Prediction of cancer treatment response from histopathology images through imputed transcriptomics

Danh-Tai Hoang[1*], Gal Dinstag[2], Leandro C. Hermida[3,4], Doreen S. Ben-Zvi[2], Efrat Elis[2], Katherine Caley[1], Stephen-John Sammut[5,6,7], Sanju Sinha[8], Neelam Sinha[8], Christopher H. Dampier[9], Chani Stossel[10], Tejas Patil[11], Arun Rajan[12], Wiem Lassoued[13], Julius Strauss[13], Shania Bailey[13], Clint Allen[14], Jason Redman[13], Tuvik Beker[2], Peng Jiang[8], Talia Golan[10], Scott Wilkinson[15], Adam G. Sowalsky[15], Sharon R. Pine[11], Carlos Caldas[7], James L. Gulley[16], Kenneth Aldape[9], Ranit Aharonov[2], Eric A. Stone[1,*], Eytan Ruppin[8,*].

[1] Biological Data Science Institute, College of Science, Australian National University, Canberra, ACT, Australia
[2] Pangea Biomed Ltd., Tel Aviv, Israel
[3] Department of Immunology, University of Pittsburgh, Pittsburgh, PA, USA
[4] Tumor Microenvironment Center, UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA, USA
[5] Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, United Kingdom
[6] The Royal Marsden Hospital NHS Foundation Trust, London, United Kingdom
[7] Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, UK
[8] Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA
[9] Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA
[10] Oncology Institute, Sheba Medical Center at Tel-Hashomer, Tel Aviv University, Tel Aviv, Israel
[11] Division of Medical Oncology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
[12] Thoracic and GI Malignancies Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA.
[13] Center for Immuno-Oncology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA.
[14] Surgical Oncology Program, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA.
[15] Laboratory of Genitourinary Cancer Pathogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA
[16] Genitourinary Malignancy Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA.
* Corresponding authors: danhtai.hoang@anu.edu.au (D.T.H.), eric.stone@anu.edu.au (E.A.S.), and eytan.ruppin@nih.gov (E.R.)

ABSTRACT

Advances in artificial intelligence have paved the way for leveraging hematoxylin and eosin (H&E)-stained tumor slides for precision oncology. We present ENLIGHT-DeepPT, the first approach for predicting response to multiple targeted and immunotherapies directly from H&E-slides. ENLIGHT-DeepPT consists of: (1) DeepPT, a new deep-learning framework that predicts genome-wide tumor mRNA expression from slides, and (2) ENLIGHT, a recently published response prediction algorithm, applied here to predict response based on the DeepPT predicted expression values, instead of measured values. DeepPT successfully predicts transcriptomics in all 16 TCGA cohorts tested and generalizes well to two independent datasets. Importantly, ENLIGHT-DeepPT successfully predicts true responders in six independent cohorts with an overall odds ratio of 2.19, increasing the baseline response rate by 38.8% among predicted responders on average. It can provide clinicians with rapid treatment recommendations, and importantly, may open up future opportunities for advancing precision oncology in developing countries.

Statement of Significance

ENLIGHT-DeepPT is the first approach to successfully predict response to multiple targeted and immune cancer therapies directly from H&E slides, without requiring any drug-specific training data. ENLIGHT-DeepPT can provide rapid treatment recommendations without requiring tumor sequencing, bringing precision oncology to low-income countries and other situations where NGS is prohibitive.

INTRODUCTION

Histopathology has long been considered the gold standard of clinical diagnosis and prognosis in cancer. In recent years, the use of tumour molecular profiling within the clinic has allowed for more accurate cancer diagnostics, as well as the delivery of precision oncology [1-3]. Rapid advances in digital histopathology have also allowed[i] the extraction of clinically relevant information embedded in tumor slides by applying machine learning and artificial intelligence methods, capitalizing on recent advancements in image analysis via deep learning [4]. Key advances are already underway, as whole slide images (WSI) of tissue stained with hematoxylin and eosin (H&E) have been used to computationally diagnose tumors [5-8], classify cancer types [7, 9-13], distinguish tumors with low or high mutation burden [14], identify genetic

mutations [6, 15-22], predict patient survival [23-28], detect DNA methylation patterns [29] and mitoses [30], and quantify tumor immune infiltration [31]. The current study is focused on developing a new approach for predicting response to treatments from WSI images. Previous work has already impressively unravelled the potential of harnessing next-generation digital pathology to predict response to therapies directly from images [32-36]. However, the predictive models built in these studies were cancer and therapy specific. As in many other supervised learning approaches, predicting response to therapy using WSI requires large datasets comprising matched imaging and response data, though the availability of such data is still fairly limited, severely restricting the applicability of this approach.

To overcome this critical challenge, we introduce here for the first time a generic methodology for generating WSI-based predictors of patients' response for a broad range of cancer types and therapies, *which does not require matched WSI and response datasets for training*. To accomplish this, we have taken a two-step approach: First, we developed DeepPT (Deep Pathology for Transcriptomics), a novel deep-learning framework for imputing (predicting) gene expression from H&E slides, which extends upon previous valuable work on this topic [37-41]. The DeepPT models are cancer type-specific and are built by analyzing matched WSI and expression data from the TCGA. Second, given the gene expression values predicted by these models, we apply our previously published approach, ENLIGHT [42], originally developed to predict patient response from *measured* tumor transcriptomics, to predict patient response from the DeepPT *imputed* transcriptomics.

We proceed to provide an overview of DeepPT's architecture and a brief recap of ENLIGHT's workings, the study design and the cohorts analyzed. We then describe the results obtained; aligned with the two steps of ENLIGHT-DeepPT. First, we study the ability to predict tumor expression, showing the success of the trained DeepPT models in predicting the gene bulk expression in 16 TCGA cohorts and in two independent, unseen cohorts. Second, aligned with our key aim, we analyze six independent clinical trial datasets of patients with different cancer types that were treated with various targeted and immune therapies, on which DeepPT was never trained. We show that ENLIGHT, without any adaptation, can successfully predict the true responders from the expression values imputed by DeepPT, using only H&E images. Overall, our results show for the first time that combining digital pathology with an expression-based response prediction approach offers a promising new way to extend the feasibility of

precision oncology to situations where tumor sequencing is not possible, as is still the case in many underprivileged regions in the world.

Finally, and importantly, we believe that *ENLIGHT-DeepPT* may make precision oncology more accessible to patients in low- and middle-income countries (LMICs) and other under-served regions. As have been noted recently, LMICs now have increasing access to cancer medicines, primarily due to the WHO's expansion of the list of essential medicines (EMLs), and different partnerships [43]. However, these encouraging developments have not yet been matched by access to cancer diagnostics, which are obviously needed to provide precision treatments as beneficially as possible. As noted in [43], there are ongoing efforts to address this gap, including the introduction of the List of Essential In Vitro Diagnostics (EDL) and List of Priority Medical Devices for Cancer Management (PMDL), but much remains to be done, which is a key goal of our study.

RESULTS

The DeepPT/ENLIGHT computational pipeline

A. Building cancer-type specific DeepPT models and their architecture: For each cancer type, a specific DeepPT model is constructed by training on formalin-fixed, paraffin-embedded (FFPE) whole slide images and their corresponding bulk gene expression profiles from TCGA patient samples. The model obtained can then be used to predict gene expression from new WSI both for internal held-out and external datasets. In contrast to previous studies aimed at predicting gene expression from WSI, which have focused on fine tuning the last layer of pre-trained convolutional neural networks (CNN), DeepPT is composed of three main components (**Methods, Figure 1a, and Supplementary Figure S1**): a CNN model for feature extraction, an auto-encoder for feature compression, and a multi-layer perceptron (MLP) for the final regression. Intuitively, the pre-trained CNN layers (trained with natural images from the ImageNet database [44, 45]) play the role of a layperson's eyes that capture the shape and color of images, while the auto-encoder component is reminiscent of the pathologists' expertise in concentrating on the most important histological features or their combinations. The MLP regression module consists of three fully connected layers in which the weights from the input layer to the hidden layer are shared among genes, enabling the model to capture shared signal among similar gene expression profiles to benefit from the advantage of multi-task learning.

Rather than training a model for each gene separately or for all genes together as was done in previous studies [38, 40], we trained simultaneously on tranches of genes with similar median gene expression values, allowing shared signal to be leveraged while preventing the model from focusing on only the most highly expressed genes. Overall, DeepPT achieves a marked increase in the accuracy and efficiency compared to current published methods for prediction of gene expression, as will be further demonstrated.

B. Predicting patient response from DeepPT-predicted expression using ENLIGHT – an overview: The predicted expression then serves as input to ENLIGHT [42], which is a transcriptomics-based algorithm that predicts individual responses to a wide range of targeted and immunotherapies based on gene expression measured from the tumor tissue (**Figure 1b**). ENLIGHT relies on the approach of analyzing functional genetic interactions (GI) around the target genes of a given therapy, originally presented in SELECT [46]. Specifically, two broad types of interactions are considered: Synthetic Lethality (SL), whereby the simultaneous loss of two non-essential genes is detrimental to the cell, and Synthetic Rescue (SR), whereby the loss of an essential gene can be compensated for through the over- or under-expression of a second "rescuer" gene. ENLIGHT's drug response prediction generates an ENLIGHT Matching Score (EMS), which evaluates the match between patient and treatment, based on the patient transcriptomics (**Figure 1b**, [42], and **Methods** for more details).

In the forthcoming analyses, we predict patient treatment response by applying ENLIGHT to the expression values predicted by DeepPT instead of those measured using RNA sequencing. We show that combining ENLIGHT with DeepPT enables robust prediction of response to treatment in independent test sets. Remarkably, this is done without adapting either DeepPT, which was trained once on TCGA samples and then 'frozen', or ENLIGHT, which is not trained on any treatment response data in general, and specifically, on none of these test clinical cohorts.

Study design and patient cohorts

The workflow describing the computational analysis performed in this study is depicted in **Figure 1c**:

(1) First, to build cancer type specific DeepPT models, we collected FFPE WSI together with matched RNAseq gene expression profiles for 16 cancer types from TCGA, in 10

broad classes: breast (BRCA), lung (LUAD and LUSC), brain (LGG and GBM), kidney (KIRC, KIRP and KICH), colorectal (COAD and READ), prostate (PRAD), gastric (STAD and ESCA), head and neck (HNSC), cervical (CESC), and pancreas (PAAD). Low quality slides were excluded, resulting in 6,269 slides from 5,528 patients. Each cancer type was processed, trained and evaluated separately. We performed a 5×5 nested cross-validation to assess the model performance: In the outer loop, the samples were randomly split into five disjoint sets. Each sample set was selected in turn as the *held-out test set* (20%), while the rest were used for training (80%). Given an outer split, in the inner loop each model was trained five times by further splitting the training set into internal *training* and *validation* sets, performing a five-fold cross validation. Applying a bagging technique, we averaged the predictions from the five different models, presenting our final prediction for each gene on the held-out test set. The outer loop was repeated five times across the five held-out test sets, resulting in 25 trained models (**Supplementary Figure S2a-b**).

(2) Then, we performed external testing of DeepPT. We applied the trained models of the respective cancer types to predict gene expression on twoexternal datasets: the TransNEO breast cancer cohort (TransNEO-Breast) consisting of 160 fresh frozen (FF) slides [47], and an unpublished brain cancer cohort (NCI-Brain) consisting of 210 FFPE slides. Both datasets contained matched expression data (see **Methods**).

(3) Our final goal is to use the predicted gene expression to predict patients' treatment response. To this end, we further applied the DeepPT models to predict gene expression on six additional datasets (see Methods and **Table S1** for full details). Predicted gene expressions for each of the six datasets served as input to ENLIGHT to predict treatment outcome.

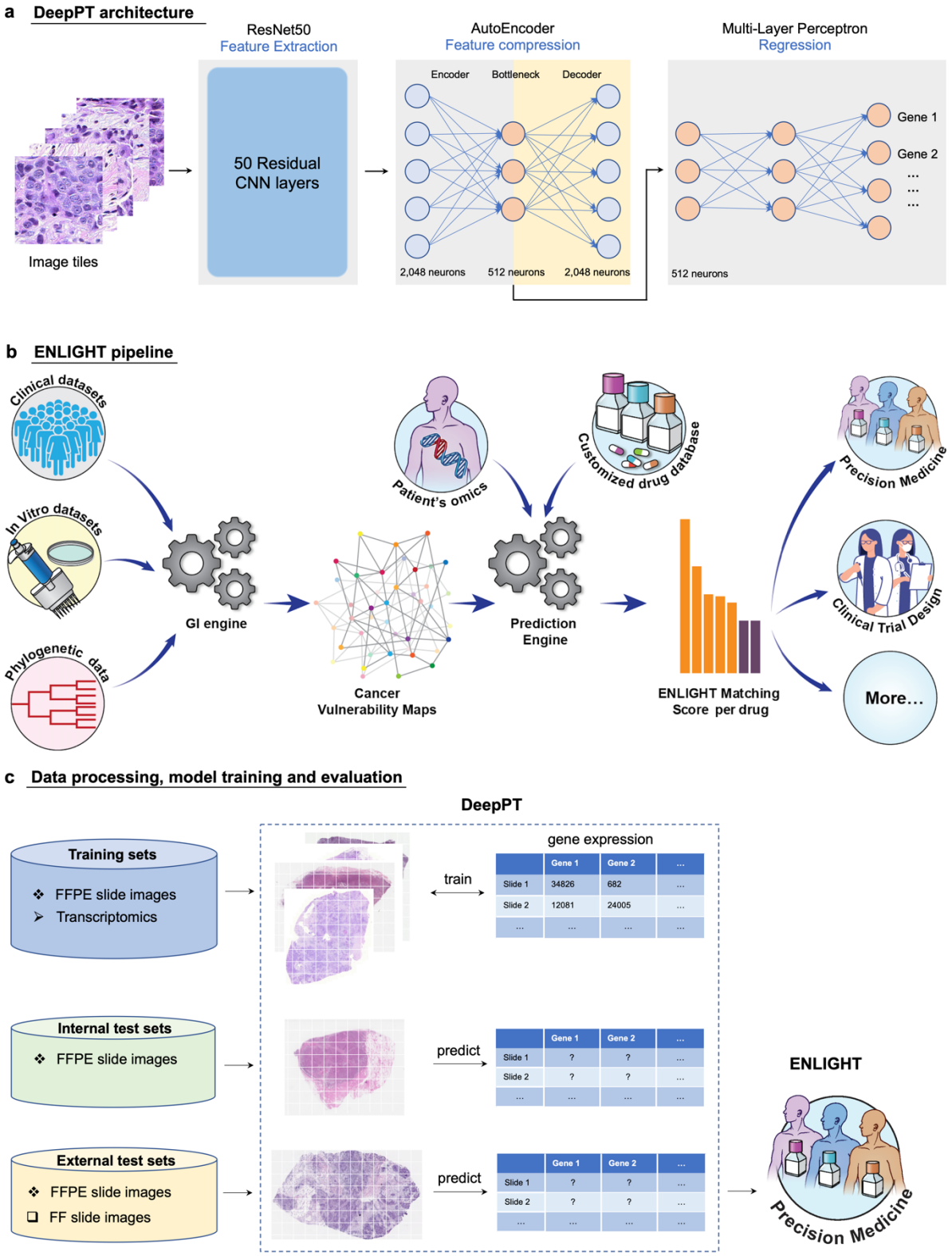

Figure 1. Study overview. (a) The three main components of DeepPT architecture, from left to right. The pre-trained ResNet50 CNN unit extracts histopathology features from tile images. The autoencoder compresses the 2,048 features to a lower dimension of 512 features. The multi-layer perceptron integrates these histopathology features to predict the sample's gene expression. **(b)** An overview of the ENLIGHT pipeline (illustration taken from [42]: ENLIGHT

starts by inferring the genetic interaction partners of a given drug from various cancer in-vitro and clinical data sources. Given the SL and SR partners and the transcriptomics for a given patient sample, ENLIGHT computes a drug matching score that is used to predict the patient response. Here, ENLIGHT uses DeepPT predicted expression to produce drug matching scores for each patient studied. **(c)** DeepPT was trained with formalin-fixed paraffin-embedded (FFPE) slide images and matched transcriptomics obtained from ten broad cancer types from the TCGA. After the training phase, the models were applied to predict gene expression on the internal (held-out) TCGA datasets and on six independent datasets on which they were never trained. The predicted tumor transcriptomics in each of these six test clinical datasets served as input to ENLIGHT for predicting the patients' response to treatment and assessing the overall prediction accuracy.

Prediction of gene expression from histopathology images

As illustrated in **Figure 1**, we constructed models predicting normalized gene expression profiles from their corresponding histopathology images for each of 10 broad TCGA cancer classes. We then applied the trained models to predict gene expression of internal held-out test sets in each of the cancer types, using five-fold cross validation. In most cancer types, thousands of genes were significantly predicted, with Holm-Sidak corrected p-values < 0.05, outperformed the recently published state-of-the-art expression prediction approach, HE2RNA by Schmauch et al [40], in most cohorts considered in this study (**Figure 2a**). To further evaluate model performance, we estimated the Pearson correlation (R) between predicted and actual expression values of each gene across the test dataset samples. In most cancer types, thousands of genes had a correlation above 0.4 (**Figure 2c and Supplementary Figure S3**). For breast cancer, for instance, DeepPT predicted 1,812 genes with mean correlation greater than 0.4, more than doubling the number of genes predicted at that correlation level reported by HE2RNA, which was 786 genes [40], testifying further to the increased accuracy of DeepPT.

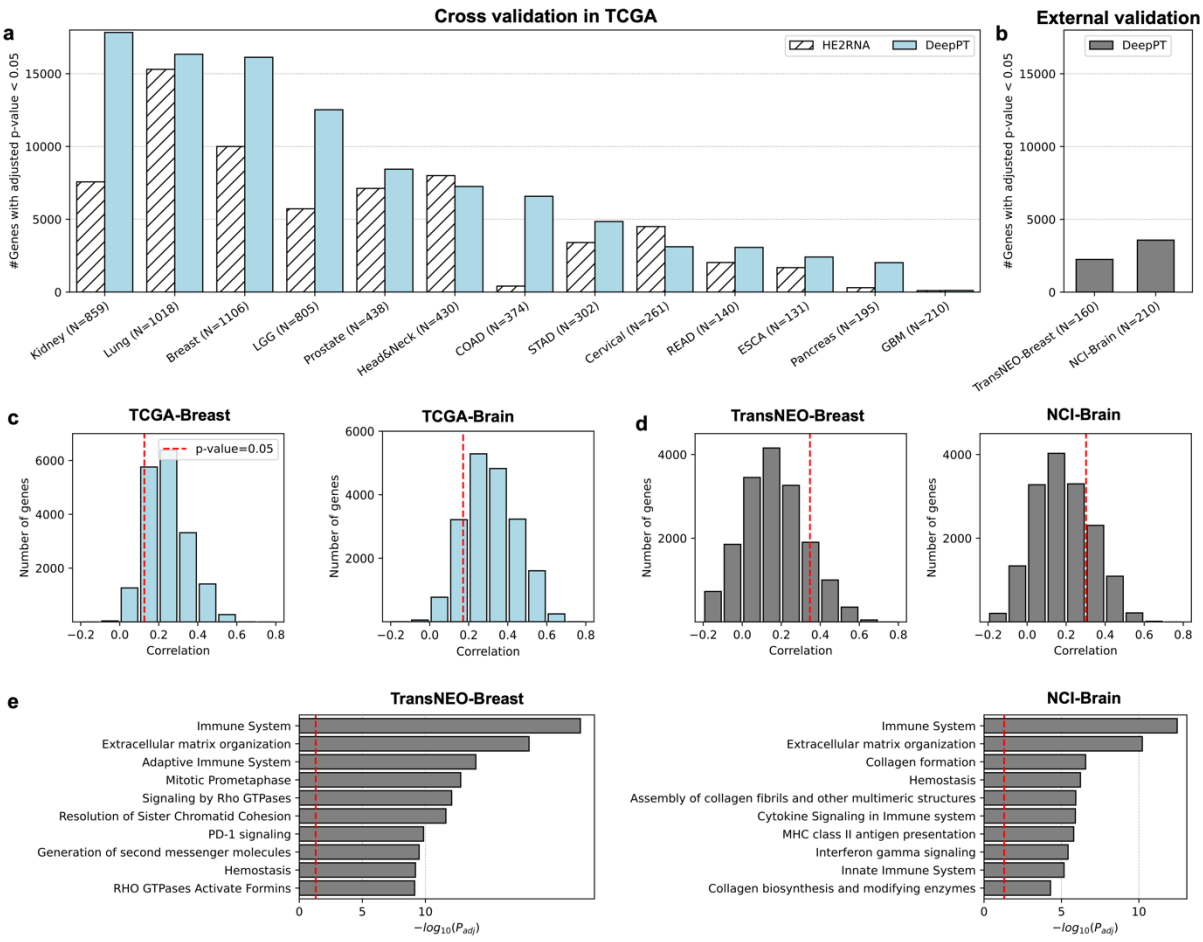

Figure 2. DeepPT prediction of gene expression from H&E slides. (a, b) Number of significantly predicted genes for each cohort, in comparison with the current state-of-the-art method, HE2RNA. For comparison against HE2RNA, the performance of each cancer subtype of brain (LGG and GBM), colorectal (COAD and READ), and gastric (STAD and ESCA) cohort are shown independently as in [40]. The performance of HE2RNA were extracted from [40]. **(c, d)** Histograms of the Pearson correlation coefficients between predicted and actual expressions for each gene in the two external validation sets. Here a DeepPT model was trained on the respective TCGA cohort and used to predict expressions for the denoted dataset. Red dashed lines represent the correlation coefficient level beyond which the results are significant (p-value < 0.05 after correction for multiple hypotheses testing). **(e)** Top enriched pathways among the well-predicted genes (R > 0.4) for the TransNEO-Breast and NCI-Brain datasets, inferred from trained DeepPT models on TCGA-Breast and TCGA-Brain, respectively. Red dashed lines represent p-value = 0.05. We then tested the prediction ability of DeepPT on two unseen independent datasets available to us, which contained matched tumor WSIs and expression. We applied the DeepPT model constructed using the TCGA-Breast cancer dataset with no further training or modifications to predict gene expression from corresponding H&E

slides of the TransNEO cohort (n=160). Notably, the two datasets were generated independently at different facilities, with two different preparation methods (TCGA slides are FFPE while TransNEO slides are FF), so the histological features extracted from these two datasets are distinct (**Supplementary Figure S4**). Despite these differences, without any further training, we found 1,489 genes that had fairly strong correlations between the predicted and measured expression values across the samples, R > 0.4 (**Figure 2d)**. Similarly, we applied the DeepPT model trained on TCGA-Brain samples to predict gene expression from unpublished NCI-Brain slides (n=210), and found that 1,324 genes had correlations of R > 0.4 across samples (**Figure 2d)**. This testifies to the considerable predictive power of DeepPT.

Of interest, a gene set enrichment analysis using the GSEApy python package [48] and Reactome database [49] of the well-predicted genes (R > 0.4) reveals an enrichment of cellular functions known to play a key role in oncogenesis, such as cell cycle, mitosis, and extracellular matrix organization. Many enriched pathways are specific to different cancer types, pointing to an interesting tumor specificity of the pathways whose expression is associated with H&E features (**Figure 2e and Supplementary Figure S5**). These results suggest that pathways associated with oncogenic processes may be more strongly associated with visible changes in the H&E slides, though the nature of these associations is tumor-specific.

Predicting treatment response from DeepPT-imputed gene expression

As described earlier, we developed ENLIGHT-DeepPT, a two-step approach to predict patients' response from WSI without any further training. Given a previously unseen tumour slide image, we first apply a pre-trained, cancer-type specific DeepPT model (built by training on the TCGA, as described above) to predict the tumor transcriptomics. Second, based on this predicted gene expression, we apply our published precision oncology algorithm, ENLIGHT [42], to predict the patient's response.

We tested the ability of ENLIGHT-DeepPT to accurately predict patient response in six clinical cohorts, treated with various targeted and immunotherapies, from which we have managed to obtain access to patients slides and response data. Those include two HER2+ breast cancer patient cohorts treated with chemotherapy plus Trastuzumab [47, 50], a BRCA+ pancreatic cancer cohort treated with PARP inhibitors (Olaparib or Veliparib), a mixed indications cohort of Lung, Cervical and H&N patients treated with Bintrafusp alfa [51], a bi-

specific antibody that targets TGFB and PDL1, a prostate cancer cohort of patients treated with neoadjuvant androgen receptor deprivation (NADT) plus enzalutamide [52, 53], and finally, an ALK+ NSCLC cohort treated with ALK inhibitors (Alectinib or Crizotinib). For each dataset, the response definition was determined by the clinicians running the respective trial (see **Methods** and **Table S2** for more details). As ENLIGHT does not predict response to chemotherapies, only the targeted agents were considered for response prediction.

For each cohort, we used the DeepPT model previously trained on the appropriate TCGA cohort, *without any changes and with no further training*, to predict the gene expression values from the H&E slide of each patient's pre-treated tumor. We then applied ENLIGHT to these predicted gene expression values to produce ENLIGHT Matching Scores (EMS) based on the genetic interaction network of the given drug [42]. Importantly, we do not restrict the GI network to include only genes with strong correlations between the actual and predicted expression values; This is done as ENLIGHT considers the combined effect of a large set of genes, averaging out the noise of individual gene expression prediction, leading to quite robust response predictions as we show below. The prediction accuracy of ENLIGHT-DeepPT on each of these six datasets individually and in aggregate is shown in **Figure 3**.

Since the ENLIGHT-DeepPT workflow is designed with clinical applications in mind, we focus our assessment of its predictive power on measures that have direct clinical importance – namely the response odds ratio of the test and its average precision. **Figure 3a** depicts the odds ratio (OR) obtained for each dataset studied; the OR denotes the ratio of the odds to respond among patients receiving ENLIGHT-matched treatments vs. the odds to respond among patients whose treatments were not ENLIGHT-matched. Patients were considered ENLIGHT-matched if their EMS scores were greater or equal to a threshold value of 0.54. This threshold was determined already in the original ENLIGHT publication [42] on independent data and was kept fixed here. Using this predefined threshold, we observe that the OR of ENLIGHT-DeepPT is higher than 1 for all but the *ARi* dataset, demonstrating that patients receiving ENLIGHT-matched treatments indeed had a higher chance to respond.

Complementing the OR measure, which is of major translational interest as it quantifies the performance at a specific decision threshold, **Figure 3b** further depicts the prediction performance of ENLIGHT-DeepPT via a complementary measure, the average precision (AP) (precision is also termed positive predictive value). AP is a measure of the precision of a prediction model across the entire range of thresholds. For a classifier to be of merit, its AP

should be higher than the overall response rate (ORR), which denotes the fraction of responders observed in each cohort. Reassuringly, the AP of ENLIGHT-DeepPT well exceeds the ORR for five of six datasets, testifying to its broader predictive power beyond that quantified by the OR.

Turning to an aggregate analysis of the performance of ENLIGHT-DeepPT, by analysing all patients together in a simulated "basket" trial in which each patient receives a different treatment (n = 266), the OR of ENLIGHT-DeepPT is 2.19 ([1.26,3.82] 95% CI, left bar in **Figure 3a**), significantly higher than 1 (p = 0.004, Fisher's exact test), and its precision is 47.9%, a 38.8% increase compared to the overall response rate of 34.5% ($p = 2.13 \times 10^{-6}$, one sided proportional test). The AP is 0.468 (left bar, **Figure 3b**), significantly higher than the baseline response rate (p = 0.001, one sided permutation test).

For one of the datasets (Trastuzumab_1), tumor gene expression measured using RNA sequencing was also available, and was previously analyzed by ENLIGHT [42]. **Figures 3c and d** compare the predictive performance using ENLIGHT-DeepPT scores to that using ENLIGHT scores calculated based on the measured expression values (denoted ENLIGHT-actual). Using the previously established threshold of 0.54, the OR of ENLIGHT-DeepPT is 3.94, which is lower than the OR of 6.14 obtained by ENLIGHT-actual, but still significantly higher than expected by chance (p = 0.02, test for OR > 1). The PPV of ENLIGHT-DeepPT was 53.3%, surprisingly higher than the PPV of 52% when using ENLIGHT-actual, and both almost double the basic ORR of 29.7% observed in this study. However, the sensitivity (the fraction of responders correctly identified) of ENLIGHT-DeepPT is markedly lower than that of ENLIGHT-actual, 42.1% vs. 68.%. **Figure 3d** also compares ENLIGHT-actual and ENLIGHT-DeepPT performance to that of the multi-omic machine learning predictor that uses DNA, RNA and clinical data, published by Sammut et al. [47], denoted here as Sammut-ML. The latter was based on *in-cohort supervised learning* to predict response to chemotherapy with or without trastuzumab among HER2+ breast cancer patients. In both analyses, we applied each method to the patients for which the relevant data was available: 65 patients had RNAseq data (ENLIGHT-actual), 64 had H&E slides (ENLIGHT-DeepPT), and 56 had RNAseq, DNAseq and clinical features needed to run the Sammut-ML predictor. To systematically compare between the predictors across a wide range of decision thresholds, and since Sammut et al. did not derive a binary classification cutoff, we used AP as the comparative rod here. As can be seen, all methods have quite comparable predictive power, with ENLIGHT-DeepPT having the

highest AP. Importantly, using only H&E slides without need for RNA or DNA data or other clinical features has a major invaluable practical advantage.

The above results demonstrate that ENLIGHT-DeepPT is robust to measurement noise, enabling expression predictions on unseen data in an unsupervised manner, without further training. Notably, the predictions for Trastuaumab$_1$ were made on FF tissue slides, which differ considerably from FFPE samples used to train the DeepPT model, further testifying to the robustness of DeepPT and ENLIGHT.

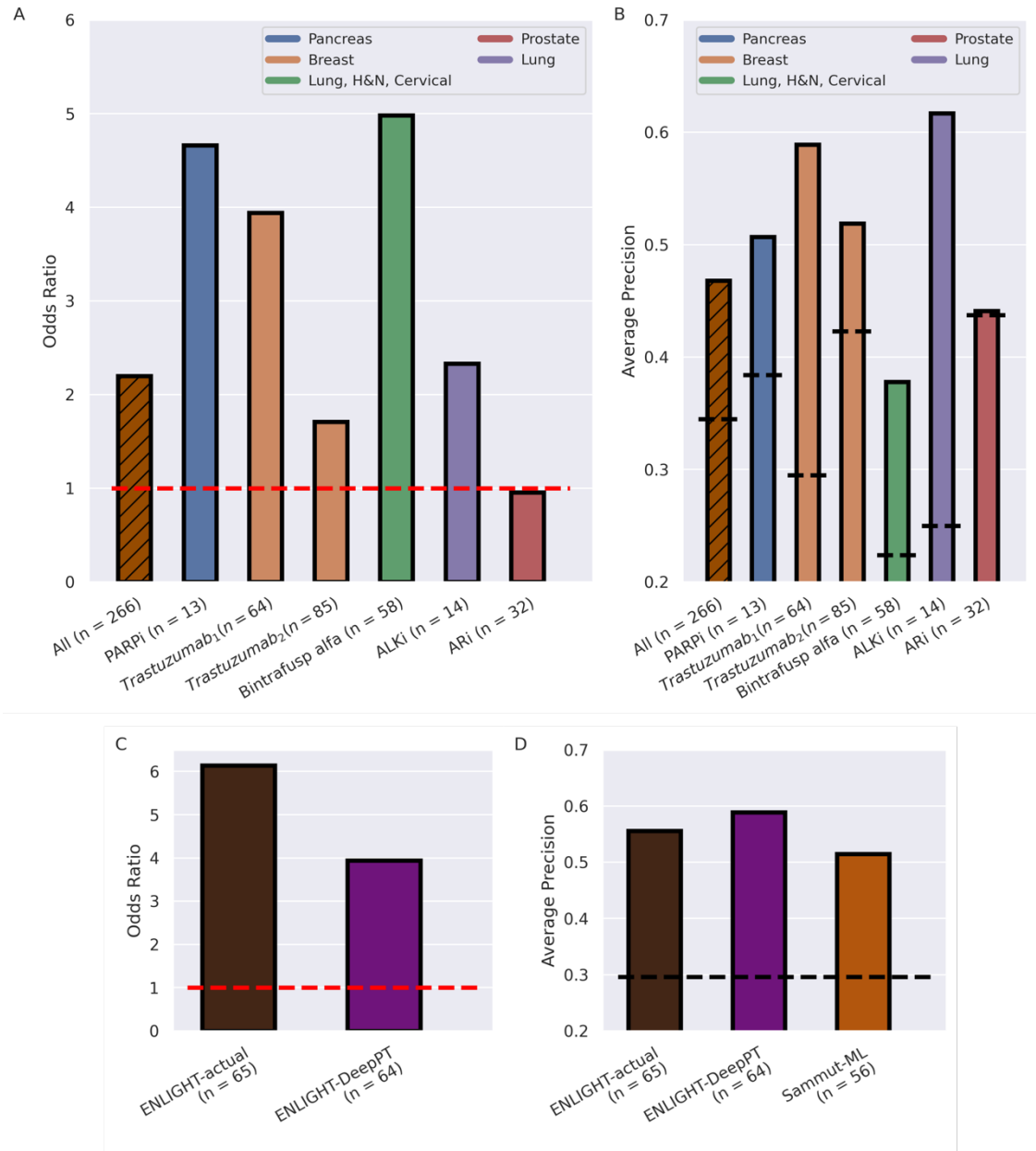

Figure 3. Predicting treatment response from H&E slides. (a) Odds Ratio (OR, Y axis) for the six datasets tested (X axis). Drug and sample sizes are denoted in the X axis labels. Red horizontal dashed line denotes an OR of 1 which is expected by chance. Bars are colour coded

according to the indication(s) of the respective cohort. **(b)** Average Precision (AP, Y axis) for the six datasets, as in **a**. Black horizontal dashed lines denote the ORR for each dataset. An AP higher than the ORR demonstrates better accuracy than expected by chance. **(c)** OR for ENLIGHT-actual and ENLIGHT-DeepPT when predicting response to Trastuzumab among HER2+ breast cancer patients. **(d)** Comparison of AP (Y axis) for both ENLIGHT based models and the Sammut-ML predictor of Sammut et al. [47]. Each method was applied to the patients for which relevant data is available (see main text). Black horizontal dashed line denotes the ORR.

DISCUSSION

Our study demonstrates that combining DeepPT, a novel deep learning framework for predicting gene expression from H&E slides, with ENLIGHT, a published unsupervised computational approach for predicting patient response from pre-treated tumor transcriptomics, could be used to form a new *ENLIGHT-DeepPT* approach for H&E-based prediction of clinical response to a host of targeted and immune therapies. We show that DeepPT significantly outperforms the current state-of-the-art method in predicting mRNA expression profiles from H&E slides. More importantly, *ENLIGHT-DeepPT* successfully predicts the true responders in several clinical datasets from different indications, treated with a variety of targeted drugs directly from the H&E images, demonstrating its potential clinical utility throughout.

Combining DeepPT with ENLIGHT is a promising approach for predicting response directly from H&E slides because it does not require response data on which to train. This is a crucial advantage compared to the more common practice of using response data to train classifiers in a supervised manner. Indeed, while sources like TCGA lack response data that would enable building a supervised predictor of response to targeted and immune treatments, applying ENLIGHT to predicted expression has successfully enabled the prediction of response to four different treatments in five datasets spanning five cancer types with considerable accuracy, without the need for any treatment data for training.

DeepPT is fundamentally different from previous computational pipelines for gene expression prediction both in its model architecture and in its training strategy. We attribute its superior performance to four key innovations: (i) All previous studies fed output from the conventional pre-trained CNN model (trained with natural images from ImageNet database) directly into their regression module, whereas we added an auto-encoder to re-train the output

of the pre-trained CNN model. This helps to familiarize the model with histological features, to exclude noise, avoid over fitting and reduce the computational demands. (ii) Our regression module is a non-linear MLP model in which the weights from the input layer to the hidden layer are shared among genes. This architecture enables the model to exploit the correlations between the expression of the genes. (iii) We trained together sets of genes with similar median gene expression values; doing so further implements a form of multitask learning and prevents the model from focusing on only the most highly expressed genes. (iv) We performed ensemble learning by taking the mean predictions across all models. This further improves the prediction accuracy quite significantly (**Supplementary Figure S2**).

DeepPT can be broadly applied to all cancer types for which H&E slides are available; however, similar to many other deep learning models, it requires a considerable number of training samples comprising matched imaging and gene expression. An interesting direction for future work would be to apply transfer learning between cohorts, to improve the predictive performance in cancer types with yet small training cohorts. In other words, it might be possible to train the model on large TCGA cohorts such as breast and lung cancer, then fine tune it for generating predictions for cancer types with smaller TCGA cohorts such as melanoma or ovarian cancer.

ENLIGHT is an unsupervised approach that leverages large-scale data in cancer to infer genetic interaction partners associated with drug targets, and then uses their expression levels to generate a matching score for each possible treatment. The ENLIGHT matching score is translated to a binary response prediction using a single pan-cancer pan-treatment threshold, above which a treatment is considered a beneficial match to the patient ("ENLIGHT-matched"). ENLIGHT was originally developed and validated using measured RNA expression data and was applied here *as-is* to the expression predicted by DeepPT. We demonstrated the predictive power and robustness of this combined approach (ENLIGHT-DeepPT) in five independent test sets. In one case where data was available, ENLIGHT-DeepPT's prediction accuracy was comparable to that achieved by ENLIGHT (applied to measured expression data) and by a supervised classifier trained for this specific task.

A striking finding of this study is the robustness of response predictions based on H&E slides when combining DeepPT and ENLIGHT. First, despite the inevitable noise introduced by the prediction of gene expression, the original ENLIGHT inferred GI partners did not require

any modifications to predict response here from the DeepPT-predicted expression. Second, though DeepPT was trained using FFPE slides, it generalized well and could be used as-is to predict expression values from FF slides. This demonstrates the applicability of DeepPT for predicting RNA expression either from FF or from FFPE slides. Nevertheless, as promising as the results presented here are, they should of course be further tested and expanded upon by applying the generic pipeline presented here to many more cancer types and treatments.

There are several limitations of our study that should be noted: (i) as explained above, while DeepPT can predict expression reliably for many genes based on a relatively small number of samples, it is still reliant on sufficient training data and future studies should aim to enlarge the number of genes whose expression is accurately predicted. (ii) ENLIGHT-DeepPT should be more extensively validated on additional indications and treatments, as data is accumulated. (iii) Here we used ENLIGHT's clinical decision threshold as is from [42]. Further research is required to fine tune the decision threshold for ENLIGHT-DeepPT once its application to a broader set of cohorts becomes feasible.

Developing a response prediction pipeline directly from H&E slides, if reasonably accurate and further carefully tested and validated in clinical settings, could obviously be of utmost benefit, as NGS results often take 4-6 weeks after initiation to return a result. Many patients who have advanced cancers require treatment immediately, and this method can potentially offer treatment options within a shorter time frame. Moreover, obtaining H&E images can be done at relatively low cost, compared to the expenses incurred by NGS. Increasing efforts to harness the rapid advances in deep learning are likely to improve precision oncology approaches, including by leveraging histopathology images. Most current studies aimed at predicting patient response have been focused on predicting genomic alterations; to date, we are not aware of studies aiming at predicting patient response to therapy from the predicted tumor expression based on histological images. Given its general and unsupervised nature, we are hopeful that *ENLIGHT-DeepPT* may have considerable impact, making precision oncology more accessible to patients in Low- and middle-income countries (LMICs), in under-served regions and in other situations where sequencing is less feasible. Specifically, affordable cancer diagnostics is critical in LMICs since their limited access to cancer diagnostics is a bottleneck for effectively leveraging the increasing access to cancer medicine [43]. While promising, one should of course cautiously note that the results presented in this

study await a broader testing and validation before they may be applied in the clinic. We are hopeful that the results presented here will expedite such efforts by others going forward.

METHODS

Data collection

- TCGA histological images and their corresponding gene expression profiles were downloaded from GDC (https://portal.gdc.cancer.gov). Only diagnostic slides from primary tumors were selected, making a total of 6,269 formalin-fixed paraffin-embedded (FFPE) slides from 5,528 patients with breast cancer (1,106 slides; 1,043 patients), lung cancer (1,018 slides; 927 patients), brain cancer (1,015 slides; 574 patients), kidney cancer (859 slides; 836 patients), colorectal (514 slides; 510 patients), prostate (438 slides; 392 patients), gastric (433 slides; 410 patients), head & neck (430 slides; 409 patients), cervical (261 slides; 252 patients), pancreatic cancer (195 slides; 175 patients).
- The TransNEO-Breast dataset consists of fresh frozen slides and their corresponding gene expression profiles from 160 breast cancer patients. Full details of the RNA library preparation and sequencing protocols, as well as digitisation of slides have been previously described [47].
- The NCI-Brain histological images and their corresponding gene expression profiles were obtained from archives of the Laboratory of Pathology at the NCI, and consisted of 210 cases comprising a variety of CNS tumors, including both common and rare tumor types. All cases were subject to methylation profiling to evaluate the diagnosis, as well as RNA-sequencing.
- The Bintrafusp alfa treated cohort consisted of 58 patients with lung cancer (9 patients), cervical cancer (16 patients), and head and neck cancer (33 patients). FFPE slides were made available from the NCI.
- The Trastuzumab_1 cohort is a subset of TransNEO-Breast dataset mentioned above, consisting of 64 patients who had received a combination of chemotherapy and Trastuzumab.

- Trastuzumab$_2$, a HER2+ breast cancer cohort treated with a combination of Trastuzumab and chemotherapy, consisted of 85 patients and their FFPE slides [36, 50]. FFPE slides were downloaded from The Cancer Imaging Archive database (TCIA, https://www.cancerimagingarchive.net).
- The ARi dataset consisted of 32 prostate cancer patients treated with neoadjuvant androgen receptor deprivation (NADT) therapy plus enzalutamide [52, 53]. FFPE slides were downloaded from the "NADT-Prostate" dataset of The Cancer Imaging Archive database (TCIA, https://www.cancerimagingarchive.net).
- The ALKi dataset consisted of 14 NSCLC ALK mutated patients treated with Crizotinib or Alectinib. Corresponding FFPE slides were made available from Colorado University.
- The PARPi dataset consisted of 13 germline BRCA mutated pancreatic cancer patients treated with PARP inhibitors. FFPE slides were made available from Sheba Medical Center.

For each dataset, the classification of patients to responders and non-responders was based on the criterion used by the clinicians running the respective trial: for Trastuzumab$_1$ and Trastuzumab$_2$, response was defined as pCR while non-responders were defined as RD. For Bintrafusp alfa, responders were defined as patients with partial response or complete response. For ARi, response was defined as patient residual cancer burden of less than 0.05. For ALKi, response was defined as more than 18 months progression free survival. For PARPi, response was defined as more than 36 months overall survival. Full details can be found in **Table S2**.

Histopathology image processing

We first used Sobel edge detection [54] to identify areas containing tissue within each slide. Because the WSI are too large (from 10,000 to 100,00 pixels in each dimension) to feed directly into the deep neural networks, we partitioned the WSI at 20x magnification into non-overlapping tiles of 512 x 512 RGB pixels. Tiles containing more than half of the pixels with a weighted gradient magnitude smaller than 10 were removed. Depending on the size of slide, the number of tiles per slide in TCGA cohort varied from 100 to 8,000 (**Supplementary Figure S6**). In contrast, TransNEO slides for example are much smaller, resulting in 100 to 1,000 tiles per slide (**Supplementary Figure S6e**). To minimize staining variation (heterogeneity and batch effects), color normalization was applied to the selected tiles.

Gene expression processing

Gene expression profiles were obtained as read counts for approximately 60,483 gene identifiers. Genes considered expressed were identified using edgeR, resulting in approximately 18,000 genes for each cancer type. The median expression across samples of each gene varied from 10 to 10,000 reads for each dataset (**Supplementary Figure S7**). To reduce the range of gene expression values, and to minimize discrepancies in library size between experiments and batches, a normalization was performed as described in our previous work [42].

DeepPT architecture

Our model architecture was composed of three main units (**Supplementary Figure 1**).

(1) Feature extraction: The pre-trained ResNet50 CNN model, trained with 14 million natural images from the ImageNet database [34] was used to extract features from image tiles. Before feeding these tiles into the ResNet50 unit, the image tiles were resized to 224 x 224 pixels to match the standard input size for the convolutional neural network. Through the feature extraction process, each input tile is represented by a vector of 2,048 derived features.

(2) Feature compression: We applied an autoencoder, which consists of a bottleneck of 512 neurons, to reduce the number of features from 2,048 to 512. This helps to familiarise the model with histological features, to exclude noise, to avoid over fitting, and finally to reduce the computational demands.

(3) Multi-Layer Perceptron (MLP) regression: The purpose of this component is to build a predictive model linking the aforementioned auto-encoded features to whole-genome gene expression. The model consists of three layers: (1) an input layer with 512 nodes, reflecting the size of the auto-encoded vector; (2) a hidden layer whose size depends on the number of genes under shared consideration; and (3) an output layer with one node per gene. The rationale behind this architecture is to leverage similarity among the genes under shared consideration, as captured by the weights connecting the input layer to the hidden layer. The weights connecting the hidden layer to the output layer model the subsequent relationship between the hidden layer and each individual gene. This follows the philosophy of multi-task learning. If the prediction of each gene's expression level represents a single task, then our strategy is to first group these tasks for shared learning, followed by optimization of each individual task. In

our default whole-genome approach, we bin genes into groups of 4,096 whose median expression levels are similar, and we use 512 hidden nodes. Because the training data is comprised of gene expression at the slide level (i.e. bulk gene expression, as opposed to at spatial resolution), we average our per-tile predictions to obtain bulk values at the slide level.

DeepPT training and evaluation

We trained and evaluated each cancer type independently. To evaluate our model performance, we applied 5x5 nested cross-validation. For each outer loop, we split the entire patients (of each cohort) into training (80%) and held-out test (20%) set. We further split the training set into internal training and evaluation set, according to five-fold cross validation. The models were trained and evaluated independently with each pair of training/validation sets. Averaging the predictions from the five different models represents our final prediction for each single gene on each held-out test set. We repeated this procedure five times across the five held-out test sets, making a total of 25 trained models. These models trained with TCGA cohorts were used to predict the expression of each gene in a given external cohort by computing the mean over the predicted values of all models.

As noted in the Model Architecture section, tranches of genes with similar median expression levels were grouped for simultaneous training and evaluation. This was done to optimize model performance and model efficiency, and contrasts with approaches in the literature that either train on each gene separately [38] or on all genes together [40]. Each training round was stopped at a maximum of 500 epochs, or sooner if the average correlation per gene between actual and prediction values of gene expression on the validation set did not improve for 50 continuous epochs. The Adam optimizer with mean squared error loss function was employed in both auto-encoder and MLP models. A learning rate of 10^{-4} and a minibatches of 32 image tiles per step were used for both the auto-encoder model and MLP regression model. To reduce overfitting, a dropout of 0.2 was used.

Data augmentation

Because the number of samples in the TCGA-Pancreas cohort is relatively small, to further reduce overfitting, we artificially increased the amount of data for this cohort by rotating the whole slide images by 90°, 180°, 270°. During the test time, the average of four symmetries represents our prediction for each slide. No data augmentation was performed for other cohorts.

Implementation details

All analysis in this study was performed in Python 3.9.7 and R 4.1.0 with the libraries including Numpy 1.20.3, Pandas 1.3.4, Scikit-learn 1.1.1, Matplotlib 3.4.3, and edgeR 3.28.0. Image processing including tile partitioning and color normalization was conducted with OpenSlide 1.1.2, OpenCV 4.5.4, PIL 8.4.0. The histopathology feature extraction was carried out using TensorFlow 2.8.0. The feature compression (autoencoder unit) and MLP regression parts were implemented using PyTorch 1.12.0. Pearson correlation was calculated using Scipy 1.5.0. Gene set enrichment analysis was performed with Gseapy 0.10.7. The identification of highly expressed genes was performed with edgeR 3.28.0.

ENLIGHT

ENLIGHT's drug response prediction comprises of two steps: (i) Given a drug, the *GI engine* identifies the clinically relevant genetic interaction (GI) partners of the drug's target gene(s). The GI engine first identifies a list of initial candidate Synthetic lethal/Synthetic rescue (SL/SR) interactions by analyzing cancer cell line dependencies based on the principle that SL/SR interactions should decrease/increase tumor cell viability, respectively, when 'activated' (e.g., in the SL case, viability is decreased when both genes are under-expressed). It then selects those pairs that are more likely to be clinically relevant by analysing a database of tumor samples with associated transcriptomics and survival data, requiring a significant association between the joint inactivation of target and partner genes and better patient survival for SL interactions, and analogously for SR interactions. (ii) The drug-specific GI partners are then used to predict a patient's response to each drug based on the gene expression profile of the patient's tumor. The ENLIGHT Matching Score (EMS), which evaluates the match between patient and treatment, is based on the overall activation state of the set of GI partner genes of the drug targets, deduced from the gene expression, reflecting the notion that a tumor would be more susceptible to a drug that induces more active SL interactions and fewer active SR interactions.

We applied ENLIGHT in its original version as described in [42]. The only modification made to the original ENLIGHT version is the exclusion of the component that, in the case of monoclonal antibodies, considers the expression of the drug target itself, as this component

highly weighs a single gene and is hence much more susceptible to errors in the prediction of expression values.

Data and Code Availability

TCGA data are available from https://portal.gdc.cancer.gov. TCIA data is available from https://www.cancerimagingarchive.net/. All DeepPT predicted expressions and relevant response data, along with the code to calculate performance measures are available in Github: https://github.com/PangeaResearch/enlight-deeppt-data. ENLIGHT scores given expression profiles (either measured directly from the tumor or predicted from slides) can be calculated using a web service at ems.pangeabiomed.com. The DeepPT prediction pipeline will be made accessible upon acceptance.

Acknowledgments

This work was supported by the Australian Research Council (ARC) (D.T.H, E.A.S) and by the Intramural Research Program of the National Institutes of Health (NIH), National Cancer Institute (NCI), Center for Cancer Research (CCR) (S.S, N.S, E.R). This work utilized the super computational resources of the Australian National Computational Infrastructure (AUNCI) and the Australian National University Merit Allocation Scheme (ANUMAS).

Declaration of interests

G.D, D.S.B, E.E, T.B, and R.A are employees of Pangea Biomed. E.R. is a co-founder of Medaware, Metabomed, and Pangea Biomed (divested from the latter). E.R. serves as a non-paid scientific consultant to Pangea Biomed under a collaboration agreement between Pangea Biomed and the NCI.

References

1. Golub, T.R., et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.* Science, 1999. **286**(5439): p. 531-537.
2. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, 2012. **486**(7403): p. 346-52.
3. Doroshow, D.B. and J.H. Doroshow, *Genomics and the History of Precision Oncology.* Surg Oncol Clin N Am, 2020. **29**(1): p. 35-49.
4. Rosenthal, J., et al., *Building Tools for Machine Learning and Artificial Intelligence in Cancer Research: Best Practices and a Case Study with the PathML Toolkit for Computational Pathology.* Molecular Cancer Research, 2022. **20**(2): p. 202-206.

5. Strom, P., et al., *Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study.* Lancet Oncol, 2020. **21**(2): p. 222-232.
6. Fu, Y., et al., *Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis.* Nature Cancer, 2020. **1**(8): p. 800-810.
7. Noorbakhsh, J., et al., *Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images.* Nat Commun, 2020. **11**(1): p. 6367.
8. Kim, J., et al., *Unsupervised discovery of tissue architecture in multiplexed imaging.* Nature Methods, 2022. **19**(12): p. 1653-1661.
9. Coudray, N., et al., *Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning.* Nat Med, 2018. **24**(10): p. 1559-1567.
10. Swiderska-Chadaj, Z., et al., *Learning to detect lymphocytes in immunohistochemistry with deep learning.* Med Image Anal, 2019. **58**: p. 101547.
11. Yu, K.-H., et al., *Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks.* Journal of the American Medical Informatics Association : JAMIA, 2020. **27**(5): p. 757-769.
12. Couture, H.D., et al., *Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype.* NPJ Breast Cancer, 2018. **4**: p. 30.
13. Campanella, G., et al., *Clinical-grade computational pathology using weakly supervised deep learning on whole slide images.* Nat Med, 2019. **25**(8): p. 1301-1309.
14. Xu, H., et al., *Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients.* bioRxiv, 2019: p. 554527.
15. Kather, J.N., et al., *Pan-cancer image-based detection of clinically actionable genetic alterations.* Nat Cancer, 2020. **1**(8): p. 789-799.
16. Qu, H., et al., *Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning.* NPJ Precis Oncol, 2021. **5**(1): p. 87.
17. Schaumberg, A.J., M.A. Rubin, and T.J. Fuchs, *H&E-stained Whole Slide Image Deep Learning Predicts SPOP Mutation State in Prostate Cancer.* bioRxiv, 2018: p. 064279.
18. Tsou, P. and C.J. Wu, *Mapping Driver Mutations to Histopathological Subtypes in Papillary Thyroid Carcinoma: Applying a Deep Convolutional Neural Network.* J Clin Med, 2019. **8**(10).
19. Chang, P., et al., *Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas.* AJNR Am J Neuroradiol, 2018. **39**(7): p. 1201-1207.
20. Kather, J.N., et al., *Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer.* Nature Medicine, 2019. **25**(7): p. 1054-1056.
21. Kim, R.H., et al., *A Deep Learning Approach for Rapid Mutational Screening in Melanoma.* bioRxiv, 2019: p. 610311.
22. Chen, M., et al., *Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning.* npj Precision Oncology, 2020. **4**(1): p. 14.
23. Mobadersany, P., et al., *Predicting cancer outcomes from histology and genomics using convolutional networks.* Proc Natl Acad Sci USA, 2018. **115**(13): p. E2970-e2979.
24. Cheng, J., et al., *Integrative Analysis of Histopathological Images and Genomic Data Predicts Clear Cell Renal Cell Carcinoma Prognosis.* Cancer Res, 2017. **77**(21): p. e91-e100.
25. Beck, A.H., et al., *Systematic analysis of breast cancer morphology uncovers stromal features associated with survival.* Sci Transl Med, 2011. **3**(108): p. 108ra113.
26. Bulten, W., et al., *Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study.* Lancet Oncol, 2020. **21**(2): p. 233-241.
27. Courtiol, P., et al., *Deep learning-based classification of mesothelioma improves prediction of patient outcome.* Nat Med, 2019. **25**(10): p. 1519-1525.

28. Boehm, K.M., et al., *Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer.* Nature Cancer, 2022. **3**(6): p. 723-733.
29. Zheng, H., et al., *Whole slide images reflect DNA methylation patterns of human tumors.* NPJ Genom Med, 2020. **5**: p. 11.
30. Wang, H., et al., *Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features.* J Med Imaging (Bellingham), 2014. **1**(3): p. 034003.
31. Turkki, R., et al., *Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples.* J Pathol Inform, 2016. **7**: p. 38.
32. Wang, X., et al., *Spatial interplay patterns of cancer nuclei and tumor-infiltrating lymphocytes (TILs) predict clinical benefit for immune checkpoint inhibitors.* Science Advances, 2022. **8**(22): p. eabn3966.
33. Hu, J., et al., *Using deep learning to predict anti-PD-1 response in melanoma and lung cancer patients from histopathology images.* Translational Oncology, 2021. **14**(1): p. 100921.
34. Johannet, P., et al., *Using Machine Learning Algorithms to Predict Immunotherapy Response in Patients with Advanced Melanoma.* Clinical Cancer Research, 2021. **27**(1): p. 131-140.
35. Zhang, F., et al., *Predicting treatment response to neoadjuvant chemoradiotherapy in local advanced rectal cancer by biopsy digital pathology image features.* Clin Transl Med, 2020. **10**(2): p. e110.
36. Farahmand, S., et al., *HER2 and trastuzumab treatment response H&E slides with tumor ROI annotations [Data set]*. 2022: The Cancer Imaging Archive.
37. He, B., et al., *Integrating spatial gene expression and breast tumour morphology via deep learning.* Nat Biomed Eng, 2020. **4**(8): p. 827-834.
38. Wang, Y., et al., *Predicting Molecular Phenotypes from Histopathology Images: A Transcriptome-Wide Expression-Morphology Analysis in Breast Cancer.* Cancer Res, 2021. **81**(19): p. 5115-5126.
39. Monjo, T., et al., *Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation.* bioRxiv, 2021: p. 2021.04.22.440763.
40. Schmauch, B., et al., *A deep learning model to predict RNA-Seq expression of tumours from whole slide images.* Nat Commun, 2020. **11**(1): p. 3877.
41. Levy-Jurgenson, A., et al., *Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer.* Sci Rep, 2020. **10**(1): p. 18802.
42. Dinstag, G., et al., *Clinically oriented prediction of patient response to targeted and immunotherapies from the tumor transcriptome.* Med, 2022.
43. Erfani, P., et al., *Leveraging Molecular Diagnostic Technologies to Close the Global Cancer Pathology Gap.* JCO Global Oncology, 2022(8): p. e2200182.
44. He, K., et al. *Deep Residual Learning for Image Recognition*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
45. Deng, J., et al. *ImageNet: A large-scale hierarchical image database*. in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009.
46. Lee, J.S., et al., *Synthetic lethality-mediated precision oncology via the tumor transcriptome.* Cell, 2021. **184**(9): p. 2487-2502.e13.
47. Sammut, S.-J., et al., *Multi-omic machine learning predictor of breast cancer therapy response.* Nature, 2022. **601**(7894): p. 623-629.
48. Fang, Z. *Gene Set Enrichment Analysis in Python*. 2021 Nov 18 [cited 2021 Dec 20]; Available from: https://pypi.org/project/gseapy/.

49. Jassal, B., et al., *The reactome pathway knowledgebase.* Nucleic Acids Res, 2020. **48**(D1): p. D498-d503.
50. Farahmand, S., et al., *Deep learning trained on hematoxylin and eosin tumor region of Interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer.* Modern Pathology, 2022. **35**(1): p. 44-51.
51. Strauss, J., et al., *Bintrafusp alfa, a bifunctional fusion protein targeting TGF-β and PD-L1, in patients with human papillomavirus-associated malignancies.* Journal for ImmunoTherapy of Cancer, 2020. **8**(2): p. e001395.
52. Wilkinson, S., et al., *Nascent Prostate Cancer Heterogeneity Drives Evolution and Resistance to Intense Hormonal Therapy.* European Urology, 2021. **80**(6): p. 746-757.
53. Karzai, F., et al., *Sequential Prostate Magnetic Resonance Imaging in Newly Diagnosed High-risk Prostate Cancer Treated with Neoadjuvant Enzalutamide is Predictive of Therapeutic Response.* Clinical Cancer Research, 2021. **27**(2): p. 429-437.
54. Kanopoulos, N., N. Vasanthavada, and R.L. Baker, *Design of an image edge detection filter using the Sobel operator.* IEEE Journal of Solid-State Circuits, 1988. **23**(2): p. 358-367.

SUPPLEMENTARY MATERIAL

Cohort	#Slides	#Patients
TCGA-Breast	1,106	1,043
TCGA-Lung	1,018	927
TCGA-Brain	1,015	574
TCGA-Kidney	859	836
TCGA-Colorectal	514	510
TCGA-Prostate	438	392
TCGA-Gastric	433	410
TCGA-Head&Neck	430	409
TCGA-Cervical	261	252
TCGA-Pancreas	195	175
TransNEO-Breast	160	160
NCI-Brain	210	210

Table S1. Number of slides and number of patients for each cohort.

Name	Indication	Drug	Background Therapy	N	# of responders	Responders criteria
PARPi	Pancreas	Olaparib or Veliparib		13	5	OS > 36 months
Trastuzumab$_1$	Breast	Trastuzumab	Chemotherapy	64	19	pCR
Trastuzumab$_2$	Breast	Trastuzumab	Chemotherapy	85	36	pCR
Bintrafusp alfa	Lung, H&E, Cervical	Bintrafusb alfa		58	13	CR, PR
ARi	Prostate	androgen receptor deprivation (ADT) therapy plus enzalutamide		32	14	RCB < 0.05 cm^3

Table S2. Details of the 5 datasets analyzed using ENLIGHT-DeepPT. OS - overall survival; pCR - pathological complete response; RD - residual disease; CR – complete response; PR – partial response; RCB – residual cancer burden

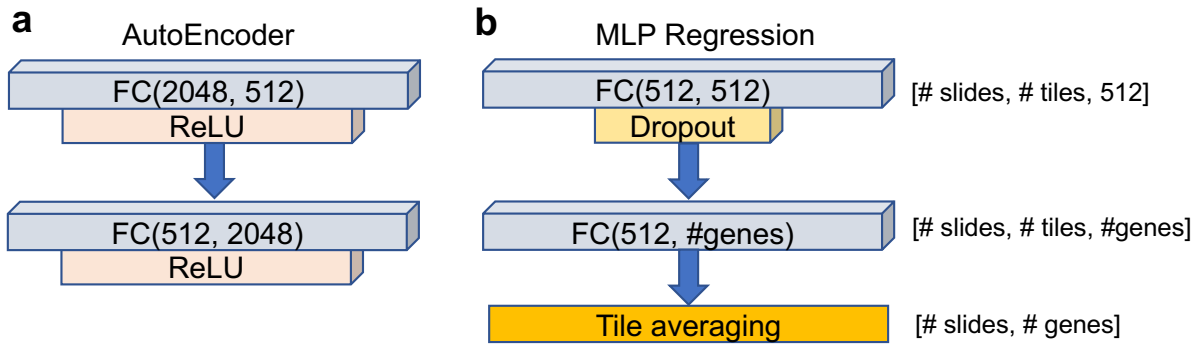

Figure S1. Model architecture in detail. (a) The feature compression subnetwork consists of an input layer of 2,048 neurons, a bottleneck of 512 neurons, and an output layer of 2,048 neurons. (b) The MLP regression subnetwork consists of an input layer of 512 neurons, a hidden layer of 512 neurons, and an output layers with the number of neurons reflecting the number of genes in each group.

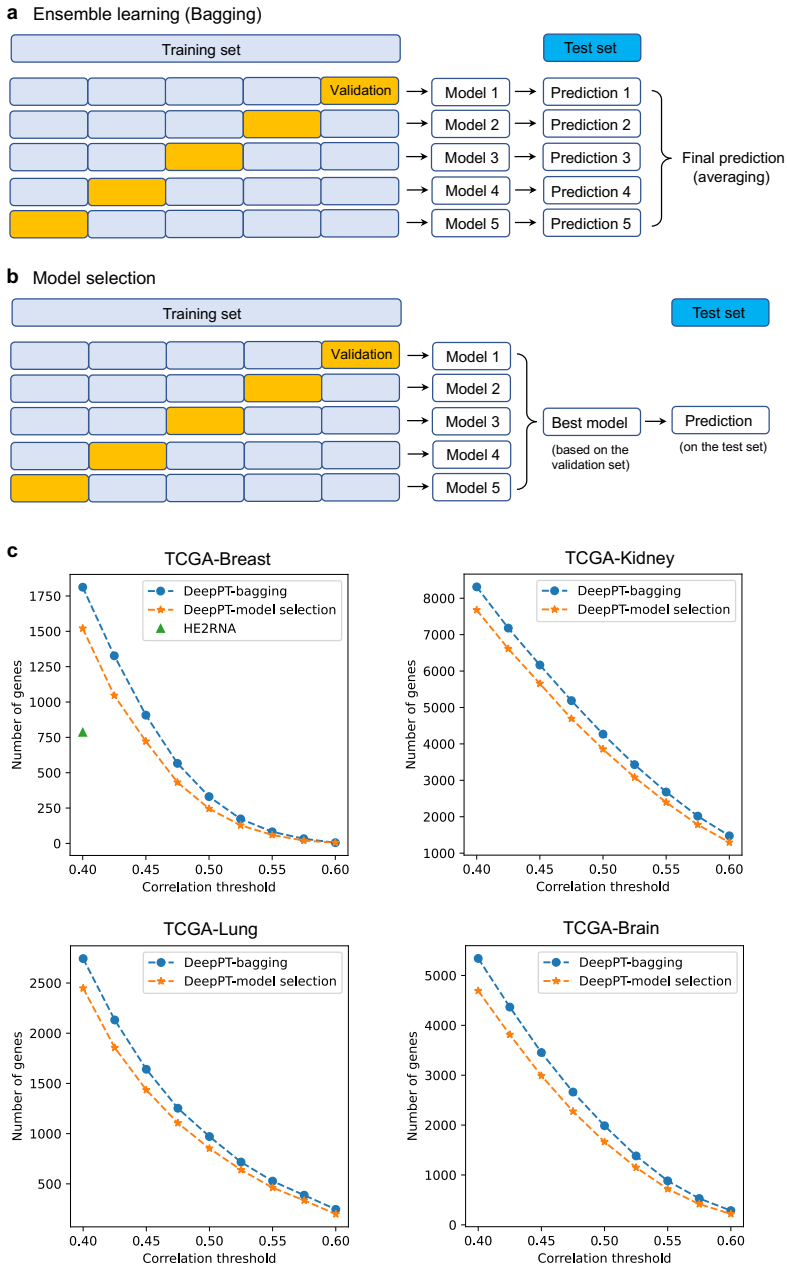

Figure S2. Training strategies and their performance. (a) In the ensemble learning strategy (bagging), five models were trained independently with five internal training-validation splits; these five model predictions were averaged to make the final prediction. **(b)** In the model selection strategy, the "best" model with the highest performance on the validation set was chosen to make prediction on the test set. **(c)** Number of genes with mean correlation over 5 folds greater than a certain threshold, obtained from these strategies. Note that with either strategy, DeepPT outperforms the current state-of-the-art approach, HE2RNA.

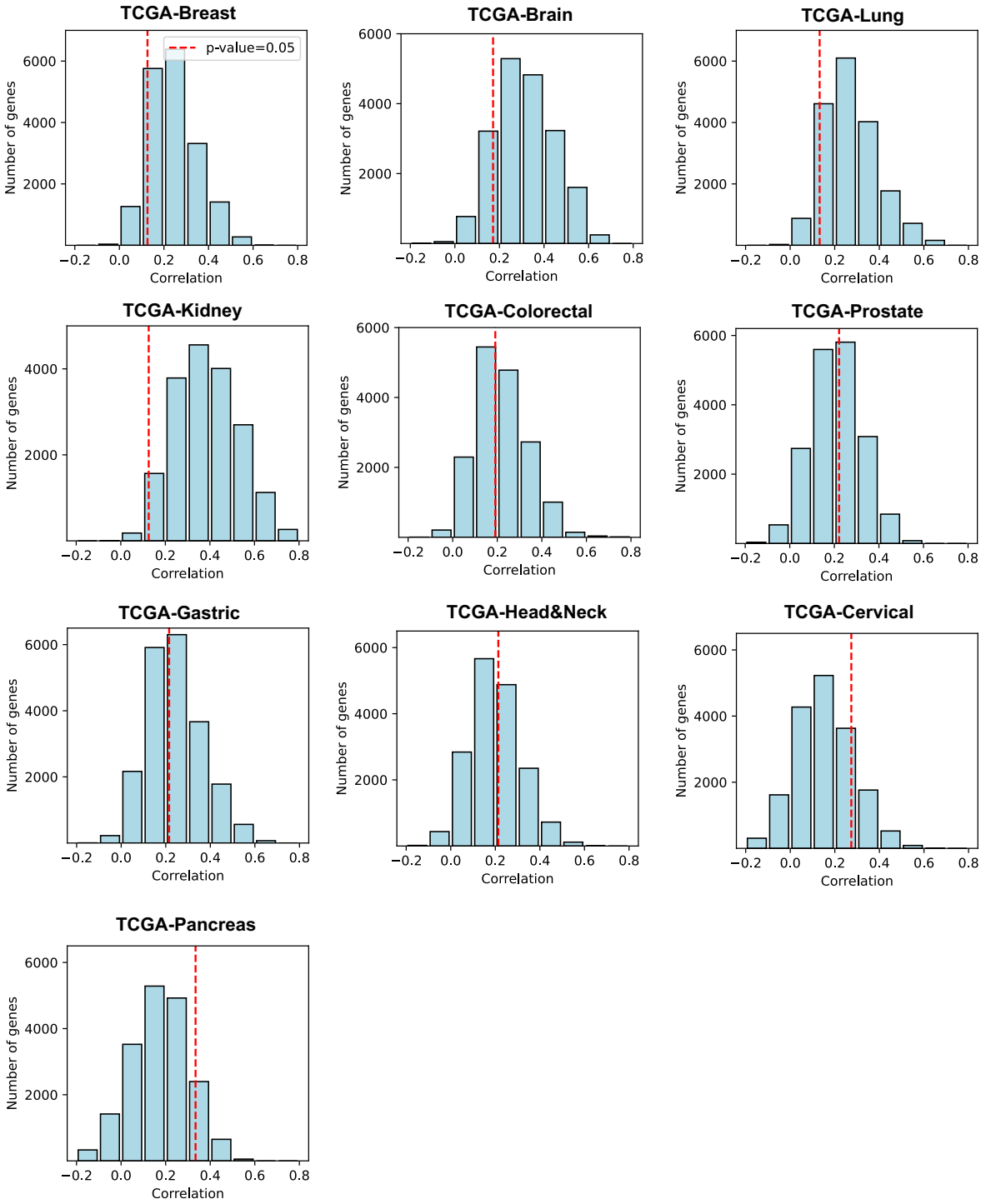

Figure S3. Histograms of the Pearson correlation coefficients between predicted and actual expression for each gene across test sets, for each cancer type cohort. Red dashed lines represent the correlation coefficient level beyond which the results are significant (p-value < 0.05 after correction for multiple hypotheses testing).

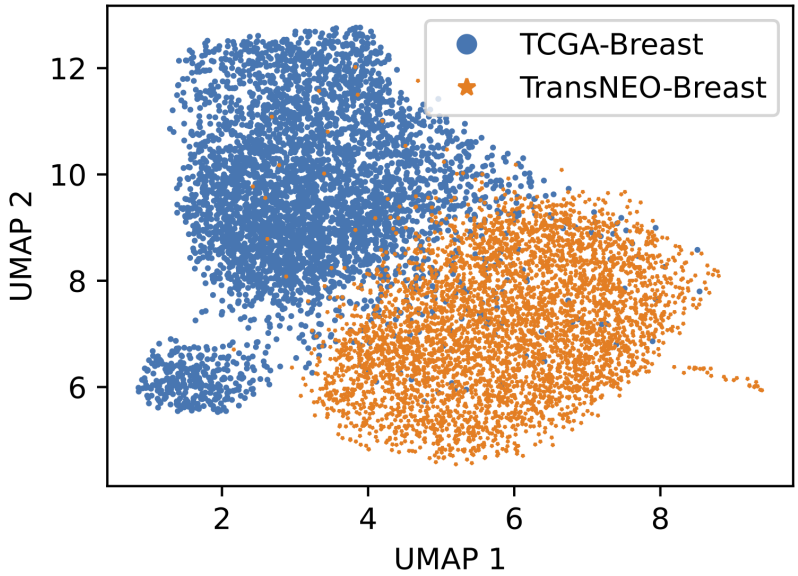

Figure S4. Difference between histopathological features extracted from TCGA-Breast tiles and TransNEO-Breast tiles. UMAP visualization of 2,048 histopathological features that were extracted by using pre-trained ResNet50 CNN. 4,000 image tiles from each dataset were selected randomly to illustrate. Each point represents each feature vector of one image tile.

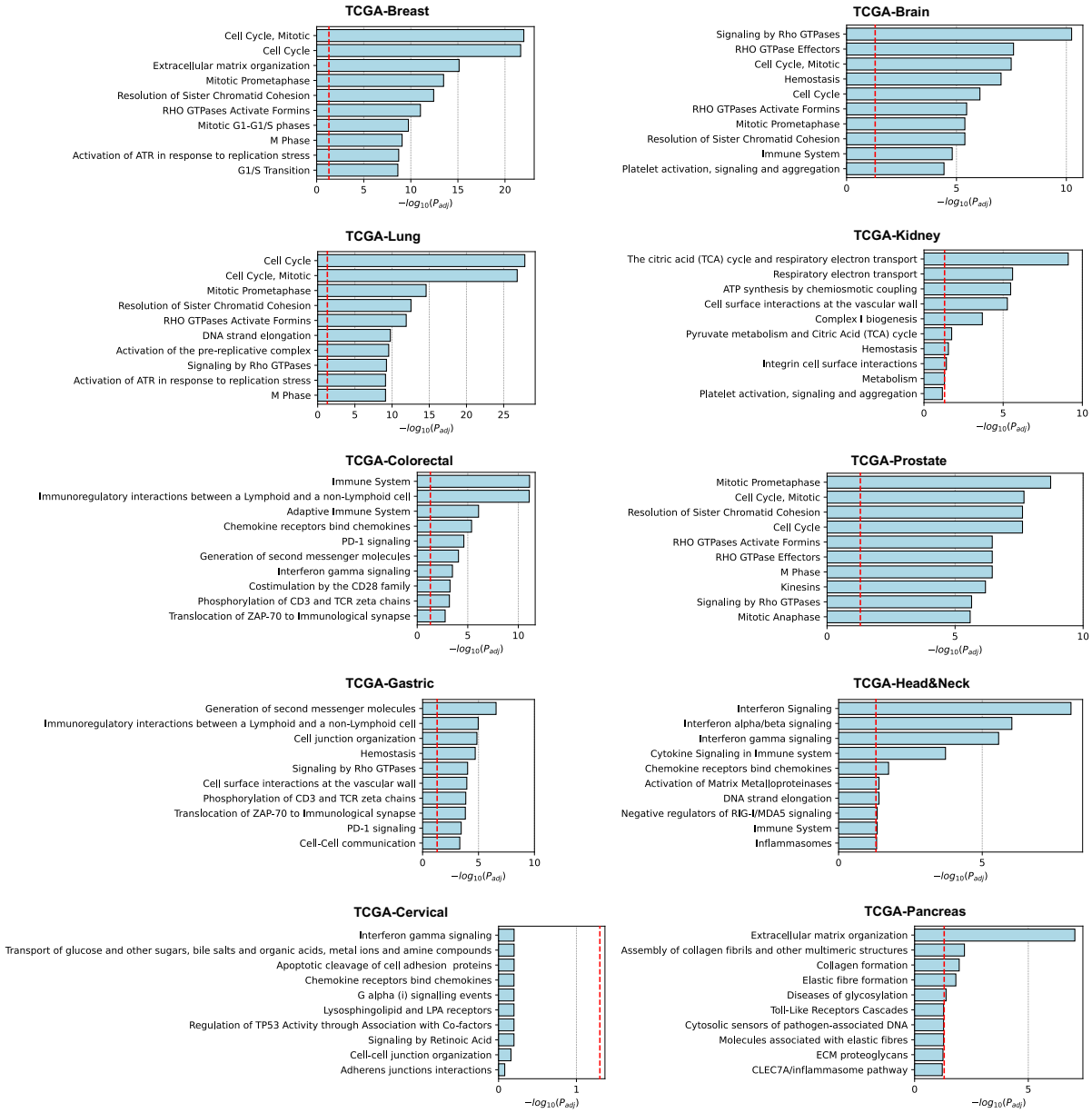

Figure S5. Top 10 enriched pathways among the well-predicted genes (with Pearson correlation above 0.4) for each cancer type cohort. Red dashed lines represent p-value = 0.05. GSEApy tool and Reactome database were used for the analysis.

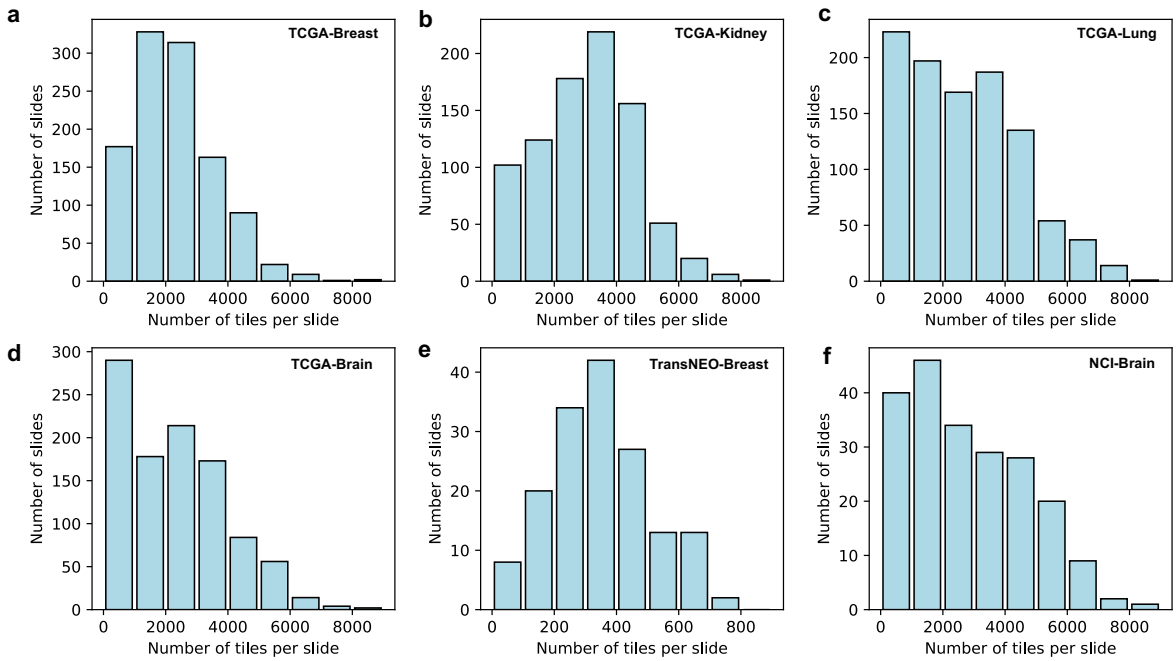

Figure S6. Histograms of the number of tiles per slide by cohort. The number of tiles in each slide image from TCGA and NCI-Brain datasets ranges from 100 to 8,000 (a, b, c, d), while the number of tiles in each TransNEO-Breast slide image is much smaller, ranging from 100 to 1,000 (e).

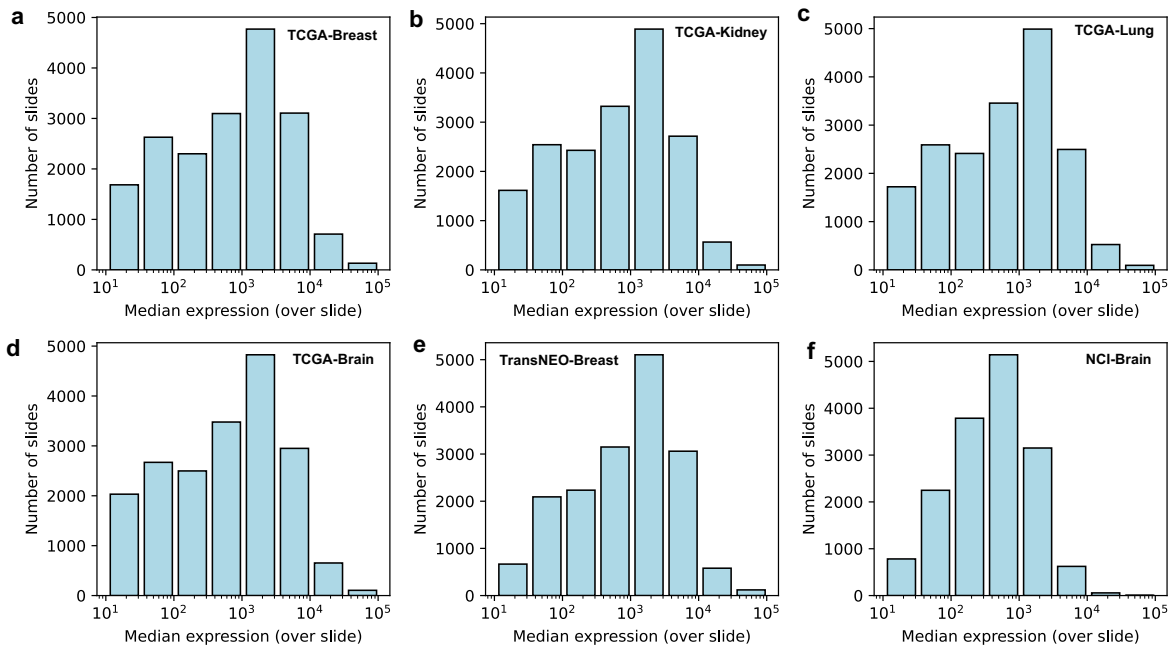

Figure S7. Histogram of median expression over slides. The median expression over samples of each gene commonly varies from 10 to 100,000 for every dataset considered in this study.

i