–310 (2022).

53. Chu, P.-Y., Tzeng, Y.-D. T., Tsui, K.-H., Chu, C.-Y. & Li, C.-J. Downregulation of ATP binding cassette subfamily a member 10 acts as a prognostic factor associated with immune infiltration in breast cancer. *Aging* **14**, 2252 (2022).

54. Choi, K. M. et al. The interferon-inducible protein viperin controls cancer metabolic reprogramming to enhance cancer progression. *J. Clin. Invest.* **132**, e157302 (2022).

55. Tzeng, Y. T. et al. Integrated analysis of pivotal biomarker of LSM1 , immune cell infiltration and therapeutic drugs in breast cancer. *J. Cell. Mol. Med.* **26**, 4007–4020 (2022).

56. Wolff, A. C. et al. Human epidermal growth factor receptor 2 testing in breast cancer: ASCO–College of American Pathologists Guideline Update. *J. Clin. Oncol.* **41**, 3867–3872 (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*

arXiv:1412.6980 (2017).

67. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Jorge Cardoso, M. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* 240–248 (2017).

68. Loshchilov, I. & Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv* arXiv:1608.03983 (2017).

69. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* arXiv:1802.03426 (2020).

70. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. *International Conference on Machine Learning* (PMLR) 12310–12320 (2021).

71. Dolezal, J. M. et al. Slideflow: Deep Learning for Digital Histopathology with Real-Time Whole-Slide Visualization. *arXiv* arXiv:2304.04142 (2023).

Acknowledgments

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

Author contributions

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. 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. conceived the experiments, B.B., M.M., J.R., and J.P. conducted the experiments. B.B., M.M., 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., H.K., S.I.C., and S.K. interpreted and validated the algorithms. B.B., M.M., S.I.C., S.M.A., and S.K. prepared an initial draft of the manuscript. All authors reviewed and approved the final version of the manuscript.

Competing interests

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

Figure legend

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

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

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

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

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

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

Fig. 5. Performance analysis of the artificial intelligence (AI) models on novel immunostains with varying interpretation cutoffs. a The receiver operating characteristic (ROC) curve by changing the cutoff over the predicted TPS and measuring false and true positive rates. In this experiment, we fixed the ground truth TPS cutoff to 1% since it is the most common and intuitive. **b** Comparing UIHC and single-cohort models across a range of 1% and the second cutoff value within the [2%, 75%] range, illustrating the 3-way classification accuracy. UIHC, universal immunohistochemistry model; H-Br, HER2 of breast; -Br, PD-L1 22C3 of breast; P-L, PD-L1 22C3 of lung; P-Bl, PD-L1 22C3 of bladder. AVG, average.

Fig. 6. Histopathologic validation of the universal immunohistochemistry (UIHC) model. a Lung cancer whole slide image (WSI) is stained with mesenchymal-epithelial transition factor (MET). The UIHC model predicts more accurate classes unlike the P-L model which confuses positively stained Tumor Cell (TC+) with negatively stained Tumor Cell (TC-). **b** Gastric cancer WSI is stained with fibroblast growth factor receptor 2 (FGFR2). P-L, PD-L1 22C3 of lung; TC, tumor cell.

Fig. 7. Qualitative analysis of the artificial intelligence (AI)-learned representation. **a** Two-dimensional (2D) projection of internal representation colored by tumor proportion score (TPS). Each patch is encoded to a 2D plot using three representations: raw pixels, self-supervised

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

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

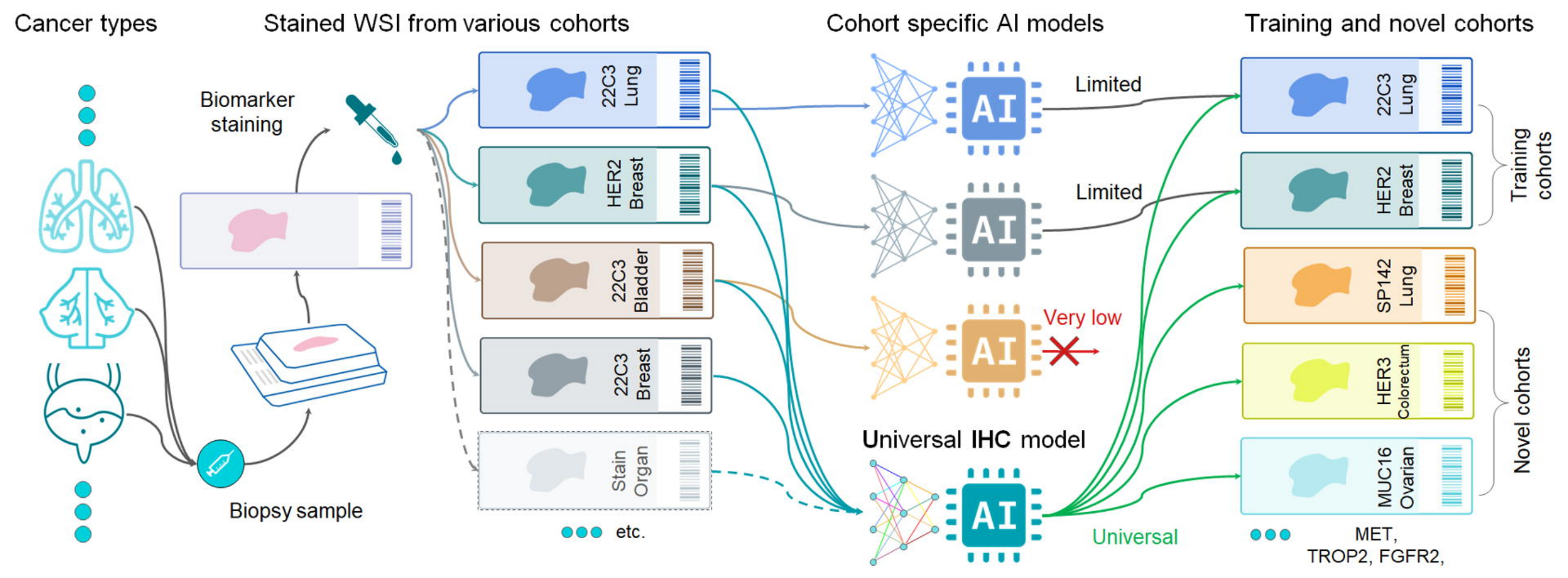
Table 1. Validation of universal immunohistochemistry (UIHC) model on cases with next-generation sequencing results.

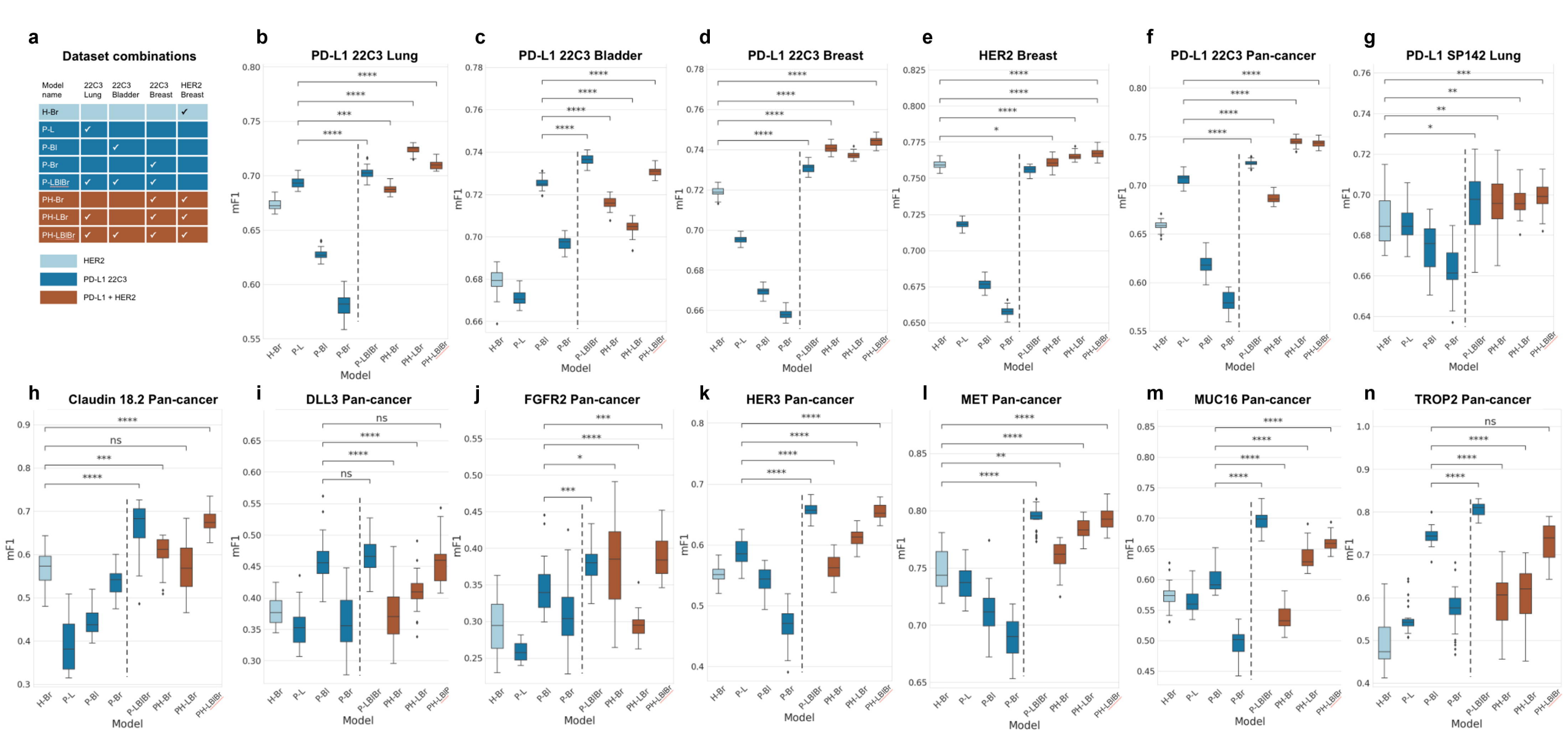
Case no.	Organ	Group	Mutation/Amplification detail	UIHC TPS	Average UIHC TPS according to the group
1	Lung	EGFR exon20ins	p.Ala763_Tyr764insPheGlnGluAla	68.6	75.7±23.2
2	Lung	EGFR exon20ins	p.Ala767_Val769dup	85.7	
3	Lung	EGFR exon20ins	p.Asp770_Asn771insGly	88.4	
4	Lung	EGFR exon20ins	p.Ser768_Asp770dup	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	

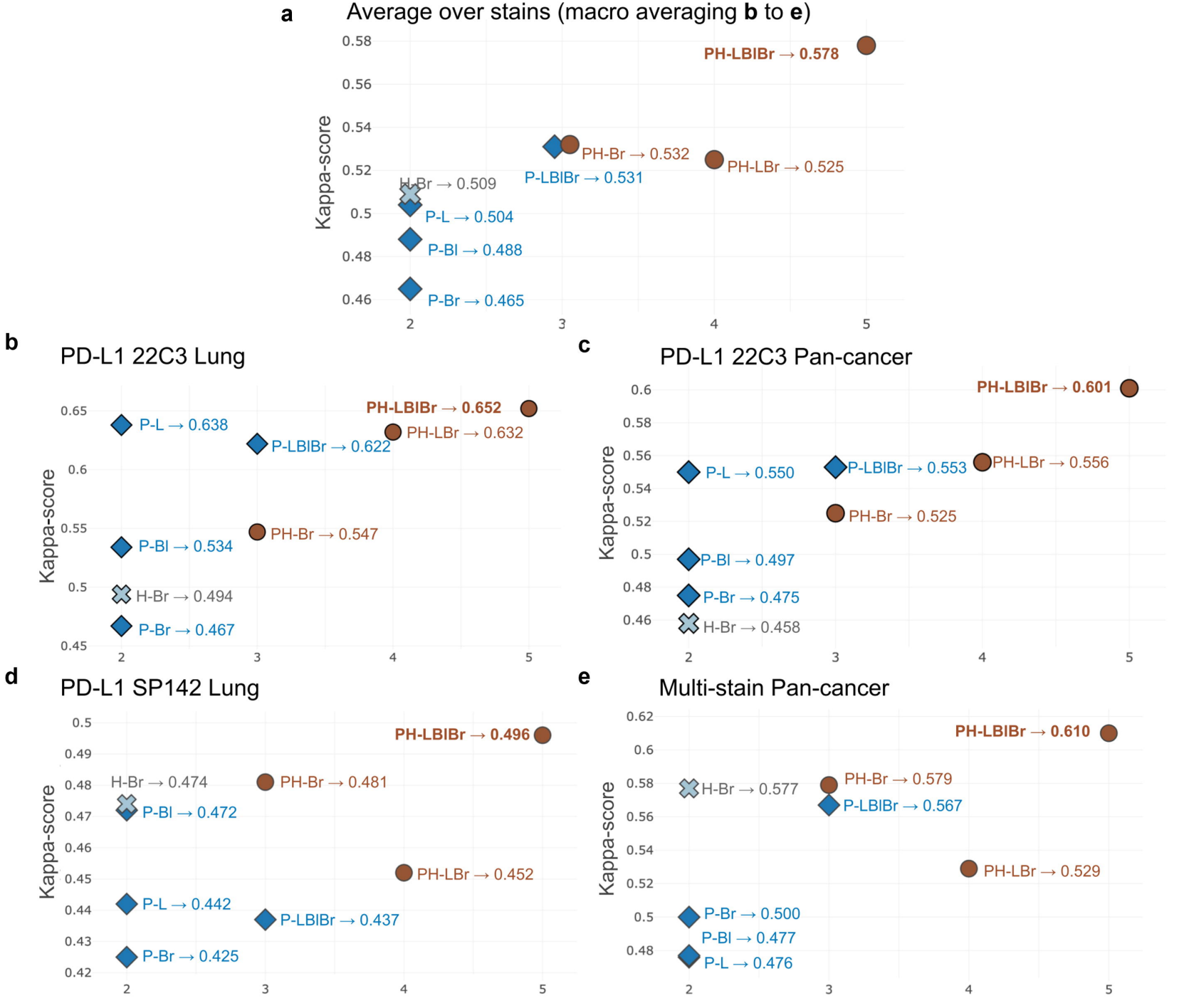
16	Lung	MET amplification	5 copies	96.5	

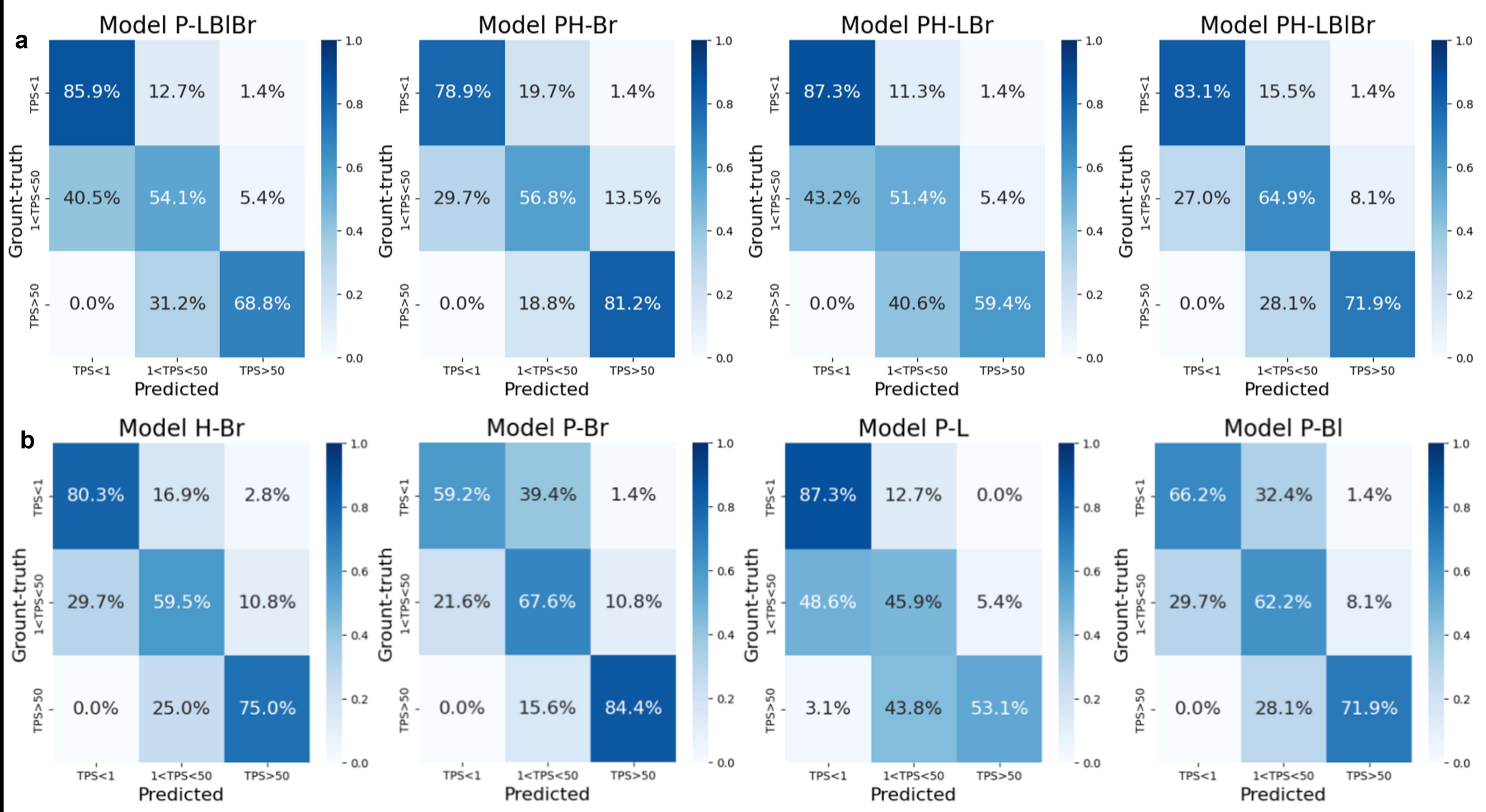
788 EGFR, epidermal growth factor receptor; MET, mesenchymal-epithelial transition

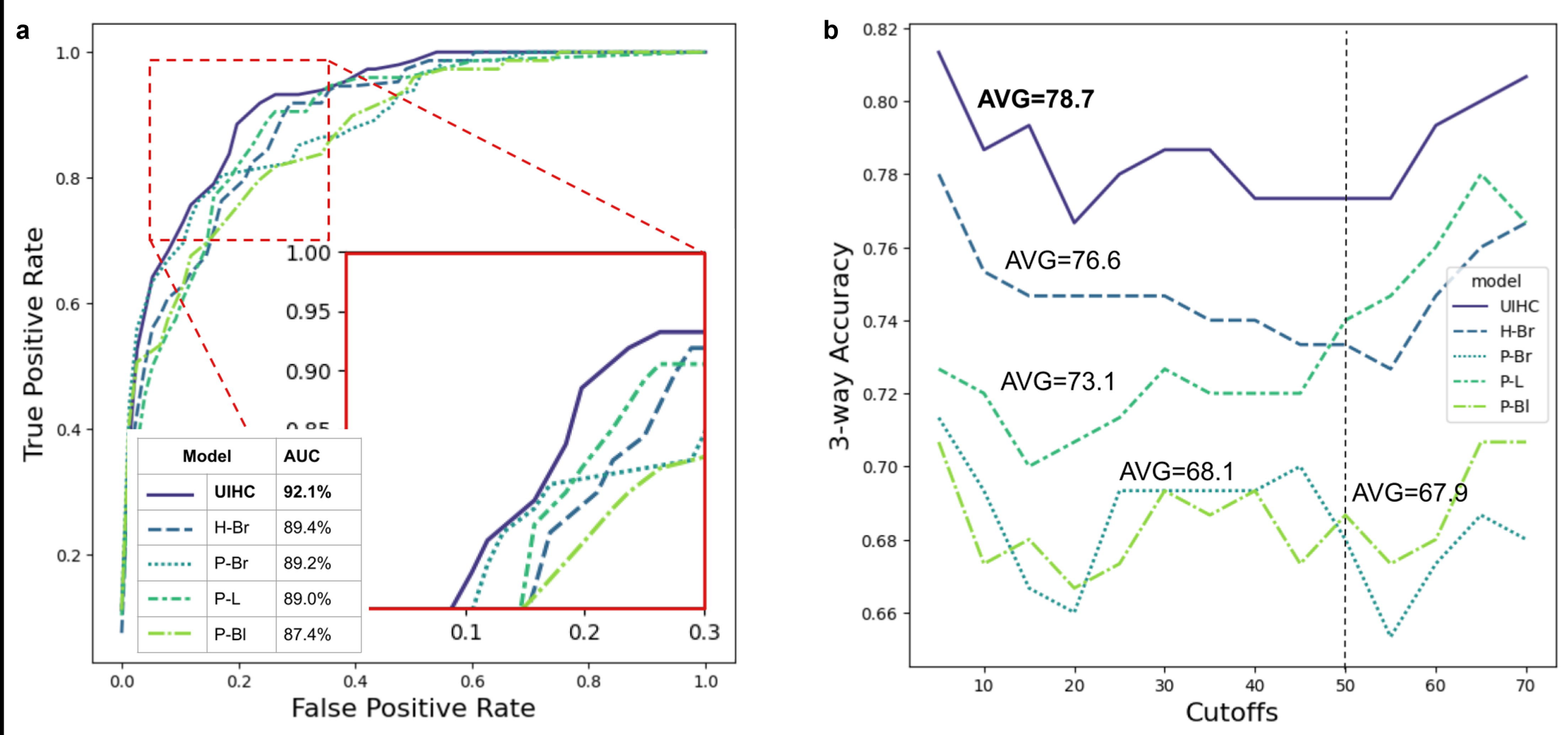
789 factor; TPS, tumor proportion score.

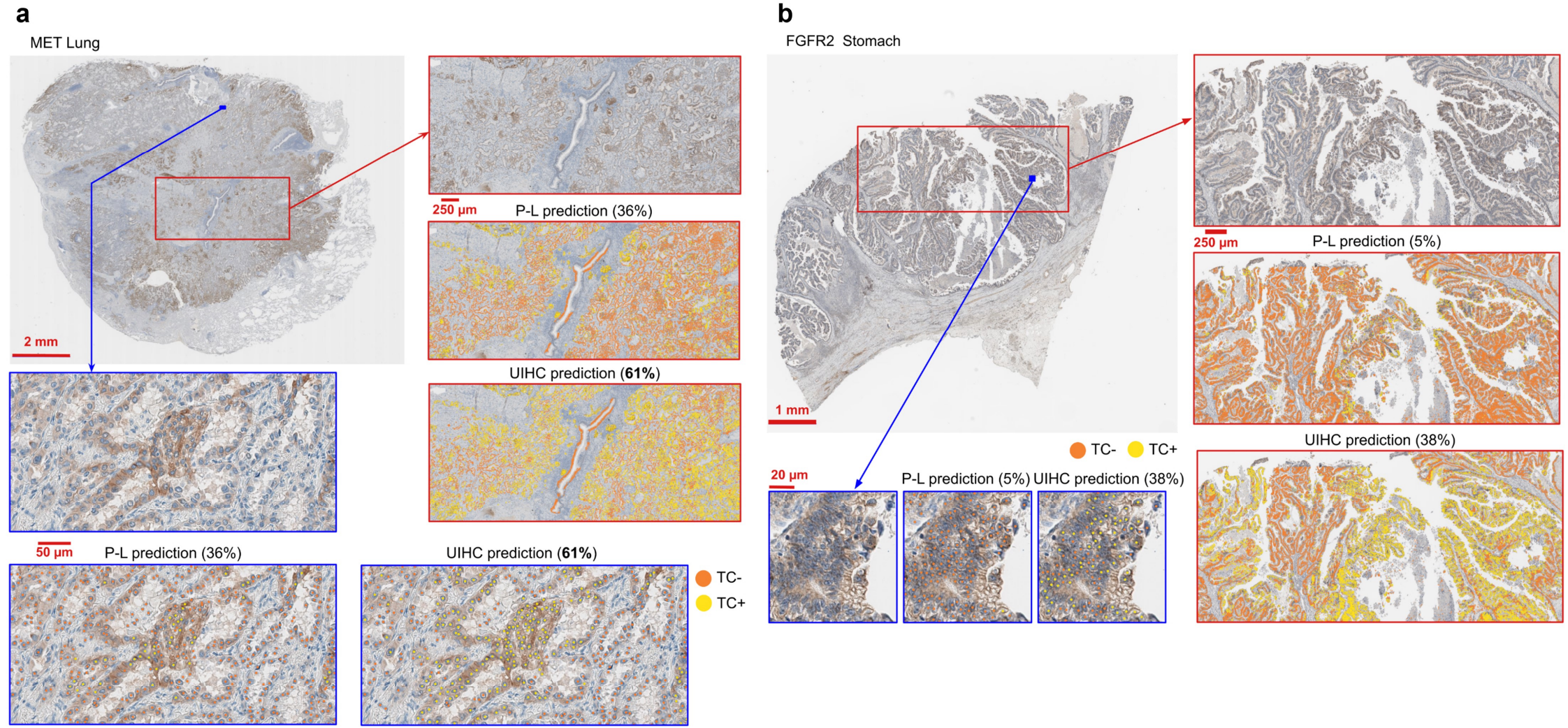

Cancer types
Stained WSI from various cohorts
Cohort specific AI models
Training and novel cohorts
Biomarker staining
Biopsy sample
22C3 Lung
HER2 Breast
22C3 Bladder
22C3 Breast
Stain Organ
etc.
AI
Limited
Limited
Very low
Universal IHC model
Universal
Training cohorts
SP142 Lung
HER3 Colorectum
MUC16 Ovarian
Novel cohorts
MET, TROP2, FGFR2,

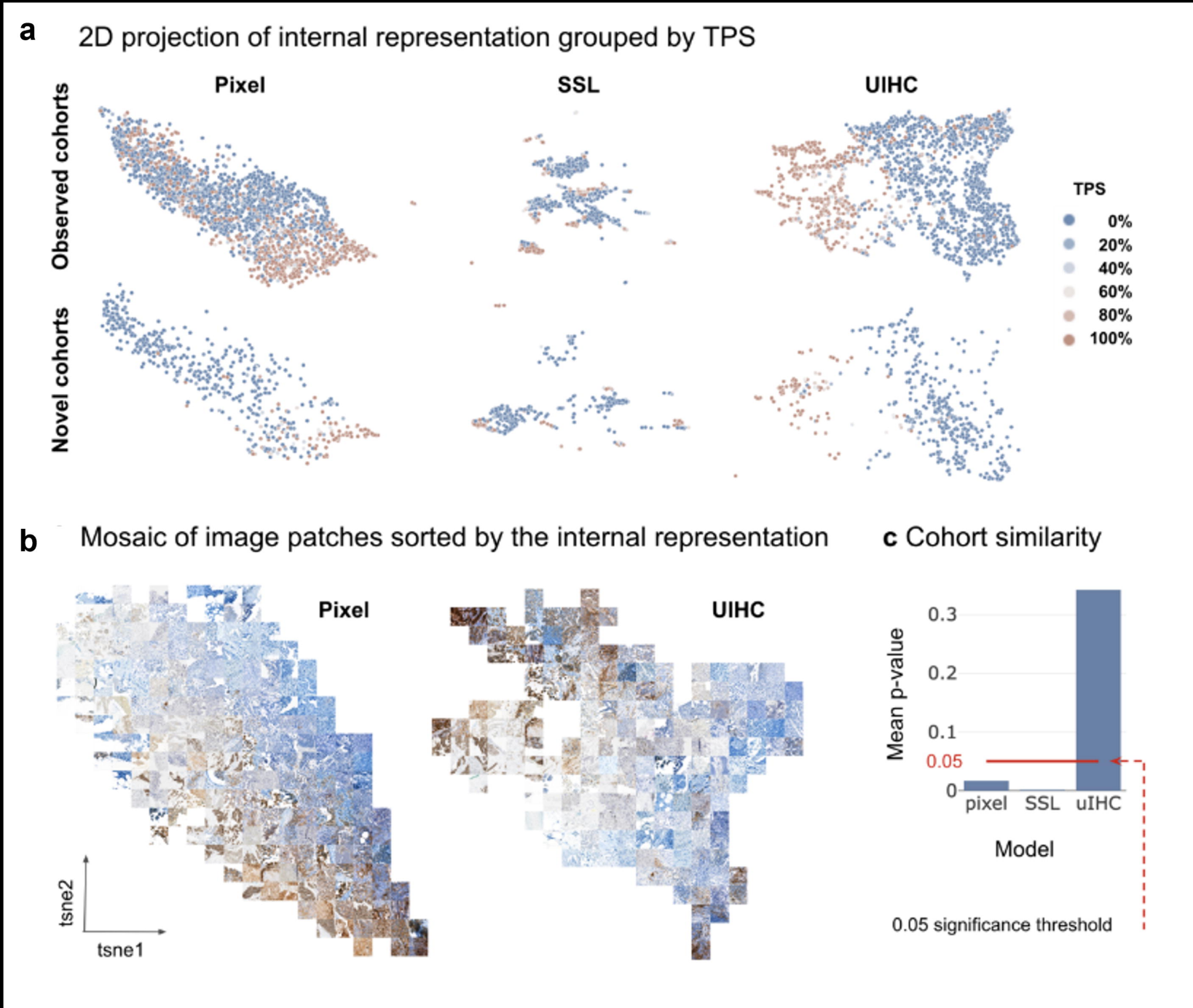

a
Dataset combinations
Model name
22C3 Lung
22C3 Bladder
22C3 Breast
HER2 Breast
H-Br
P-L
P-Bl
P-Br
P-LBlBr
PH-Br
PH-LBr
PH-LBlBr
HER2
PD-L1 22C3
PD-L1 + HER2
b
PD-L1 22C3 Lung
c
PD-L1 22C3 Bladder
d
PD-L1 22C3 Breast
e
HER2 Breast
f
PD-L1 22C3 Pan-cancer
g
PD-L1 SP142 Lung
h
Claudin 18.2 Pan-cancer
i
DLL3 Pan-cancer
j
FGFR2 Pan-cancer
k
HER3 Pan-cancer
l
MET Pan-cancer
m
MUC16 Pan-cancer
n
TROP2 Pan-cancer
mF1
Model

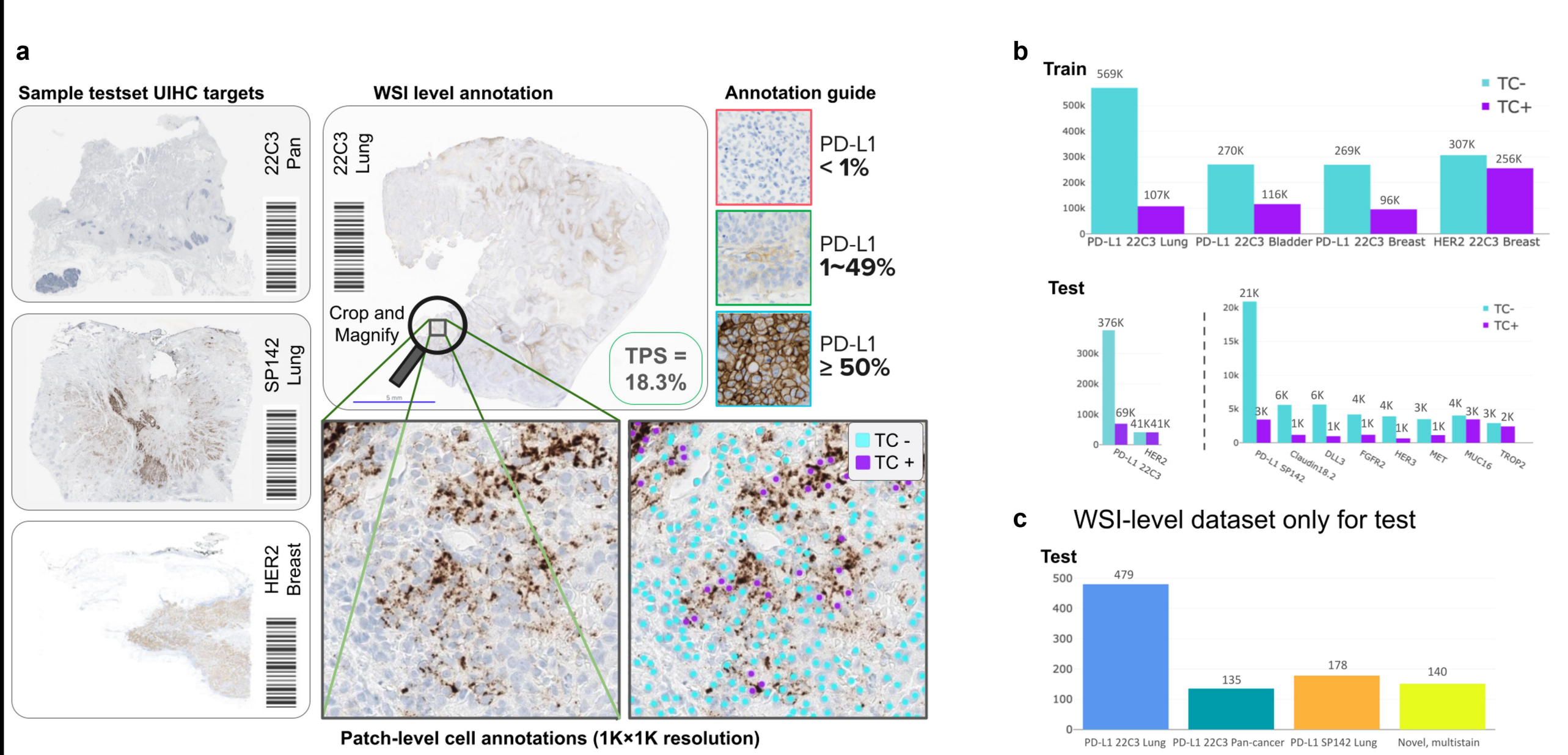

a
Average over stains (macro averaging b to e)
Kappa-score
PH-LBIBr → 0.578
PH-Br → 0.532
PH-LBr → 0.525
P-LBIBr → 0.531
H-Br → 0.509
P-L → 0.504
P-BI → 0.488
P-Br → 0.465
b
PD-L1 22C3 Lung
PH-LBIBr → 0.652
PH-LBr → 0.632
P-L → 0.638
P-LBIBr → 0.622
PH-Br → 0.547
P-BI → 0.534
H-Br → 0.494
P-Br → 0.467
c
PD-L1 22C3 Pan-cancer
PH-LBIBr → 0.601
P-L → 0.550
P-LBIBr → 0.553
PH-LBr → 0.556
PH-Br → 0.525
P-BI → 0.497
P-Br → 0.475
H-Br → 0.458
d
PD-L1 SP142 Lung
PH-LBIBr → 0.496
H-Br → 0.474
PH-Br → 0.481
P-BI → 0.472
PH-LBr → 0.452
P-L → 0.442
P-LBIBr → 0.437
P-Br → 0.425
e
Multi-stain Pan-cancer
PH-LBIBr → 0.610
H-Br → 0.577
PH-Br → 0.579
P-LBIBr → 0.567
PH-LBr → 0.529
P-Br → 0.500
P-BI → 0.477
P-L → 0.476

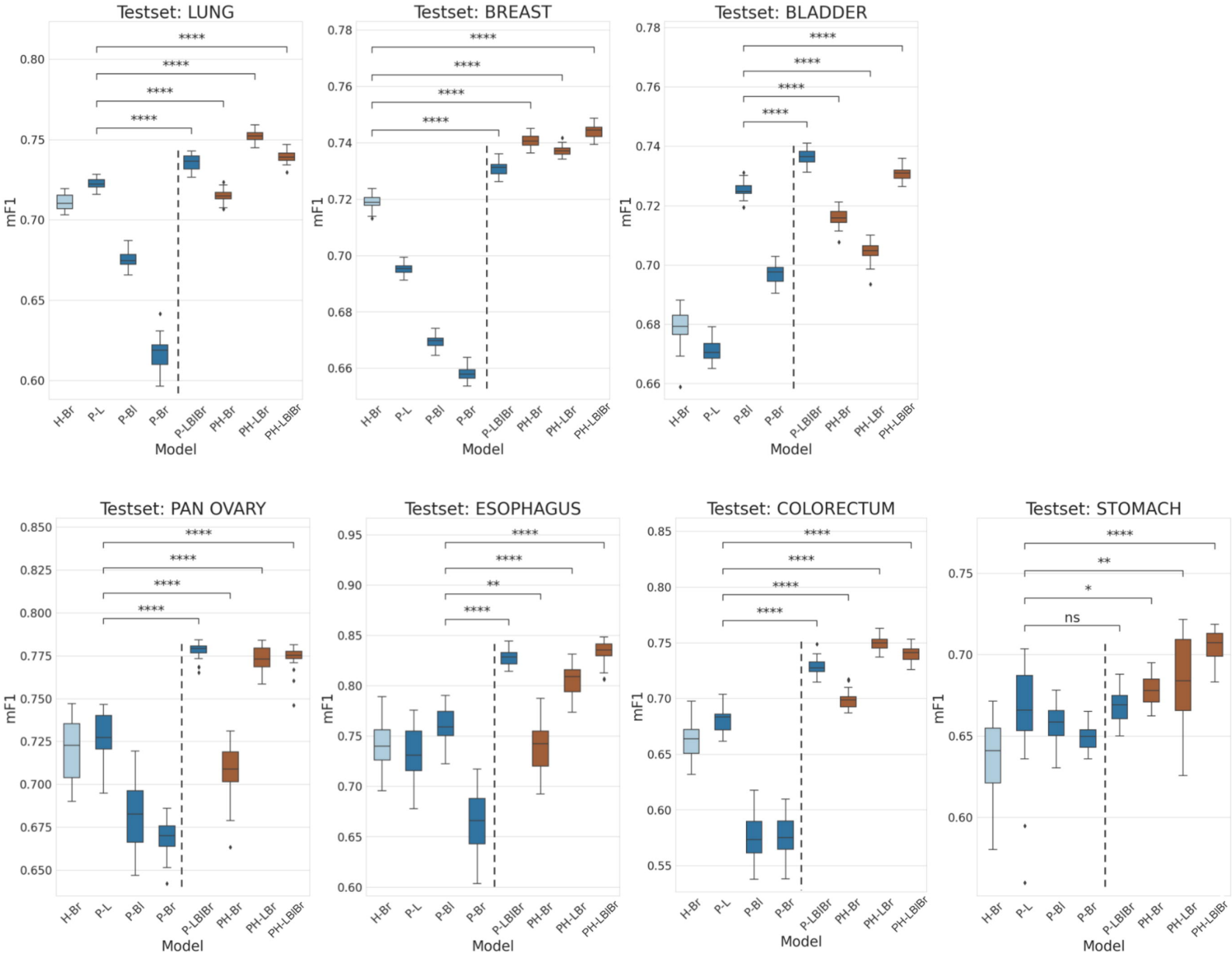

a
Model P-LBlBr
85.9% 12.7% 1.4%
40.5% 54.1% 5.4%
0.0% 31.2% 68.8%
Model PH-Br
78.9% 19.7% 1.4%
29.7% 56.8% 13.5%
0.0% 18.8% 81.2%
Model PH-LBr
87.3% 11.3% 1.4%
43.2% 51.4% 5.4%
0.0% 40.6% 59.4%
Model PH-LBlBr
83.1% 15.5% 1.4%
27.0% 64.9% 8.1%
0.0% 28.1% 71.9%
b
Model H-Br
80.3% 16.9% 2.8%
29.7% 59.5% 10.8%
0.0% 25.0% 75.0%
Model P-Br
59.2% 39.4% 1.4%
21.6% 67.6% 10.8%
0.0% 15.6% 84.4%
Model P-L
87.3% 12.7% 0.0%
48.6% 45.9% 5.4%
3.1% 43.8% 53.1%
Model P-Bl
66.2% 32.4% 1.4%
29.7% 62.2% 8.1%
0.0% 28.1% 71.9%
Grount-truth
TPS<1
1<TPS<50
TPS>50
Predicted
1.0
0.8
0.6
0.4
0.2
0.0

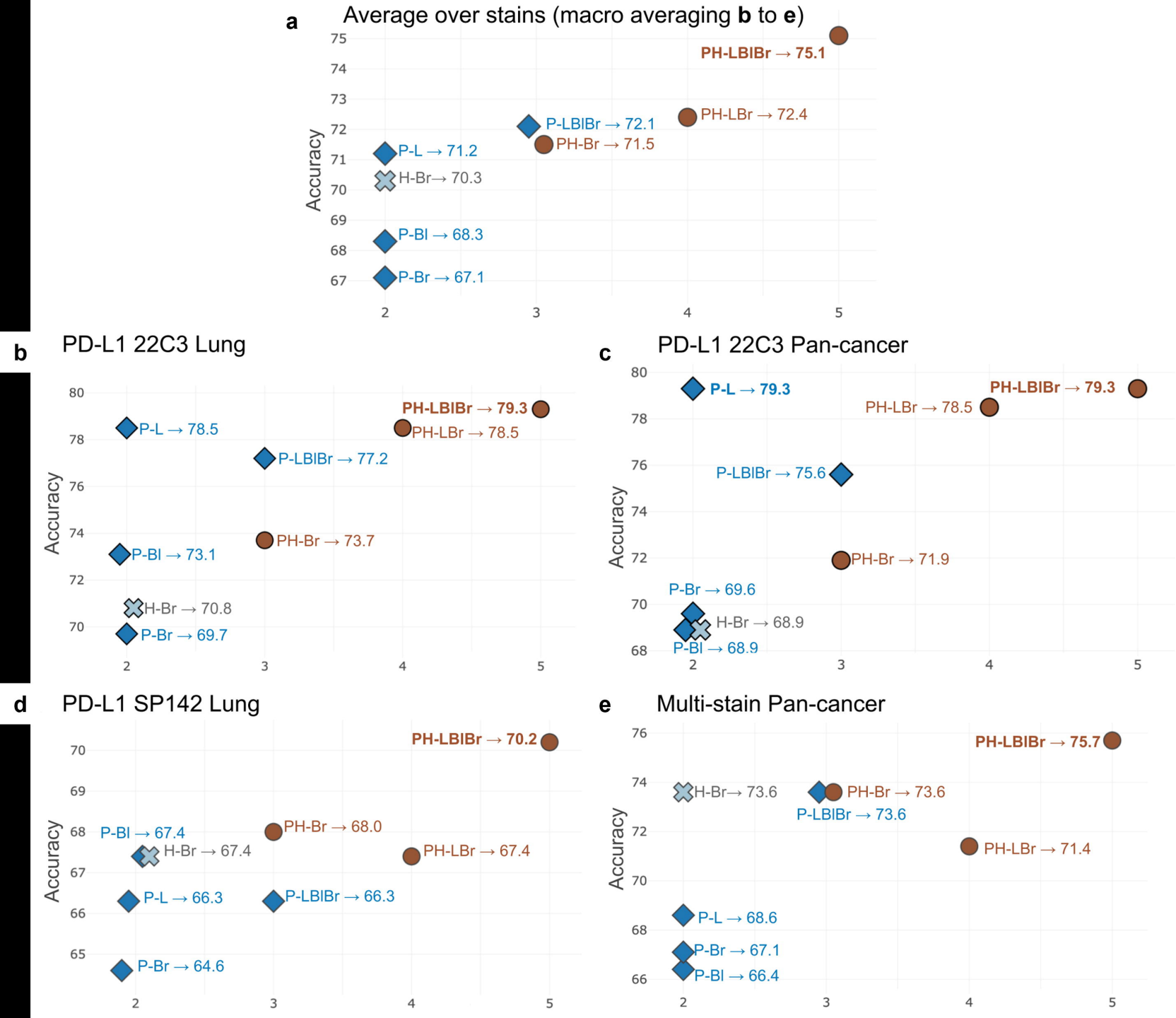

a
True Positive Rate
False Positive Rate
Model
AUC
UIHC
92.1%
H-Br
89.4%
P-Br
89.2%
P-L
89.0%
P-Bl
87.4%
b
3-way Accuracy
Cutoffs
AVG=78.7
AVG=76.6
AVG=73.1
AVG=68.1
AVG=67.9
model
UIHC
H-Br
P-Br
P-L
P-Bl


a
MET Lung
2 mm
250 µm
P-L prediction (36%)
UIHC prediction (61%)
50 µm
P-L prediction (36%)
UIHC prediction (61%)
TC-
TC+
b
FGFR2 Stomach
1 mm
250 µm
P-L prediction (5%)
UIHC prediction (38%)
TC-
TC+
20 µm
P-L prediction (5%)
UIHC prediction (38%)


a
2D projection of internal representation grouped by TPS
Pixel
SSL
UIHC
Observed cohorts
Novel cohorts
TPS
0%
20%
40%
60%
80%
100%
b
Mosaic of image patches sorted by the internal representation
Pixel
UIHC
tsne2
tsne1
c Cohort similarity
Mean p-value
0.3
0.2
0.1
0.05
0
pixel
SSL
uIHC
Model
0.05 significance threshold


a
Sample testset UIHC targets
22C3 Pan
SP142 Lung
HER2 Breast
WSI level annotation
22C3 Lung
Crop and Magnify
TPS = 18.3%
Annotation guide
PD-L1 < 1%
PD-L1 1~49%
PD-L1 ≥ 50%
TC -
TC +
Patch-level cell annotations (1K×1K resolution)
b
Train
569K
107K
270K
116K
269K
96K
307K
256K
TC-
TC+
PD-L1 22C3 Lung
PD-L1 22C3 Bladder
PD-L1 22C3 Breast
HER2 22C3 Breast
Test
376K
69K
41K
41K
PD-L1 22C3
HER2
21K
3K
6K
1K
6K
1K
4K
1K
4K
1K
3K
1K
4K
3K
3K
2K
PD-L1 SP142
Claudin18.2
DLL3
FGFR2
HER3
MET
MUC16
TROP2
c
WSI-level dataset only for test
Test
479
135
178
140
PD-L1 22C3 Lung
PD-L1 22C3 Pan-cancer
PD-L1 SP142 Lung
Novel, multistain

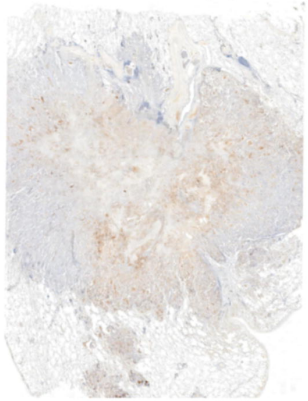

Testset: LUNG
Testset: BREAST
Testset: BLADDER
Testset: PAN OVARY
Testset: ESOPHAGUS
Testset: COLORECTUM
Testset: STOMACH
mF1
Model
H-Br
P-L
P-Bl
P-Br
P-LBlBr
PH-Br
PH-LBr
PH-LBlBr
****
**
*
ns


a
Average over stains (macro averaging b to e)
Accuracy
PH-LBIBr → 75.1
PH-LBr → 72.4
P-LBIBr → 72.1
PH-Br → 71.5
P-L → 71.2
H-Br→ 70.3
P-BI → 68.3
P-Br → 67.1
b
PD-L1 22C3 Lung
PH-LBIBr → 79.3
PH-LBr → 78.5
P-L → 78.5
P-LBIBr → 77.2
PH-Br → 73.7
P-BI → 73.1
H-Br → 70.8
P-Br → 69.7
c
PD-L1 22C3 Pan-cancer
P-L → 79.3
PH-LBIBr → 79.3
PH-LBr → 78.5
P-LBIBr → 75.6
PH-Br → 71.9
P-Br → 69.6
H-Br → 68.9
P-BI → 68.9
d
PD-L1 SP142 Lung
PH-LBIBr → 70.2
PH-Br → 68.0
P-BI → 67.4
H-Br → 67.4
PH-LBr → 67.4
P-L → 66.3
P-LBIBr → 66.3
P-Br → 64.6
e
Multi-stain Pan-cancer
PH-LBIBr → 75.7
H-Br→ 73.6
PH-Br → 73.6
P-LBIBr → 73.6
PH-LBr → 71.4
P-L → 68.6
P-Br → 67.1
P-BI → 66.4

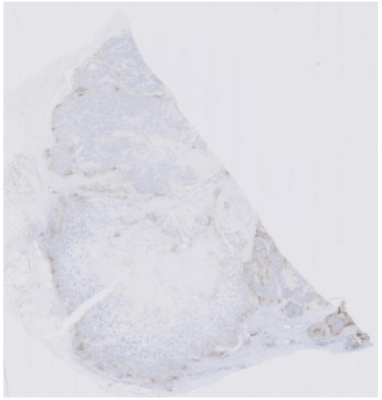
a

IHCs samples from training cohorts (TPS %)

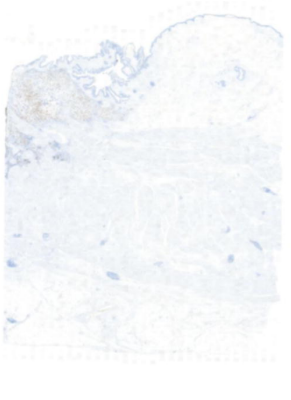
PD-L1 22C3 lung (40%)



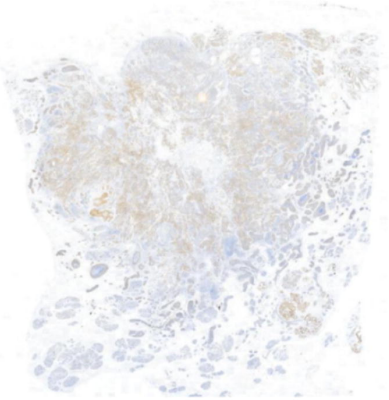
PD-L1 22C3 breast (0%)



PD-L1 22C3 bladder (0%)



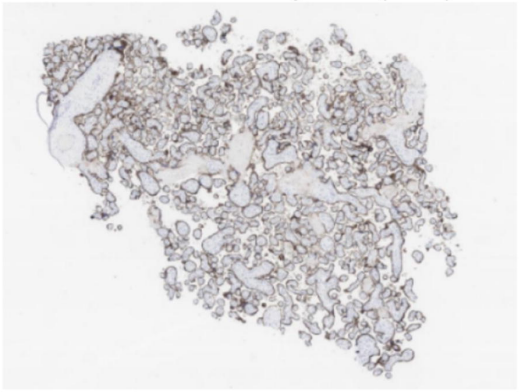
HER2 breast (90%)



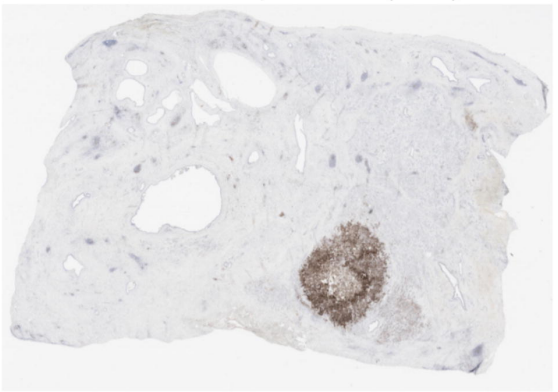
b

IHCs samples from novel cohorts (TPS %)

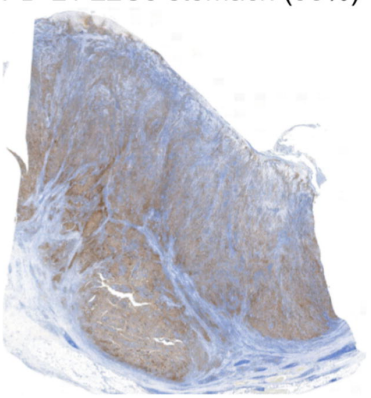
PD-L1 22C3 biliary tract (85%)



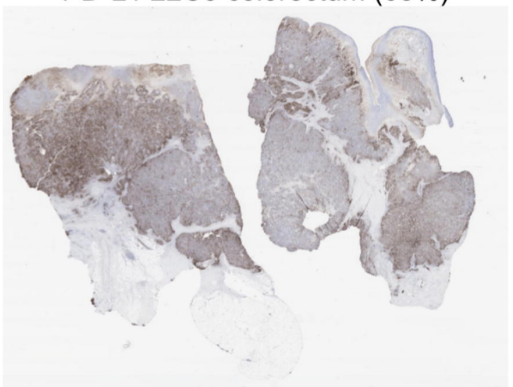
PD-L1 22C3 pancreas (70%)



PD-L1 22C3 stomach (93%)



PD-L1 22C3 colorectum (68%)

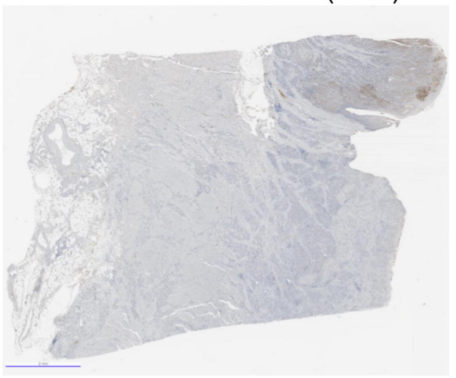


PD-L1 22C3 Liver (93%)

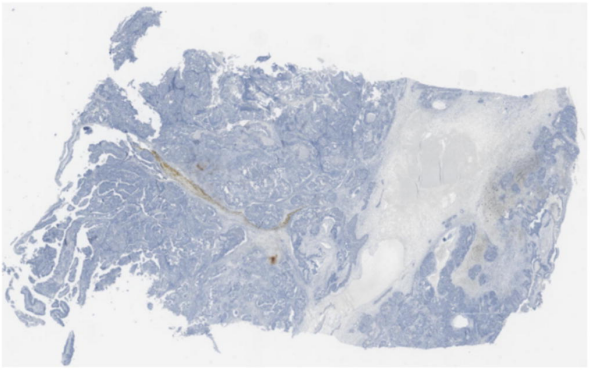
PD-L1 22C3 prostate (50%)

Novel IHCs samples (TPS)

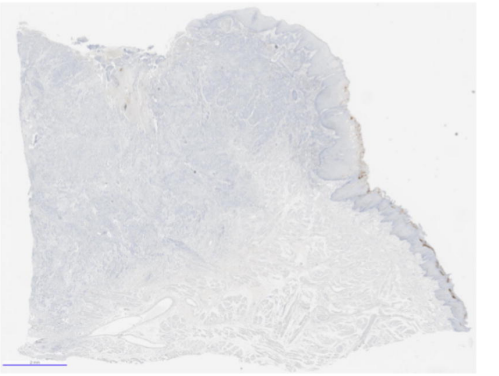
Claudin 18.2 stomach (13%)



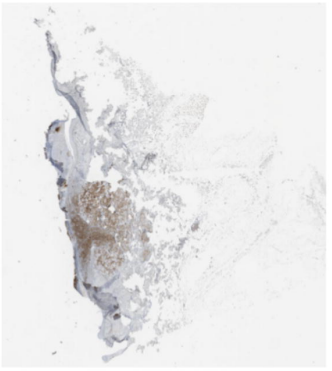
MUC16 cervix (0%)



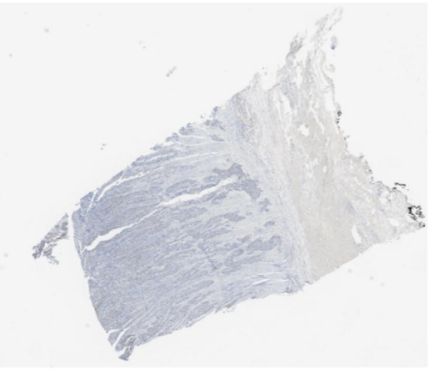
HER3 head & neck (0%)



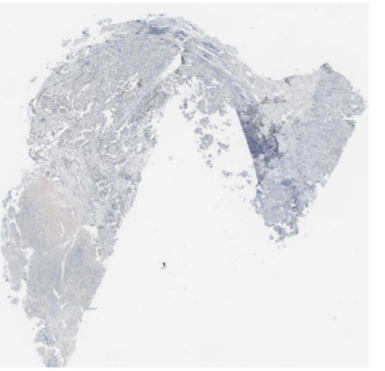
DLL3 melanoma (75%)



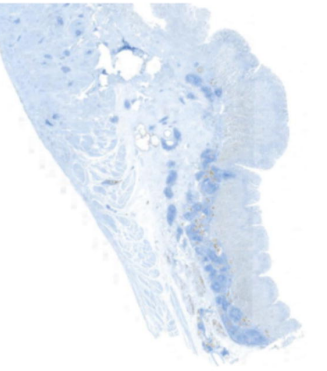
MET esophagus (4%)



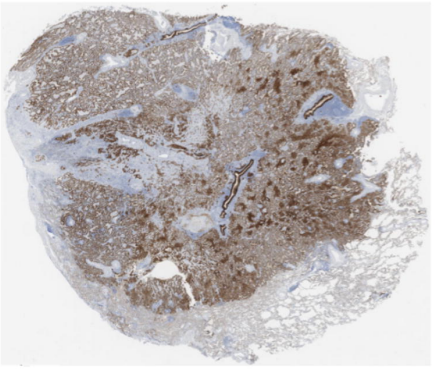
FGFR2 bladder (13%)



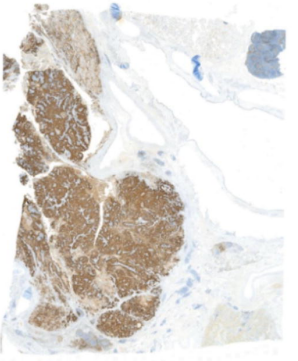
E-Cadherin colorectum (100%)



TROP2 lung (96%)



SP142 lung (100%)



SP263 lung (43%)