npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-024-00770-z

A universal immunohistochemistry analyzer for generalizing AI-driven assessment of immunohistochemistry across immunostains and cancer types

Biagio Brattoli[1,6], Mohammad Mostafavi[1,6], Taebum Lee[1,6], 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] ✉ & Seokhwi Kim[4,5] ✉

Immunohistochemistry (IHC) is the common companion diagnostics in targeted therapies. However, quantifying protein expressions in IHC images present a significant challenge, due to variability in manual scoring and inherent subjective interpretation. Deep learning (DL) offers a promising approach to address these issues, though current models require extensive training for each cancer and IHC type, limiting the practical application. We developed a Universal IHC (UIHC) analyzer, a DL-based tool that quantifies protein expression across different cancers and IHC types. This multi-cohort trained model outperformed conventional single-cohort models in analyzing unseen IHC images (Kappa score 0.578 vs. up to 0.509) and demonstrated consistent performance across varying positive staining cutoff values. In a discovery application, the UIHC model assigned higher tumor proportion scores to MET amplification cases, but not MET exon 14 splicing or other non-small cell lung cancer cases. This UIHC model represents a novel role for DL that further advances quantitative analysis of IHC.

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]. For example, tumor proportion score (TPS), which represents the percentage of positive tumor cells for a particular IHC among all tumor cells, is known to be a favorable index for treatment response in patients with a certain score or higher.

Recently, artificial intelligence (AI) models that use deep learning (DL) 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 DL models is heavily constrained by their reliance on training cohorts that typically contain at least several hundred or often more WSI cases of a single cancer type and single immunostain matched to the desired indication. Moreover, these training sets are manually labeled ('annotated') on a cellular basis by pathologists with each individual slide taking several hours for annotation depending on complexity[10,11].

[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. [6]These authors contributed equally: Biagio Brattoli, Mohammad Mostafavi, Taebum Lee. ✉e-mail: sooickcho@lunit.io; seokhwikim@ajou.ac.kr

THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA

Importantly, there is an additional limitation of 'domain-shift', where current DL models for IHC cannot recognize elements - either immunostain or cancer type - that are absent from the data set used to train the model. This limitation indicates for each immunostain-cancer type combination, a matched IHC training set must be created and annotated with accompanying significant time and resource cost, which is particularly relevant when evaluating multiple candidates for antibody 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 DL model that is proficient in interpreting IHC results without 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 and without need for matching training set. Eight models were defined by their training regimes which consist of exposure to varying single or multiple cohorts. These cohorts were 'patches' (discrete areas) taken from WSI of three cancer types (lung, breast, and urothelium), which were also immunostained for PD-L1 22C3 or Human Epidermal growth factor Receptor 2 (HER2). For each patch, pathologists annotated positively stained tumor cells (TC+) and negatively stained tumor cells (TC−). On the basis of this annotation, the DL models were trained to detect TC+ and TC−, and then could calculate the TPS across the entire region. Models trained on single cohorts served as the benchmark and were similar to prior work, whereas models trained with multiple cohorts were the innovation described in this study[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

Single-cohort-derived models (SC-models) were trained on one dataset (cohort) and multiple-cohort-derived models (MC-models) were trained on multiple datasets. These datasets contained NSCLC, urothelial carcinoma, and breast cancer cases stained with 22C3 and breast cancer datasets stained with HER2 (Fig. 1). Figures 2a and 3a show the combination of different datasets to develop the eight DL models utilized in this study. Figures 2b–g and 3b–h show patch-level performance of various DL models on different IHC datasets. SC-models exhibit favorable performance within test sets matched for the immunostain and cancer type used for training as evidenced by the cell detection (TC− or TC+) performance (median F1-score [min, max]) of the P-L model (a DL model trained on 22C3 stained NSCLC cases) on the 22C3 Lung test set (0.693 [0.686, 0.705], Fig. 2b), the P-U model (trained on 22C3 of urothelium) on the 22C3 Urothelial test set (0.725 [0.719, 0.731], Fig. 2c), the P-B (22C3 of breast) on the 22C3 Breast test set (0.599 [0.590, 0.607], Fig. 2d), and the H-B model (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 (the P-LUB model [22C3 of lung, urothelium, and breast], PH-B [22C3 and HER2 of breast], PH-LB [22C3 of lung and HER2 of breast], PH-LUB [22C3 and HER2 of lung, urothelium, 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, Supplementary Table 1).

In assessing test sets that contained 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 22C3 Pan-cancer set in Fig. 2f, MC-models trained with more cancer types (P-LUB) and/or an additional stain (PH-LB and PH-LUB) outperformed the P-L model (P-L, 0.708 [0.694, 0.719], P-LUB, 0.722 [0.716, 0.730], $p < 0.001$; PH-LB, 0.745 [0.735, 0.753], $p < 0.001$; PH-LUB, 0.743 [0.735, 0.752], $p < 0.001$), which was the best performing SC-models.

In the other novel cohorts with unseen immunostains such as PD-L1 SP142 (Fig. 2g), 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), MUCin-16 (MUC16), and TROPhoblast cell-surface antigen 2 (TROP2), MC-models generally performed better than SC-models (Fig. 3a–h, Supplementary Table 2). Most representatively identified in MET Pan-cancer, all the MC-models outperformed the single best performing SC-model, the H-B model (H-B, 0.744 [0.719, 0.781], P-LUB, 0.795 [0.773, 0.810], $p < 0.001$; PH-B, 0.762 [0.725, 0.776], $p < 0.001$; PH-LB, 0.783 [0.767, 0.799], $p < 0.001$; PH-LUB, 0.792 [0.776, 0.815], $p < 0.001$) (Fig. 3f). This tendency for MC-models to outperform SC-models was also observed when the data was categorized by cancer type (lung, breast, urothelium, 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 DL models at the WSI level were subsequently assessed using the test sets outlined in Supplementary Table 3. The ground truth (GT) were categorized based on the TPS, representing the percentage of viable tumor cells that show membranous staining for PD-L1 out of the total number of viable tumor cells present in the sample by pathologists'

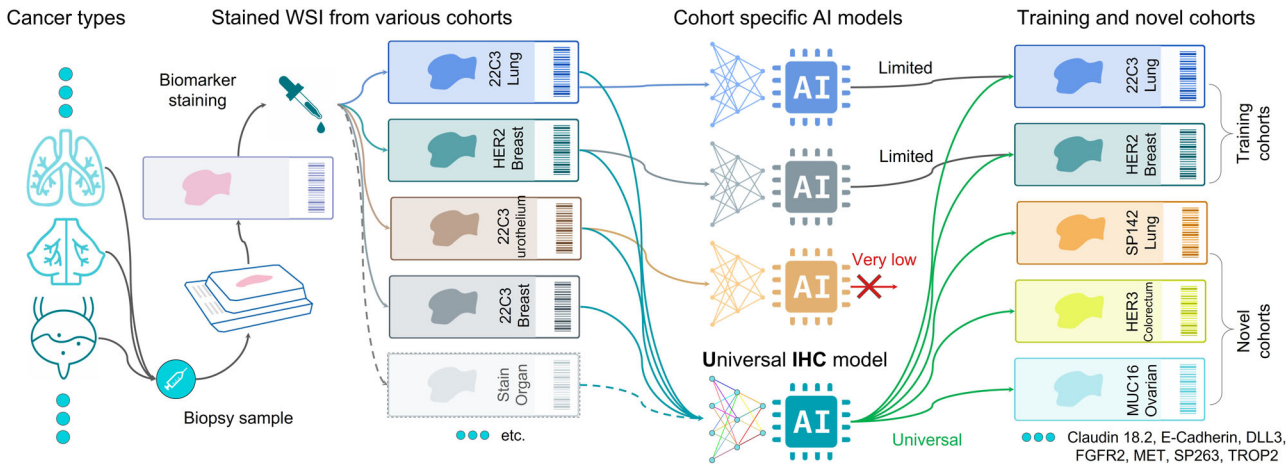

Fig. 1 | Overview of the universal immunohistochemistry (UIHC) artificial intelligence (AI) model development. Single-cohort-derived models (SC-models) were trained using one dataset, while multiple-cohort-derived models (MC-models) were trained using multiple datasets, including lung, urothelial carcinoma, 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 Claudin 18.2, DeLta-Like 3 (DLL3), E-Cadherin, Fibroblast Growth Factor Receptor 2 (FGFR2), Human Epidermal growth factor Receptor 3 (HER3), Mesenchymal-Epithelial Transition factor (MET), MUcin-16 (MUC16), PD-L1 SP142, PD-L1 SP263, and TROPhoblast cell-surface antigen 2 (TROP2).

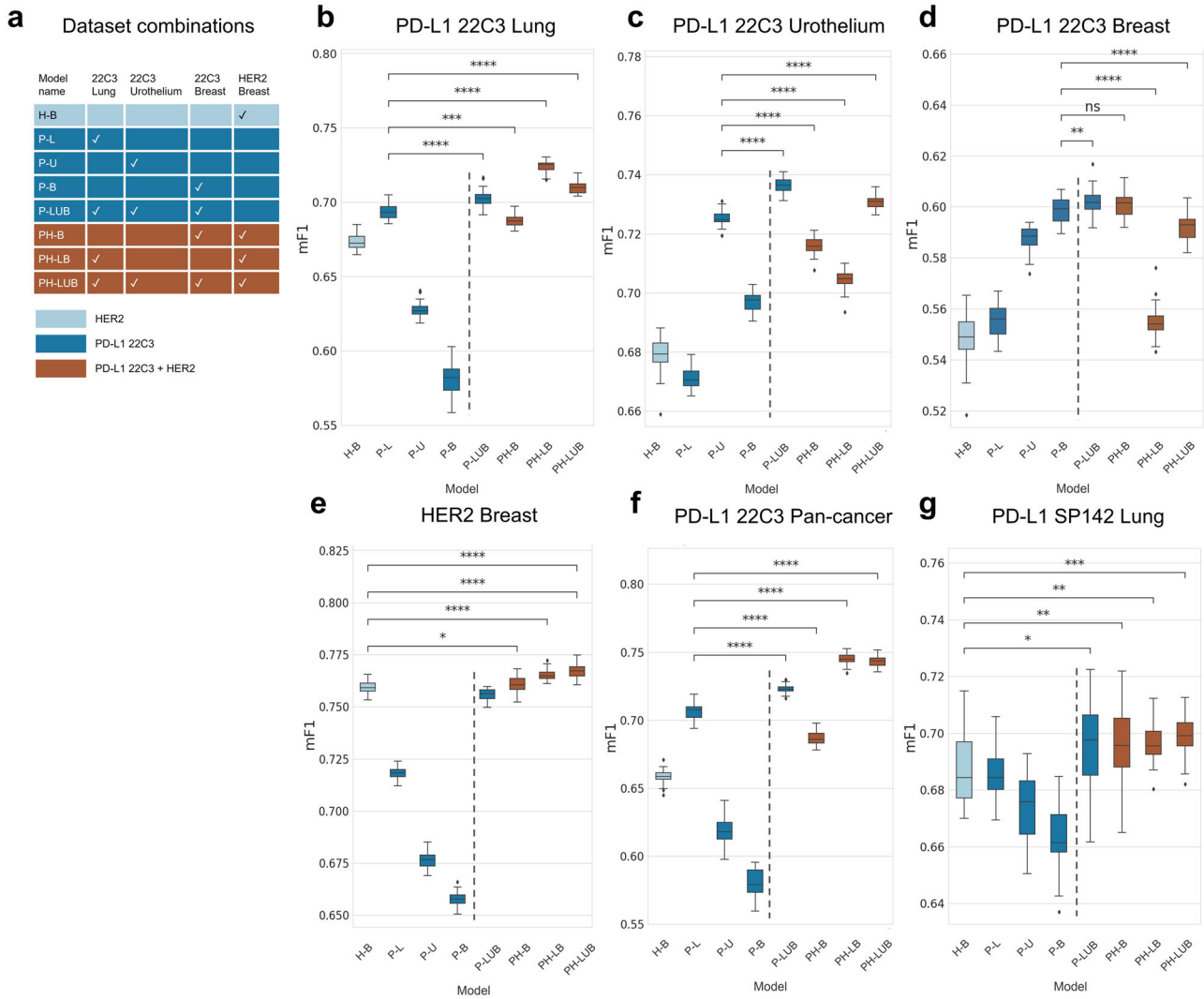

Fig. 2 | Patch-level quantitative analysis on training cohorts or comparable datasets. a List of eight deep learning models trained on different cohort combinations. H-B, HER2 of breast; P-L, PD-L1 22C3 of lung; P-B, 22C3 of breast; P-LUB, 22C3 of lung, urothelium, and breast; PH-B, 22C3 and HER2 of breast; PH-LB, 22C3 of lung and HER2 of breast; PH-LUB, 22C3 and HER2 of lung, urothelium, and breast. The different stain combinations (e.g. 22C3 or HER2 is utilized or not), are visualized by color. Performance of the eight models in training cohorts, where the stain type may be utilized during training – **b** 22C3 in lung cancer, **c** 22C3 in urothelial cancer, **d** 22C3 in breast cancer, **e** HER2 in breast cancer. **f** 22C3 in pan-cancer and **g** PD-L1 SP142 of lung are not used during training. PD-L1, Programmed Death-Ligand 1; HER2, Human Epidermal growth factor Receptor 2; mF1, mean F1 score. (****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; ns, not significant).

scoring. The models' performance was evaluated by accurately assigning the WSIs to the corresponding GT group (TPS < 1%; 1–49%; ≥50%). Figure 4, Supplementary Fig. 2, Supplementary Tables 4 and 5 represent WSI-level IHC quantification performance of various DL models compared to pathologists. Among the eight models, the PH-LUB 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. 4a, Supplementary Fig. 2a). The best SC-model overall was the H-B model, 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 22C3 Lung WSI test set, the PH-LUB model was the only model to outperform the P-L model, with a Cohen's kappa score of 0.652 versus 0.638, and an accuracy of 0.793 versus 0.785 (Fig. 4b, Supplementary Fig. 2b). For the 22C3 Pan-cancer WSI test set and the SP142 Lung WSI test set, P-LUB also outperformed all SC-models, including the P-L model (Fig. 4c, d, Supplementary Fig. 2c, d). Notably, in the multi-stain pan-cancer test set, the PH-LUB model consistently outperformed all SC-models (P-L, P-U, P-B, H-B) and MC-models with less diversity in training cohorts (PH-B, P-LUB, and PH-LB), achieving a Cohen's kappa score of 0.610 and an accuracy of 0.757 (Fig. 4e, Supplementary Fig. 2e). Confusion matrices indicated that the PH-LUB 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%) (Supplementary Fig. 3a, b). Due to its consistently high performance across test sets, PH-LUB was designated as the UIHC model.

Performance analysis of UIHC on novel immunostains for different cutoffs

For most immunostains commonly utilized in clinical practice, such as MET, TROP2, and MUC16, the absence of consensus scoring systems poses a challenge for quantitation. 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 GT TPS from multi-stain test set (Fig. 4e), analogous to the well-established approach used for PD-L1 staining, while varying the DL 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

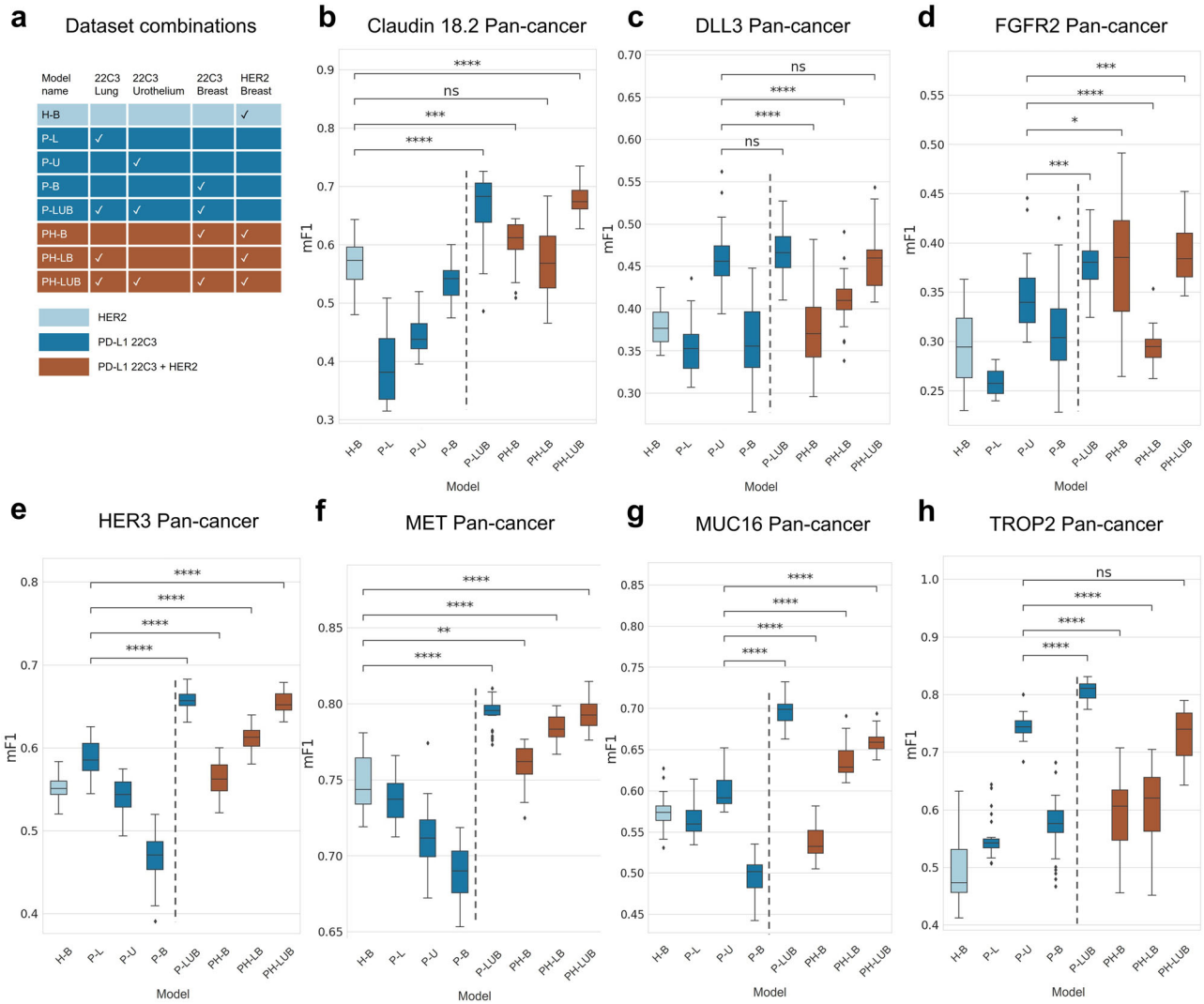

Fig. 3 | Patch-level quantitative analysis on novel cohorts. a List of eight deep learning models trained on different cohort combinations. **b**–**h** Performance of the eight models in novel cohorts where none of the test immunostain types has ever been utilized during the training phase by any of the models. **b** Claudin 18.2, **c** DeLta-Like 3 (DLL3), **d** Fibroblast Growth Factor Receptor 2 (FGFR2), **e** Human Epidermal growth factor Receptor 3 (HER3), **f** Mesenchymal-Epithelial Transition factor (MET), **g** MUCin-16 (MUC16), **h** TROPhoblast cell-surface antigen 2 (TROP2). mF1, mean F1 score. (****$p < 0.0001$; ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; ns, not significant).

our DL 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 relative to SC models, irrespective of the specific cutoff applied for novel stain types.

Pathology validation of efficacy of UIHC model versus SC-model

Representative discrepancy cases between the SC-models (under-performing) and the UIHC model (superior-performing) were subjected to WSI-level histopathological validation by board-certified pathologists to assess the accuracy of the models in detecting IHC-positive cells. In a case involving MET-stained NSCLC, the P-L model incorrectly classified the majority of tumor cells as negative (TPS 36%) (Fig. 6a). Conversely, the UIHC model accurately identified MET-positive tumor cells (TPS 61%), yielding results similar to the average TPS assessment of 75% by pathologists (GT TPS). 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 DL 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, GT 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 a 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

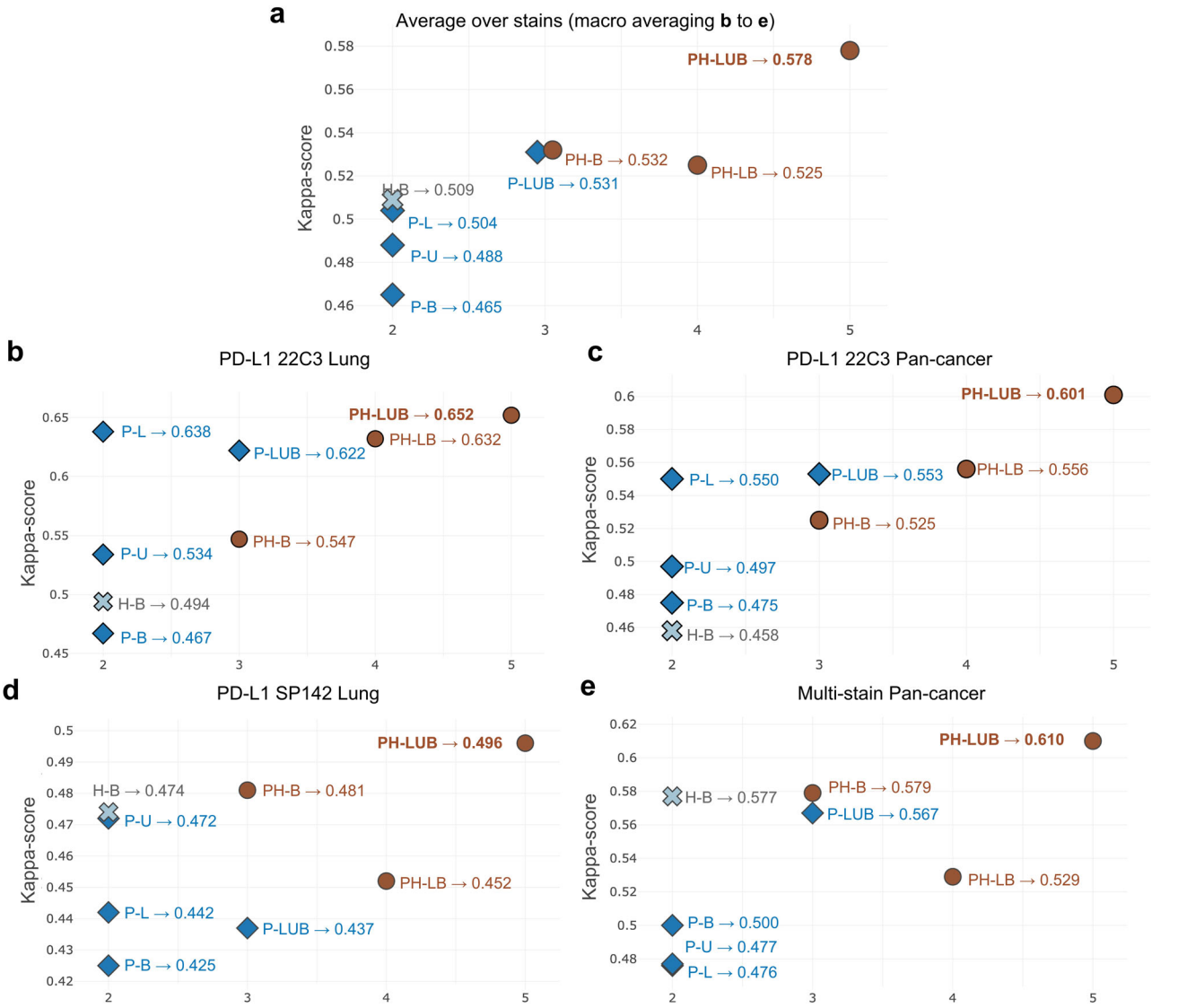

Fig. 4 | Whole slide image (WSI)-level quantitative analysis of the deep learning (DL) models. The quantitative analysis is based on comparing the tumor proportion score (TPS) score in different training settings. The reported Cohen's Kappa scores are computed using the pathologists' labeled category as ground truth. **a** Macro-averaged Cohen's Kappa scores of the eight DL models over all the stains. **b** Cohen's Kappa scores of the DL models in PD-L1 22C3 Lung dataset. **c** Cohen's Kappa scores of the DL models in 22C3 Pan-cancer dataset. **d** Cohen's Kappa scores of the DL models in PD-L1 SP142 Lung dataset. **e** Cohen's Kappa scores of the DL models in multi-stain Pan-cancer dataset. The x-axis presents the summation of the number of utilized stain types and the organ types of each cohort when training (e.g. PH-B [22C3 and HER2 of breast] is 3 as it has 2 stains and 1 cancer type). H-B, HER2 of breast; P-L, 22C3 of lung; P-B, 22C3 of breast; P-LUB, 22C3 of lung, urothelium, and breast; PH-B, 22C3 and HER2 of breast; PH-LB, 22C3 of lung and HER2 of breast; PH-LUB, 22C3 and HER2 of lung, urothelium, and breast. PD-L1, Programmed Death-Ligand 1; HER2, Human Epidermal growth factor Receptor 2.

our approach facilitated the development of an AI-powered analyzer capable of generalizing 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.

Figure 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).

Applying UIHC for discovery in oncogenically driven NSCLC

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]. MET exon 14 skipping mutation have been posited to increase the amount of c-MET protein by removing exon14 as encoding for a region signaling degradation, and thus increasing protein amount and/or receptor activity[19]. The UIHC model assigned higher TPS to the MET amplification group compared to the other groups (Table 1). The UIHC model yielded TPS of 94.5 ± 2.0 for the three MET amplification cases, 77.1 ± 17.7 for the six MET exon 14 skipping mutation cases, and 75.7 ± 23.2 for the seven EGFR exon 20 insertion mutation cases.

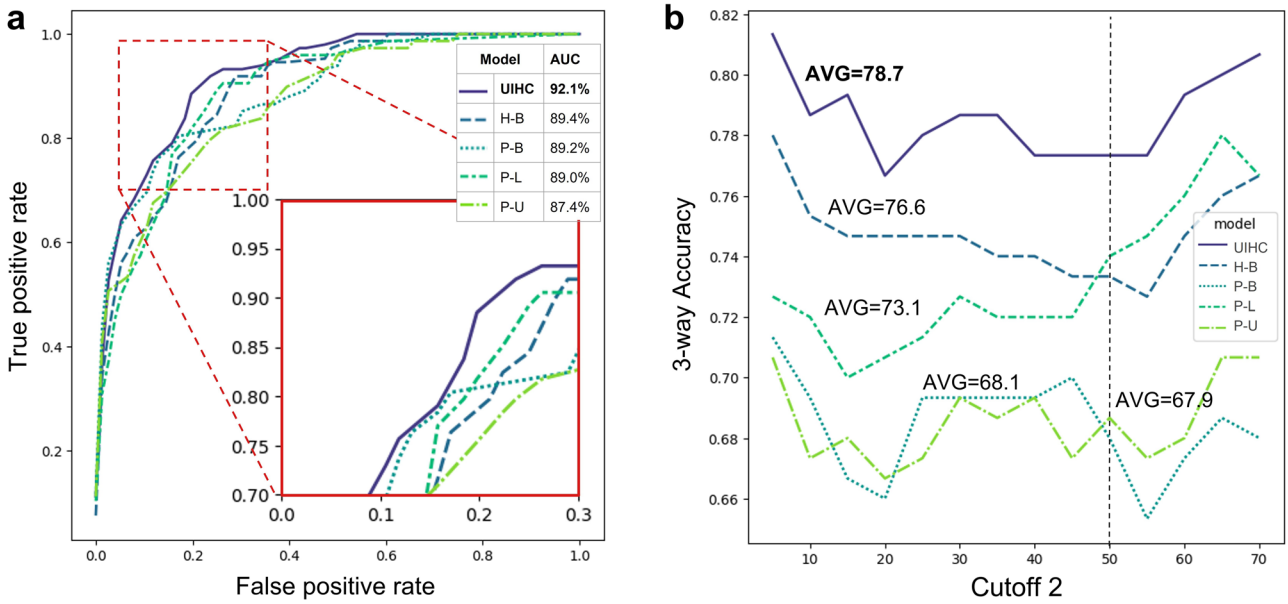

Fig. 5 | Performance analysis of the deep learning (DL) models on novel immunostains with varying interpretation cutoffs. a The receiver operating characteristic (ROC) curve by changing the cutoff over the predicted tumor proportion score (TPS) and measuring false and true positive rates. In this experiment, 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 cutoff 1 fixed at 1% and the second cutoff value within the [2%, 75%] range, illustrating the 3-way classification accuracy. Multi-stain Pan-cancer test set was used for both a and b experiments. UIHC, universal immunohistochemistry model; H-B, Human Epidermal growth factor Receptor 2 (HER2) of breast; P-B, Programmed Death-Ligand 1 (PD-L1) 22C3 of breast; P-L, PD-L1 22C3 of lung; P-U, PD-L1 22C3 of urothelium; AVG, average.

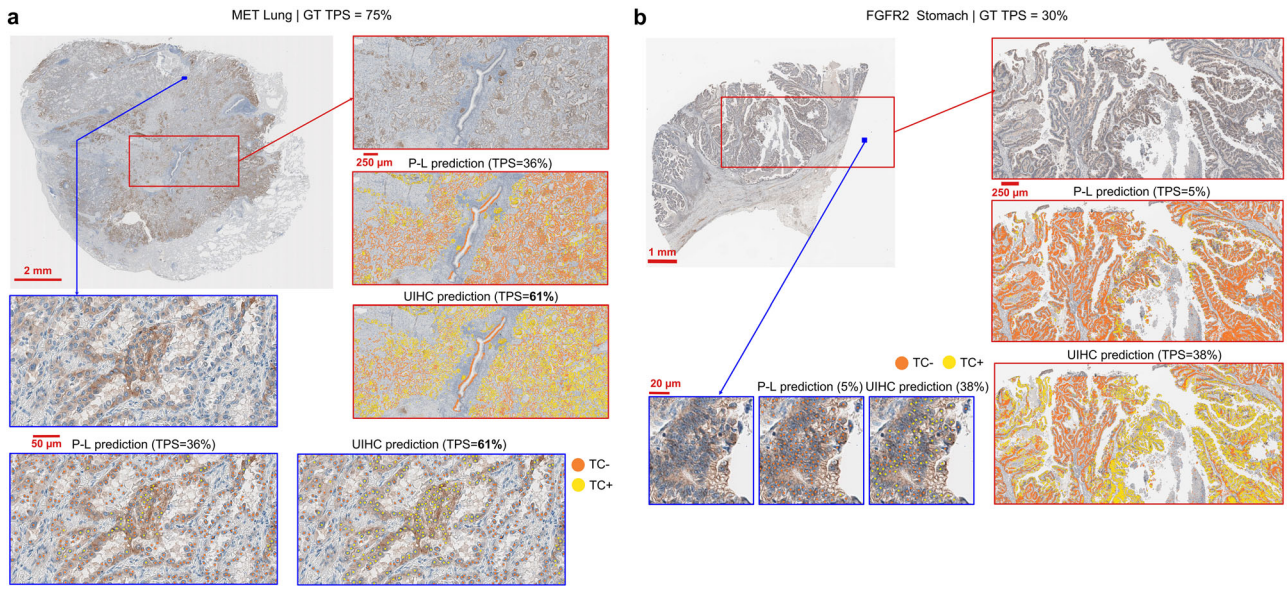

Fig. 6 | Histopathologic validation of the universal immunohistochemistry (UIHC) model. a Example of the prediction by P-L and UIHC for positively stained tumor cells (TC+) and negatively stained tumor cells (TC-) on a lung cancer whole slide image (WSI) stained with mesenchymal-epithelial transition factor (MET). The UIHC prediction yields values closer to the pathologists' ground truth TPS value, as shown in the upper. **b** Example from a gastric cancer WSI stained with fibroblast growth factor receptor 2 (FGFR2). GT TPS, ground truth tumor proportion score; P-L, Programmed Death-Ligand 1 (PD-L1) 22C3 of lung; TC tumor cell.

Discussion

In this study, we demonstrated that UIHC trained with multiple cancer types and IHC, the MC-model, is not only superior to SC-model for analyzing matched sets but also exhibited the capability to analyze never-before-seen immunostains and cancer types.

Emerging therapeutic agents that target surface proteins on tumor cells are critically important to 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)[20–22]. Recently approved drugs that exemplify this are trastuzumab deruxtecan, an ADC targeting HER2, and tarlatamab, a bispecific molecule targeting DLL3 and CD3[23–25].

IHC stands as an essential component in cancer diagnosis, and as such the pathologist's reading remains the gold standard for determining the expression level of a target protein[4,26–30]. Nonetheless, discrepancies between pathologists and poor reproducibility can hinder precise evaluation[10,15,16,31–34].

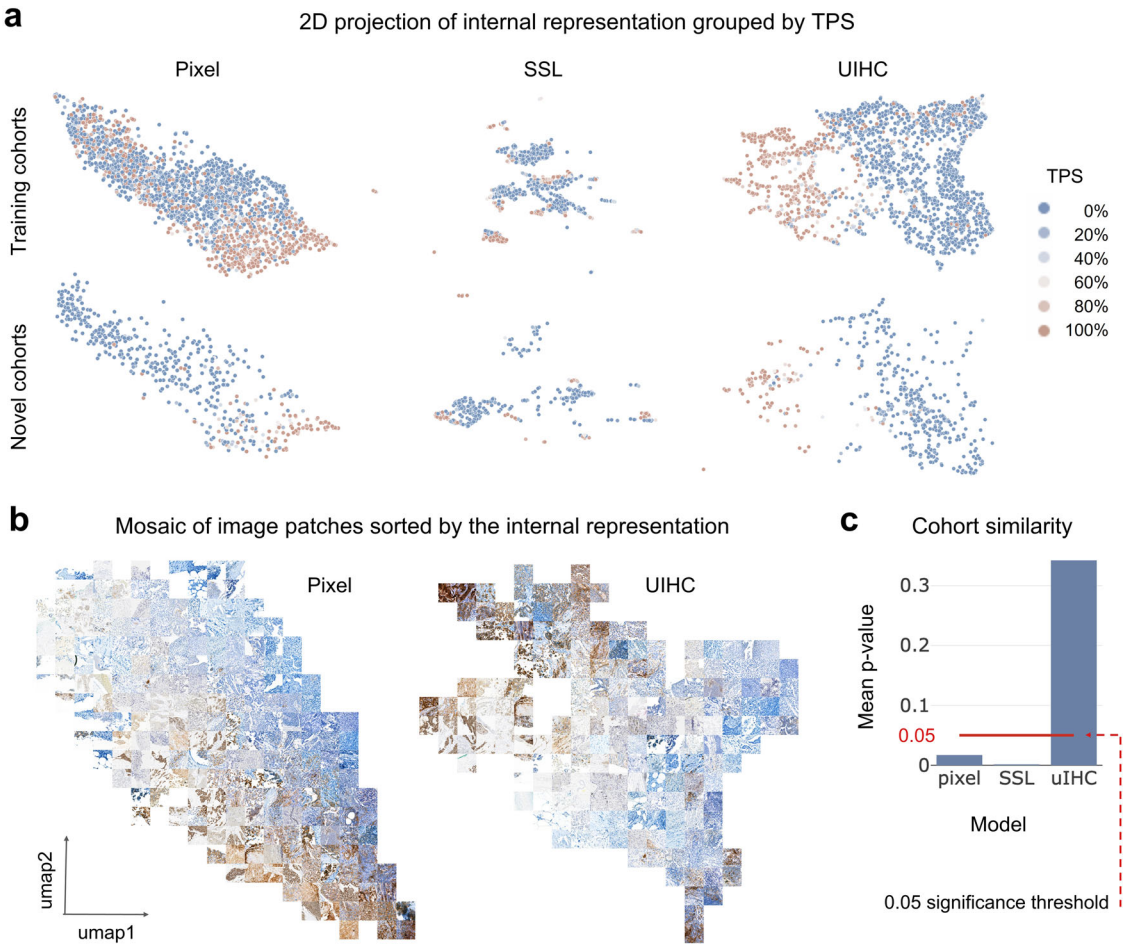

Fig. 7 | Qualitative analysis of the deep learning (DL)-learned representation. **a** Two-dimensional (2D) projection of internal representation colored by tumor proportion score (TPS). Each patch is encoded to a 2D plot using three representations: raw pixels, self-supervised learning model (SSL), and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

To develop IHC as a predictive biomarker, efforts have been made to standardize IHC and to enable quantitative evaluations. Relative to traditional computational methods for quantitative image evaluation, recently introduced DL models have exhibited notable advantages primarily due to their capacity to directly discern intricate patterns within IHC images whereas traditional methods requires pathologic analysis before computational evaluation[35,36]. These DL models can analyze PD-L1 and HER2 expression but require training on a large, manually annotated training cohort specific to each of these stains[10,11,15,16,37–39]. Moreover, such DL models have domain shift limitations 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[12,13,40].

In the present study, DL models underwent training using either a single cohort or multiple cohorts. The MC-models trained on multiple cohorts, 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 can only be attributed to the augmented training data. Compared to the H-B model, the PH-B model showed better performance on 22C3 breast and HER2 breast, indicating the impact of increasing the volume of training data. However, the superiority of PH-B over PH-LB in 22C3 urothelium, which was trained with a larger cohort than PH-B, 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 improving model performance. This phenomenon is exemplified in 22C3 Pan-cancer, where PH-LB, encompassing variations in both cancer type and immunostain, outperforms P-LUB, which only varies in cancer type, or PH-B, 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. In the case of FGFR2, some tumor cells exhibited a cytoplasmic staining pattern rather than the expected membranous staining pattern[41,42]. All models other than the UIHC model showed poor performance in TPS scoring. However, the UIHC model accurately assessed only the tumor cells displaying membranous staining, even in complex scenarios where tumor cells with cytoplasmic staining patterns were present. This was not the intended purpose of this study, so the relatively good performance of the UIHC model on cytoplasmic staining patterns will need to be validated in subsequent studies.

Recent AI-related research disciplines can be divided into the two main branches of model-centric and data-centric AI[43]. 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

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	

EGFR epidermal growth factor receptor, *MET* mesenchymal-epithelial transition factor, *TPS* tumor proportion score.

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[44–46]. However, we reserve this name for a multi-modal system that goes beyond histopathology and combines multiple medical data types[47].

To demonstrate the possible clinical utility of the current analyzer, we assessed c-MET expression in NSCLC to address a question of c-MET biology. 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)[48,49]. 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[50]. As theorized, c-MET amplifications lead to high expression of c-MET as seen in previous studies[48,51]. In contrast, tumors with METex14m had similar expression to exon 20 insertion NSCLC driven tumors. These preliminary findings should be replicated in a larger cohort but are relevant to the stratifying patients who could benefit from large molecule therapeutics targeting c-MET such as amivantanab[23].

There are some limitations of this study. For this model, detection of IHC expression detection is confined to tumor cells, but does not encompass other cell types, i.e. lymphocytes and macrophages. However, the UIHC model is capable of learning to assess these other cell types if given the training sets annotated for this cell type. 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 (data not shown)[52]. 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. Finally, this study did not directly compare the UIHC model with a model based on novel cohorts. This was due to the insufficient data available in the novel cohorts to develop robust DL models, as well as the absence of existing data from these cohorts in clinical practice, which reflects a real-world limitation.

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[35]. This innovative approach not only facilitates a broad spectrum of novel biomarker investigations but also holds the potential to assist in the development of pioneering therapeutics.

Methods

Histopathology dataset for annotation

The dataset used to develop the model consists of a total of 3046 WSIs including lung (NSCLC), urothelial carcinoma, and breast cancer cases stained for PD-L1 22C3 pharmDx IHC (Agilent Technologies, Santa Clara, CA, US) and breast cancer WSIs stained for anti-HER2/neu (4B5) (Ventana Medical Systems, Tucson, AZ, US), as reported in previous studies (Fig. 1, Supplementary Table 6)[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 0.04 mm^2 (patch, i.e. tile) is extracted.

To evaluate and compare the models, we collected patch-level test sets from ten different stain types: 22C3 (lung, urothelium, breast, liver, prostate, colorectum, stomach, biliary tract, and pancreas), HER2 (breast), SP142 (lung), various immunostain types including Claudin 18.2, DLL3, FGFR2, HER3, MET, MUC16, and TROP2 (pan-cancer). The test sets of 22C3 lung, urothelium, and breast originated from the same cohort of training and tuning sets mentioned above (internal test set in Supplementary Table 6), which could be referred to as training domain. Patches of 22C3 other than lung, urothelium, 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 was never shown to the DL model during training, we collected 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)[53–56]. Detailed

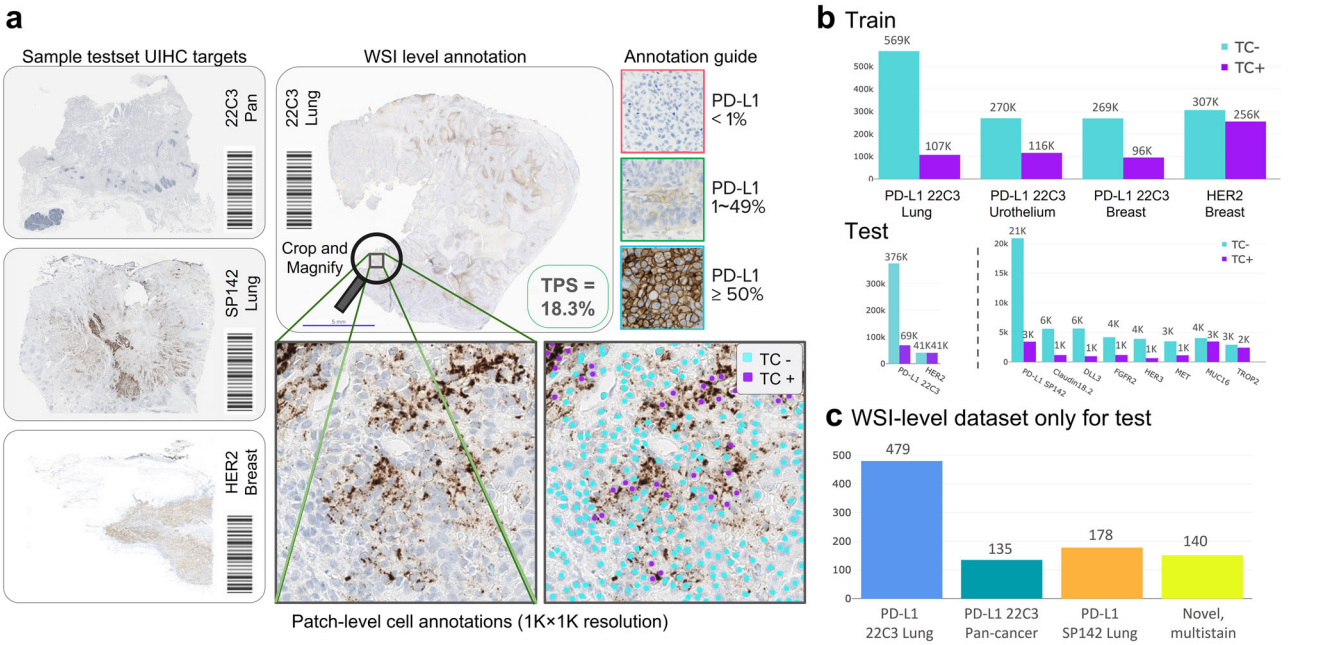

Fig. 8 | Data pipeline for Universal Immunohistochemistry (UIHC) artificial intelligence (AI) model. a Example of annotation process; patches were extracted from whole slide images (WSIs), then cells were manually annotated by expert pathologists. WSIs were split into 0.04 mm^2 patches (resized to 1024 × 1024 pixels at 0.19 microns-per pixel). **b** Patch-level annotation counted 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.

information on antibodies for various immunostain types is provided in Supplementary Table 7. 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 were selected and then resized to 1024 × 1024 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 were 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 DL model development

We defined 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)[57]. 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. Tens of pathologists marked the coordinates of where TC+ or TC- existed in a 0.04 mm^2 patch mentioned above[10,15,16]. The interpretation of tumor cell positivity by pathologists was determined by following the guidelines for PD-L1 or HER2[52,58]. 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 8, 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 9.

WSI-level test sets for DL 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[59]. Therefore, a comprehensive comparison of our model performance on WSI was conducted with the key output of WSI-level TPS[60,61].

We collected four WSI-level test sets: 22C3 lung ($n = 479$), 22C3 pan-cancer ($n = 135$), SP142 lung ($n = 178$) and a novel, multi-stain test set ($n = 140$) as presented in Fig. 8c and Supplementary Table 3. The test set containing 22C3 lung cancer was used in previous publications[10]. The 22C3 Pan-cancer contained 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 SP142 lung ($n = 178$) was derived from the same cohort of 22C3 lung cancer. IHC in the multi-stain test set included MET, MUC16, HER3, TROP2, DLL3, FGFR2, Claudin 18.2, PD-L1 SP263, and E-Cadherin across ten cancer types. Except for 22C3 lung cancer, they all corresponded to novel domains. Representative image samples from both training and novel groups are illustrated in Supplementary Figs. 4 and 5.

The multi-stain test set contains following stains: Claudin 18.2 ($n = 18$), DLL3 ($n = 16$), E-Cadherin ($n = 10$), FGFR2 ($n = 18$), HER3 ($n = 15$), MET ($n = 25$), MUC16 ($n = 16$), SP263 ($n = 10$), and TROP2 ($n = 12$) across ten cancer types (lung, breast, urothelium, cervix, colorectum, esophagus, liver, lung, melanoma, stomach). Within the multi-stain test set, except for SP263 which is applied only on lung cancer, other staining antibodies ($n = 130$) are used for: stomach ($n = 39$), urothelium (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 WSI-level test sets 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). They evaluated the TPS as a numerical scale between 0 and 100. Supplementary Table 10 describes the rate of concordant cases (all three pathologists agreed in same TPS group) among pathologists.

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 were trained with single-cohort (standard approach) or multiple-cohort data (innovation). Each model's parameters were tuned using their domain-specific tuning (validation) set. Using combinations of the above cohorts, we produced eight models as described in Fig. 2a. SC-models (H-B, P-L, P-U, and P-B) were trained on a single cohort[10,15,16]. MC-models were trained on multiple cohorts. First, cohorts from the same IHC (P-LUB) or same cancer types (PH-B) were trained. To validate the effect of mixing stain type to improve generalization, P-L and H-B were combined into one model (PH-LB). Finally, all cohorts were combined (PH-LUB). Among these candidate models, we aimed to identify the model that exhibits the highest degree of generalization for designation as a UIHC model.

Models were 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 DL models presented in this study.

Label pre-processing

Inspired by the previous work that trains the cell detection model with point annotations, we defined cell detection as a segmentation task[13,62]. At training time, we provided the cell labels as a segmentation map by drawing a disk centered on each cell point annotation. We used a fixed radius of ~ 1.3 µm, corresponding to 7 pixels at a resolution of 0.19 MPP. Finally, we assigned the value of pixels within each disk based on the class of a cell, '1' for TC−, '2' for TC+. '0' was assigned for the remaining pixels.

Inference post-processing

Given that we treated cell detection as a segmentation task, a post-processing phase was needed to extract 2D coordinates and classes of predicted cells from the probability map output by the network. We applied *skimage.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[63]. Lastly, we obtained each cell's class and probability value in the cell segmentation map through *argmax*. This probability was used as the confidence score.

Network architecture and training details

For all of our models, we used DeepLabV3+ as our base architecture with a ResNet34 encoder, which is a popular architecture specifically designed for the segmentation task[64,65]. During training, we augmented the patches with a set of standard data augmentation methods for computer vision. In particular, we utilized center crop, horizontal and vertical flip, rotation, gaussian noise, color jittering, and gray scaling. Random values were sampled for each augmentation every time an image is loaded. Network parameters were initialized by Kaiming initialization[66]. The model was optimized using the Adam optimizer[67]. Dice loss was used to train the model[68]. The initial learning rate was set to $1e{-}4$, adapted using the cosine learning rate scheduler[69]. All the models were trained for 150 epochs and evaluated at every 10 epochs to choose the best checkpoint on a hold-out tuning set. An epoch that showed the highest mF1 score on the tuning set was chosen as the best epoch and used for all evaluation purposes. All of the models were trained and evaluated with the same machine specifications as follows: 4 NVIDIA Tesla T4 GPUs each with 16GB of GPU memory and 216GB of RAM. Supplementary Fig. 6a and Supplementary Table 11 summarize the details mentioned above.

Inference details on WSIs

For WSI inference we used the full WSI for TPS calculation, excluding white background and in-house control tissue regions. The WSI was divided into 1024 × 1024 pixels of non-overlapping patches with an MPP of normalized 0.19 (following the training data), which were fed to our network, producing a prediction map with the same size as the input. All outputs were then combined to obtain a prediction map for the full WSI. Supplementary Fig. 6b presents the algorithm for TPS calculation.

Metrics for DL model performance

The performance evaluation of the DL 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 DL model. Its definition requires the number of true positives (TP), false positives (FP), and false negatives (FN) (1).

$$F1\,score = 2\,X\frac{precision\,X\,recall}{(precision + recall)} = \frac{\#TP}{\#TP + 0.5\,X\,(\#FP + \#FN)} \tag{1}$$

Since these metrics have been developed for the classification task, it is not obvious how to measure them in the context of object detection. Therefore, a hit criterion is defined as follows. For each cell class, we determined the TP, FP, and FN with the following process,

1. Sort cell predictions by their confidence score.
2. Starting from a cell prediction with the highest confidence score, check whether any GT cell is within a valid distance (25 pixels in 0.19 microns per pixel (MPP)) from the cell prediction.
   - If there is no GT cell within a valid distance, the cell prediction is counted as an FP.
   - If there are one or more GT cells within a valid distance, the cell prediction is counted as TP. The nearest GT cell is matched with the cell prediction and ignored from the further process.
3. Go back to 2. until the cell prediction with the lowest confidence score is reached.
4. The remaining GT cells that are not matched with any cell prediction are counted as FN.

After the above process, we aggregated the total number of TP, FP, and FN per cell class over all samples; then, we computed the per-class F1 score. The mean F1 (mF1) score across cell classes (TC− and TC+) was used as the final score. Each result were reproduced ×30 times using Monte-Carlo dropout at test time, an algorithm developed to study the robustness of a deep learning model over small variations[70]. In Fig. 2 and 3, for each test set we compare the statistical significance (p-value < 0.05) between the best SC-model (left of the dotted lines) with all other multiple-cohort models. p-value has been calculated using the Wilcoxon signed-rank test implemented by *scipy.stats*[71,72].

TPS is calculated by the following Eq. (2):

$$TPS = 100\,X\frac{TC+}{TC- + TC+} \tag{2}$$

PD-L1 expression is subgrouped according to the TPS cutoff 1% and 50%, i.e., classified into TPS < 1%, 1% ≤ TPS < 50%, and TPS ≥ 50%. For simplicity, TPS has been used as a general WSI-level metric to compare all IHC quantification across AI models. For each WSI, three board-certified pathologists assign a TPS score following the official protocol[60]. A category (TPS < 1%, 1% ≤TPS < 50%, or TPS ≥ 50%) is then assigned to the slide by applying the cutoff, for example for PD-L1 Lung. The WSI-level GT TPS was based on the consensus of three board-certified pathologists. For example, if two pathologists determined a WSI as TPS < 1% while one pathologist determined it as TPS 1−49%, the WSI was assigned to TPS < 1%.

The standard accuracy is computed as follows, given N number of WSIs in a test set (e.g., HER2 Breast), y and $\hat{y}$ are respectively GT and predicted category (3).

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}[yi = \hat{y}i] \tag{3}$$

The performance is also evaluated by comparing the TPS categories from pathologists to the DL model using Cohen's Kappa.

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[73]. 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 $1024 \times 1024 \times 3$ to $32 \times 32 \times 3$ 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 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[74]. 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 DL models (UIHC and SSL), each patch ran through the ResNet34 encoder producing a $16 \times 16 \times 512$ tensor of shape Height × Width × Channels. The output tensor was 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 obtained a matrix of $N \times 512$ ($N \times 3072$ for *Pixel*). Finally, we could project the 2858×512 matrix to $N \times 2$ by using UMAP, a popular non-linear dimensionality reduction algorithm[73]. This 2D matrix can be easily plotted as a scatter plot using *matplotlib* and *seaborn*. To calculate cohort similarities, we projected all patches into 1D points and collected all points from one cohort to produce a single distribution per cohort so that a statistical test can be applied. Then we computed the Wilcoxon test between all cohort pairs, producing a similarity matrix of size [# cohorts × # cohorts] containing p-values. Then we averaged the upper-triangular matrix shown in the bar plot. In addition, the mosaic of image patch was drawn by discretizing the latent representations and replacing each point with the corresponding original patch image[75]. For each discretized point in space, the median patch was 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 DL 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/).

Received: 27 June 2024; Accepted: 22 November 2024;
Published online: 03 December 2024

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* **9**, 339–354 (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. *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*. **45**, 4396–4415 (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*, **85**, 81–91 (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. Socinski, M. A. et al. *MET* Exon 14 Skipping Mutations in Non–Small-Cell Lung Cancer: An Overview of Biology, Clinical Outcomes, and Testing Considerations. JCO Precis. *Oncology* **20**, 00516 (2021).
20. 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).
21. 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* **26**, e12446 (2017).
22. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).
23. 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).
24. Ahn, M.-J. et al. Tarlatamab for Patients with Previously Treated Small-Cell Lung Cancer. *N. Engl. J. Med.* **389**, 2063–2075 (2023).
25. Cortés, J. et al. Trastuzumab Deruxtecan versus Trastuzumab Emtansine for Breast Cancer. *N. Engl. J. Med.* **386**, 1143–1154 (2022).
26. 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).
27. Sharma, P. & Allison, J. P. The future of immune checkpoint therapy. *Science* **348**, 56–61 (2015).
28. Vu, T. & Claret, F. X. Trastuzumab: updated mechanisms of action and resistance in breast cancer. *Front. Oncol.* **2**, 62 (2012).
29. 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).
30. 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).
31. 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).
32. 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).
33. Jonmarker Jaraj, S. et al. Intra- and interobserver reproducibility of interpretation of immunohistochemical stains of prostate cancer. *Virchows Arch* **455**, 375–381 (2009).
34. 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).
35. Van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
36. Rakha, E. A., Vougas, K. & Tan, P. H. Digital technology in diagnostic breast pathology and immunohistochemistry. *Pathobiology* **89**, 334–342 (2022).
37. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision* 843–852 (IEEE, 2017).
38. Mahajan, D. et al. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)* 181–196 (ECCV, 2018).
39. Kann, B. H., Hosny, A. & Aerts, H. J. Artificial intelligence for clinical oncology. *Cancer Cell* **39**, 916–927 (2021).
40. Tizhoosh, H. R. & Pantanowitz, L. Artificial intelligence and digital pathology: challenges and opportunities. *J. Pathol. Inform.* **9**, 38 (2018).
41. 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).
42. 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).
43. Zha, D. et al. Data-centric Artificial Intelligence: A Survey. *arXiv* **2303**, 10158 (2023).
44. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
45. Gerardin, Y. et al. Foundation AI models predict molecular measurements of tumor purity. *Cancer Res.* **84**, 7402–7402 (2024).
46. Campanella, G., Vanderbilt, C. & Fuchs, T. Computational Pathology at Health System Scale–Self-Supervised Foundation Models from Billions of Images. In *AAAI 2024 Spring Symposium on Clinical Foundation Models* (AAAI, 2024).
47. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
48. 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).
49. 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).
50. 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).
51. 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).
52. 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).
53. 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).
54. 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).
55. Choi, K. M. et al. The interferon-inducible protein viperin controls cancer metabolic reprogramming to enhance cancer progression. *J. Clin. Invest.* **132**, e157302 (2022).
56. 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).
57. 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).
58. Paver, E. C. et al. Programmed death ligand-1 (PD-L1) as a predictive marker for immunotherapy in solid tumours: a guide to immunohistochemistry implementation and interpretation. *Pathology* **53**, 141–156 (2021).
59. Ciga, O. et al. Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Sci. Rep.* **11**, 8894 (2021).
60. Saito, Y. et al. Inter-tumor heterogeneity of PD-L1 expression in non-small cell lung cancer. *J. Thorac. Dis.* **11**, 4982 (2019).
61. Reck, M. et al. Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **375**, 1823–1833 (2016).
62. Ryu, J. et al. OCELOT: Overlapped Cell on Tissue Dataset for Histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 23902–23912 (IEEE, 2023).
63. Van der Walt, S. et al. scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).

64. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* 801–818 (ECCV, 2018).
65. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
66. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* 1026–1034 (IEEE, 2015).
67. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* https://arxiv.org/abs/1412.6980 (2017).
68. 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. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* 240–248 (Springer, 2017).
69. Loshchilov, I. & Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv,* https://arxiv.org/abs/1608.03983 (2017).
70. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning* (PMLR) 1050–1059 (PMLR, 2016).
71. Wilcoxon, F. Individual Comparisons by Ranking Methods. In *Breakthroughs in Statistics*, 96–202 (Springer, 1992).
72. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
73. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, https://arxiv.org/abs/1802.03426 (2020).
74. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning* (PMLR) 12310–12320 (PMLR, 2021).
75. Dolezal, J. M. et al. Slideflow: Deep Learning for Digital Histopathology with Real-Time Whole-Slide Visualization. *arXiv,* https://arxiv.org/abs/2304.04142 (2023).

Acknowledgements

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

B.B., T.L., J.R., J.P., S.P., S.S., D.Y., S.M.A., K.P., C.-Y.O., and S.I.C. are employees of Lunit and/or have stock/stock options in Lunit. M.M. and S.P. were employees of Lunit.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-024-00770-z.

Correspondence and requests for materials should be addressed to Soo Ick Cho or Seokhwi Kim.