Large-scale systematic feasibility study on the pan-cancer predictability of multi-omic biomarkers from whole slide images with deep learning

Authors:

Salim Arslan[#*1], Debapriya Mehrotra[1,2], Julian Schmidt[1], Andre Geraldes[1], Shikha Singhal[1,3], Julius Hense[1], Xiusi Li[1], Cher Bass[1,4], Pandu Raharja-Liu[#1]

Affiliations:

[1] Panakeia Technologies, London, UK

[2] Department of Pathology, Barking, Havering and Redbridge University (BHR) NHS Trust, Romford, UK

[3] Department of Pathology, The Royal Wolverhampton NHS Trust, Wolverhampton, UK

[4] School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

# Authors contributed equally

* email: salim@panakeia.ai

Abstract:

We assessed the pan-cancer predictability of multi-omic biomarkers from haematoxylin and eosin (H&E)-stained whole slide image (WSI) using deep learning and standard evaluation measures throughout a systematic study. A total of 13,443 deep learning (DL) models predicting 4,481 multi-omic biomarkers across 32 cancer types were trained and validated. The investigated biomarkers included genetic mutations, transcriptomic (mRNA) and proteomic under- and over-expression status, metabolomic pathways, established markers relevant for prognosis, including gene expression signatures, molecular subtypes, clinical outcomes and response to treatment. Overall, we established the general feasibility of predicting multi-omic markers across solid cancer types, where 50% of the models could predict biomarkers with the area under the curve (AUC) of more than 0.633 (with 25% of the models having AUC larger than 0.711). Aggregating across the omic types, our deep learning models achieved the following performance: mean AUC of 0.634 ±0.117 in predicting driver SNV mutations; 0.637 ±0.108 for over-/under-expression of transcriptomic genes; 0.666 ±0.108 for over-/under-expression of proteomes; 0.564 ±0.081 for metabolomic pathways; 0.653 ±0.097 for gene signatures and molecular subtypes; 0.742 ±0.120 for standard of care biomarkers; and 0.671 ±0.120 for clinical outcomes and treatment responses. The biomarkers were shown to be detectable from routine histology images across all investigated cancer types, with aggregate mean AUC exceeding 0.62 in almost all cancers. In addition, we observed that predictability is reproducible within-marker and less dependent on sample size and positivity ratio, indicating a degree of true predictability inherent to the biomarker itself.

Introduction

With the advancements in computational pathology and ever-growing digitised datasets of tissue samples accompanied with genomic, epigenomic, transcriptomic, and proteomic data, there is an increasing interest in studying the alterations at different levels of the molecular landscape, with the aim of better understanding oncogenesis and cancer progression [1]–[3]. In-depth analysis of the associations between the molecular aberrations and the tumour microenvironment has enabled the development of targeted therapies that changed the course of treatment in various cancer types, such as breast cancer [4], lung cancer [5] and melanoma [6]. From a cancer prevention and screening perspective, genetic profiling is now considered an important tool [7], especially for those who possess a higher risk of developing cancer due to genetic factors. For instance, carriers of an altered BRCA1/2 gene can benefit from preventative treatments, which in turn, is likely to reduce their risk of breast cancer development [8].

Standard molecular and genomic profiling methods such as immunohistochemistry (IHC), next-generation sequencing (NGS), and in situ hybridization (ISH) can accurately characterize the molecular profile of a patient but may account for a large proportion of the laboratory delays in routine clinical workflow as they take time to prepare, process, and analyse [9], [10]. Moreover, ISH and NGS are expensive tests that may not be routinely accessible. In addition, test results and accuracy of outcome may depend on the efficacy of equipment and training of biomedical scientists, for which inter-expert variability constitutes another challenge. Reproducibility may also be a limiting factor, given that there still exists a lack of consensus on the interpretation of test results for some biomarkers, such as Human Epidermal Growth Factor Receptor 2 (HER2) using immunohistochemistry [11]. As biomarker discovery continues to advance, the number of additional tests is increasing, yielding even greater complexity and cost in histopathology workflows.

Meanwhile, there has been accumulating evidence suggesting that routinely available histology images stained with hematoxylin and eosin (H&E) may contain information that can be used to infer molecular profiles directly from tissue biopsies [12], [13]. Deep learning (DL), a newly emerging image analysis tool, can effectively reveal differences in morphological phenotypes in malignancies which are often too subtle to be determined visually. DL offers a means to gain more insight into the underlying biological events associated with molecular changes, which in turn, enables the prediction of molecular profiles directly from H&E-stained whole slide images [12]–[16]. DL-based methods have been used to infer molecular alterations in various cancer types, including breast [10], [17]–[19], colorectal [14], [16], [20], [21], lung [13], [22], prostate [23], liver [24], skin [25], and thyroid [26]. In addition, several studies have shown the potential of DL systems to predict survival outcome and therapy response [19], [27]–[31]. A review by Echle et al. provided an extensive review of the clinically-relevant biomarkers predictable from routinely available histology images [32] .

Recent pan-cancer studies have explored the links between genetic/molecular alterations and histomorphological features in H&E images and showed that in almost all malignancies deep learning (DL) methods can be used to infer a plethora of biomarkers directly from routine histology, including genetic mutations, transcriptomic profiles, tumour types, molecular signatures and subtypes and conventional pathology biomarkers [12], [15], [16], [33], [34]. Expanding upon previous studies, we conducted a large-scale, systematic study to test the predictability of biomarkers across the central

dogma of molecular biology from genomic, transcriptomic, proteomic, metabolomic to various clinically-relevant downstream biomarkers (e.g. standard of care biomarkers, molecular subtypes, clinical outcomes, response to treatment) from routine diagnostic slides using publicly available data from the The Cancer Genome Atlas (TCGA) project and standard evaluation measures throughout a structured pipeline (**Figure 1**). Our study aims to show the general feasibility of profiling biomarkers directly from H&E-stained whole-slide images (WSI) with deep learning across 32 cancer types, including breast cancer, lung adenocarcinoma, colorectal cancer, gastric carcinoma, prostate cancer, endometrial carcinoma and other major malignancies which have been studied comprehensively as part of the TCGA project. The specifics of the pre-processing and prediction methods are available in **Extended Methods**. In total, 13,443 models were trained across 4,481 distinct biomarkers, with the following breakdown: 1,950 mutation driver single nucleotide polymorphism (SNV) markers, 1,030 transcriptome expression level markers, 576 proteomic expression level markers, 450 metabolomic pathway markers, 270 gene signature and molecular subtype markers, 160 markers related to clinical outcomes and treatment responses and 45 standard-of-care markers.

Overall, we found that multi-omic biomarkers and clinically-relevant features can be predicted directly from histo-morphology. Detecting mutations from histology was mostly feasible for the majority of genes tested and frequently mutated genes like *TP53* were predictable across multiple cancer types. To some degree, it was possible to predict the under-/over-expression status of many genes at the transcriptome level. Similarly, detecting histomorphological changes that might be associated with molecular alterations affecting the expression of proteins was feasible to a certain extent. Overall, metabolomic biomarkers were somewhat less predictable compared to those of the other omic types, but certain pathways were still detectable from routine histology images. We observed that the differences in the tumour microenvironment can visually be determined with deep learning, enabling the prediction of molecular subtypes directly from H&E-stained whole slide images. From a clinical management perspective, our findings indicated that deep learning could detect the footprints of well-established clinical biomarkers in histopathological images. Finally, we showed that predicting certain markers with prognostic importance and response to treatment was possible to some degree. The overview of each biomarker type's feasibility can be seen in **Results** while individual discussions of relevant markers and cancer types are available in **Discussion**.

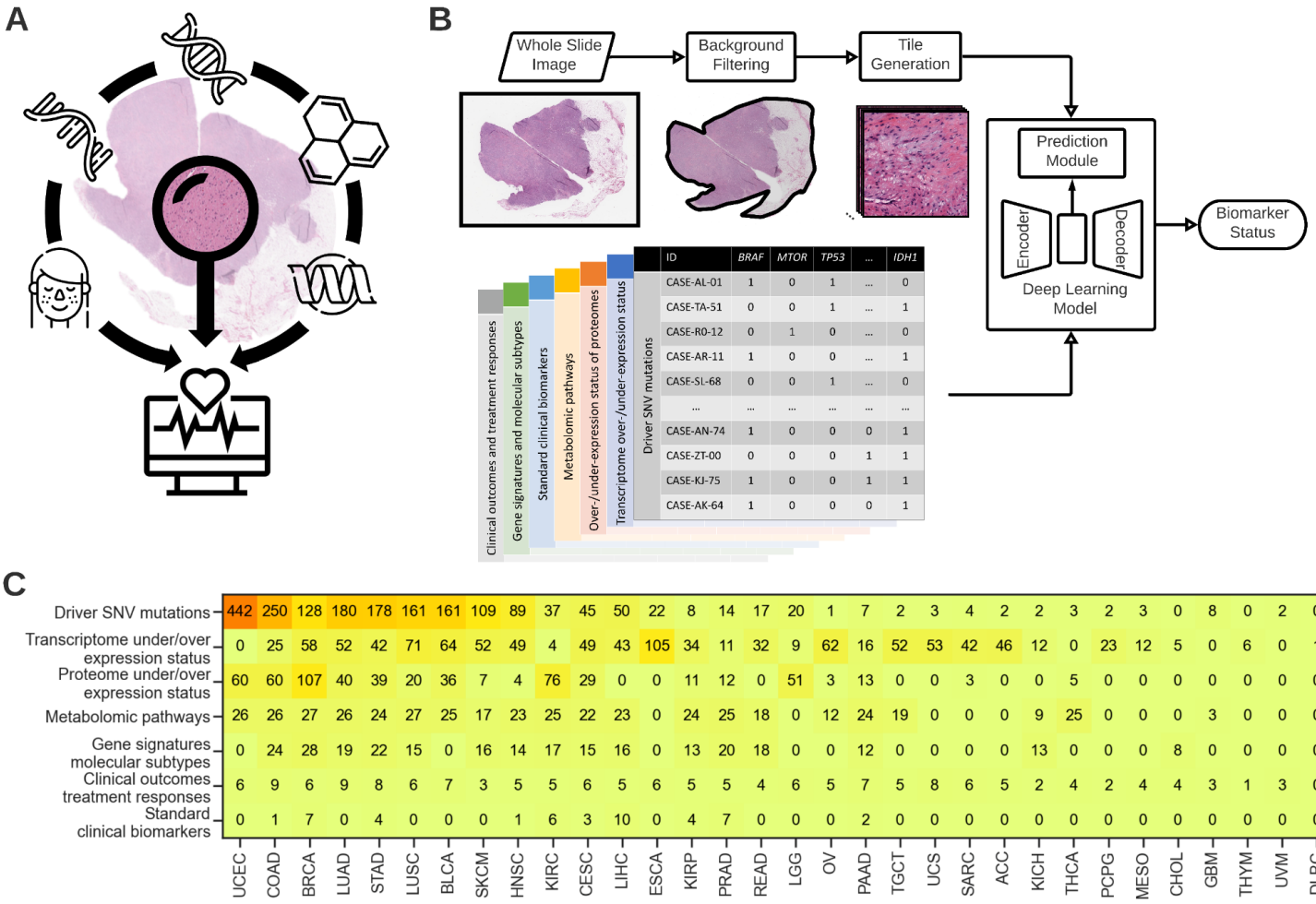


Figure 1: A: Graphical summary of the proposed study. We assess the general feasibility of predicting a plethora of biomarkers from genomic, transcriptomic, proteomic, metabolomic to various clinically-relevant biomarkers (e.g. standard of care features, molecular subtypes, clinical outcomes, response to treatment) using H&E-stained whole slide images and deep learning. **B**: Overview of the pre-processing and training pipeline. A proprietary convolutional neural network (CNN) was used for predicting molecular profiles from H&E images (see **Extended Methods: Pre-processing pipeline and training details**). A single CNN was end-to-end trained from scratch for each biomarker. Each slide was parcellated into a set of 256x256 tiles and those that did not contain any tissue were discarded. The remaining tiles were assigned with a ground-truth molecular profile (see **Extended Methods: Biomarker acquisition**). **C:** The number of biomarkers per cancer type shown in a heatmap. Each row corresponds to a single biomarker type, namely from top to bottom, driver SNV mutations, under/over-expression of transcriptomic genes, protein under/over-expression status, metabolomic pathways, gene signatures and molecular subtypes, clinical outcomes and treatment responses, and finally, standard clinical biomarkers.

Results

Multi-omic biomarkers and clinically-relevant features can be predicted directly from histo-morphology: We showed the general feasibility of profiling multi-omic biomarkers, including clinically-relevant prognostic markers using histo-morphological characteristics of standard hematoxylin and eosin (H&E) whole-slide imaging (WSI). Overall, our deep learning models performed significantly better than random based on a t-test with a p-value of 9.602e-1596 (practically 0). More than half of the models were validated at an Area Under the Curve (AUC) of

0.633 (**Figure 2A-left**). The observed AUC was greater than or equal to 0.711 for 25% of the models and above 0.829 for 5%. The top 1% models (n=135) returned an AUC of at least 0.904. Since we trained three models for each biomarker in a cross-validation configuration (see **Extended Methods: Experimental setup**), it was possible to assess the intra-marker variability of model performance (**Figure 2A-middle**, orange histogram). For the majority of the biomarkers, the predictability was highly significant with the standard deviation being less than 0.1 AUC, and the difference between the minimum and maximum performance being less than 0.2 AUC (**Figure 2A-left**, blue histogram). These indicate the general feasibility of predicting most of the biomarkers to a good extent from whole slide images with deep learning.

Most of the biomarkers showed better-than-random performance across all omics/biomarker types (**Figure 2B**, **Table 1**). The lowest average performance was seen in the prediction of metabolic pathways (AUC 0.564 ±0.081), and the highest performing models were from the standard clinical features (AUC 0.742 ±0.120). These two biomarker types also demonstrated the lowest and the most significant deviation, respectively. Performance reproducibility across different models of the same biomarker type showed similar trends compared to the overall distribution (**Figure 2D**), with the standard deviation for all omics being around 20% and a significant increase in variance considering the minimum-maximum performance (**Figure 2E**). We observed highly significant performance differences across almost all omics pairs, with pairwise t-tests returning p values of < 0.05 (**Figure 2C**).

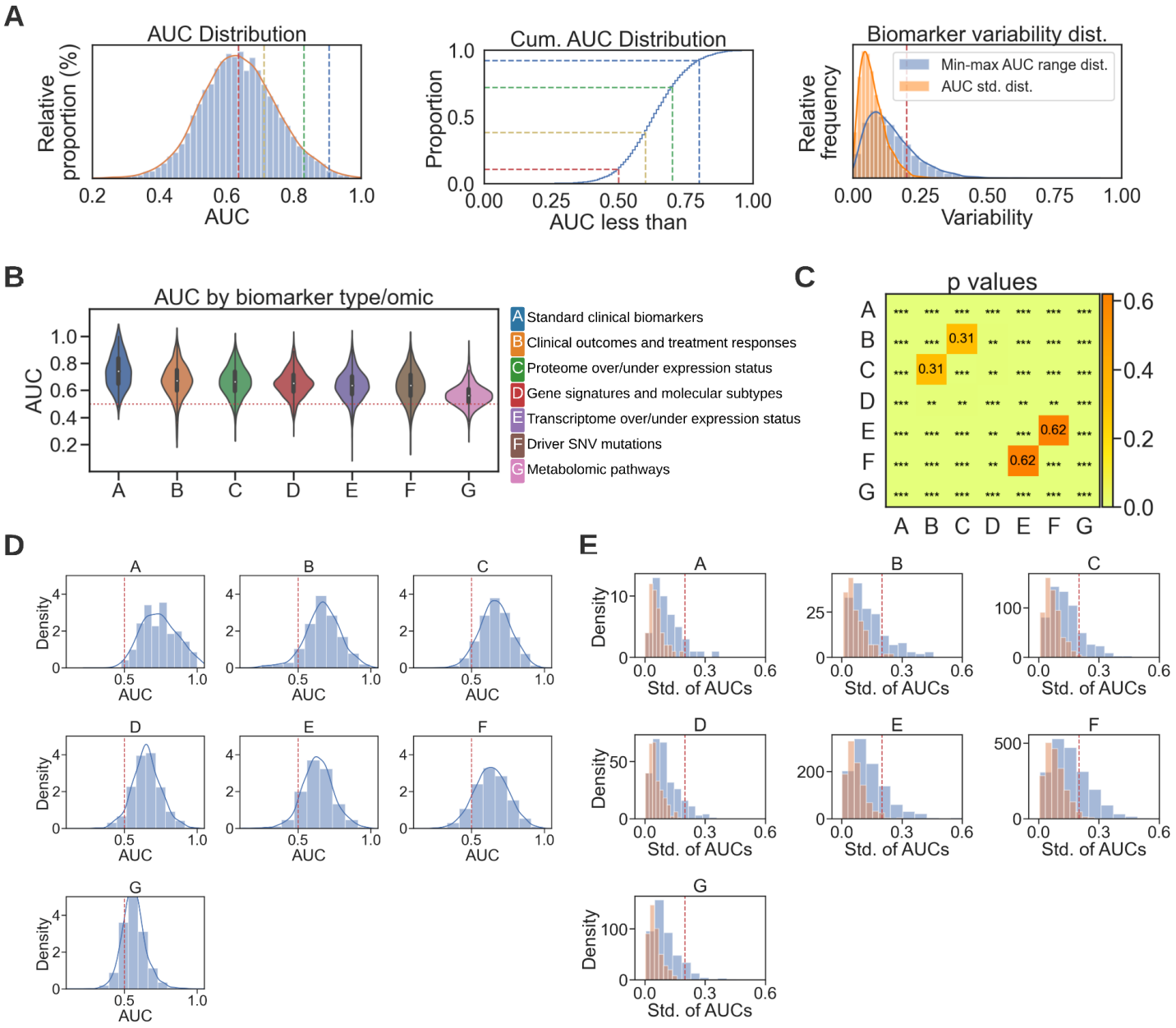


Figure 2: **A** - (*Left*) Histogram distribution and kernel density estimation of AUC values for all biomarkers. (*Middle*) Cumulative AUC distribution shows the proportion of models that have AUC less than the shown markers at 0.5, 0.6, 0.7, and 0.8. *(Right)*: Standard deviation and min-max range distribution of model performance in AUC. **B** - Violin plots showing the AUC distribution of each biomarker type. **C** - Statistical significance of performance differences across all omics pairs, obtained via pairwise t-tests (no star: non-significance, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 1e\text{-}05$). **D** - Histogram distribution and kernel density estimation of AUC values for all biomarkers in each omic type. **E** - Standard deviation (orange histogram) and min-max range distribution (blue histogram) of model performance in AUC.

Biomarker type / omic	Description	Mean AUC (± std.)
Driver SNV mutations	SNV mutations in driver genes associated with FDA-approved therapies or known-relevancy for specific treatments.	0.634 ±0.117
Over/under-expression of transcriptomic genes	Under- and/or over-expression status in driver genes at the transcriptome level.	0.637 ±0.108
Over/under-expression of proteomes	Under- and/or over-expression status of proteins encoded by driver genes.	0.666 ±0.108
Metabolomic pathways	Metabolic pathways under-/over-representation.	0.564 ±0.081
Gene signatures and molecular subtypes	Biomarkers that are highly relevant for prognosis and targeted therapies, including molecular subtypes and gene expression signatures, as compiled in a related study [12].	0.653 ±0.097
Standard clinical biomarkers	Established biomarkers that are routinely used in clinical management, as described in a related study [12].	0.742 ±0.120
Clinical outcomes and treatment responses	Clinical outcomes such as overall survival and treatment responses to a therapy or drugs.	0.671 ±0.120

Table 1: Average performance and standard deviation for all omic types, alongside their short description.

Pan-cancer predictability of multi-omic biomarkers from histology: To show the performance of biomarker prediction per cancer type, we plotted the distribution of AUC values for all studies in **Figure 3** and provide the average performance with standard deviations in **Extended Data Table 1**. Overall, all of the studies showed a significantly better performance than random (statistical significance with respect to pairwise t-tests is indicated on top of each violin plot in **Figure 3A**). The lowest general performance was obtained in uterine carcinosarcoma (UCS) with a mean AUC of 0.586 (±0.158), and the highest performing models were in mesothelioma (MESO), where an average AUC of 0.693 (±0.137) was measured. The standard deviation within each group ranged between 0.086 and 0.164, with pheochromocytoma and paraganglioma (PCPG) and glioblastoma multiforme (GBM) showing the lowest and largest intra-cancer variation, respectively. Variability across different folds of each biomarker was mostly stable, with standard deviations centring around 0.05 in most of the studies (**Figure 3A-Bottom**).

To assess how predictability changes for each malignancy depending on the type of biomarker, we analysed the performance and deviation of each cancer-omic pair (**Figure 3B**). In general, almost all studies showed a better-than-random average performance across all biomarker types, except for metabolomic pathways in glioblastoma multiforme (GBM). Performance differences were mostly significant when the AUC values in each subgroup were compared to random predictions with the same underlying distribution (based on pairwise t-tests returning p values of at least 0.05). Intra-group performance was highly stable with more than half of the sub-groups having a standard deviation of less than 10%. The largest variance was observed in the prediction of driver mutations in uterine carcinosarcoma (UCS) and clinical outcomes in kidney chromophobe (KICH), both with a standard deviation of 0.23.

Among the cancer studies targeting standard clinical features, kidney renal papillary cell carcinoma (KIRP, AUC 0.805 ±0.13, $p < 0.01$) and stomach cancer (STAD, AUC 0.805 ±0.084, $p < 1e\text{-}05$) had the top average performance, followed by kidney renal clear cell carcinoma (KIRC), breast adenocarcinoma (BRCA), and colon cancer (COAD), each with a mean AUC over 0.7. For genomic, transcriptomic, and proteomic biomarkers, the performances within individual cancer types were primarily consistent with their corresponding general trend. The highest performances were observed in thyroid carcinoma (THCA) and sarcoma (SARC) for the prediction of proteomic expression status with average AUCs around 0.78; in lower-grade glioma (LGG) and testicular germ cell tumours (TGCT) for the predictability of transcriptomic biomarkers with AUCs slightly above 0.7; and in kidney renal clear cell carcinoma (KIRC), pheochromocytoma/paraganglioma (PCPG), and thyroid carcinoma (THCA) for the detection of genetic alterations in driver genes with average AUCs ranging from 0.705 to 0.779. Multiple cancer types showed a relatively good performance with AUCs above 0.7 especially when considering the predictability of the clinical outcomes and treatment responses, where the performances of such predictions were among the highest across studies. Among them, the most notable studies were kidney renal papillary cell carcinoma (KIRP), adrenocortical carcinoma (ACC), glioblastoma multiforme (GBM), and kidney chromophobe (KICH) with average AUCs reaching as high as 0.777. Cancer-specific performance for the prediction of metabolomic pathways was on par with its overall low-performance, with none of them yielding an average AUC above 0.6.

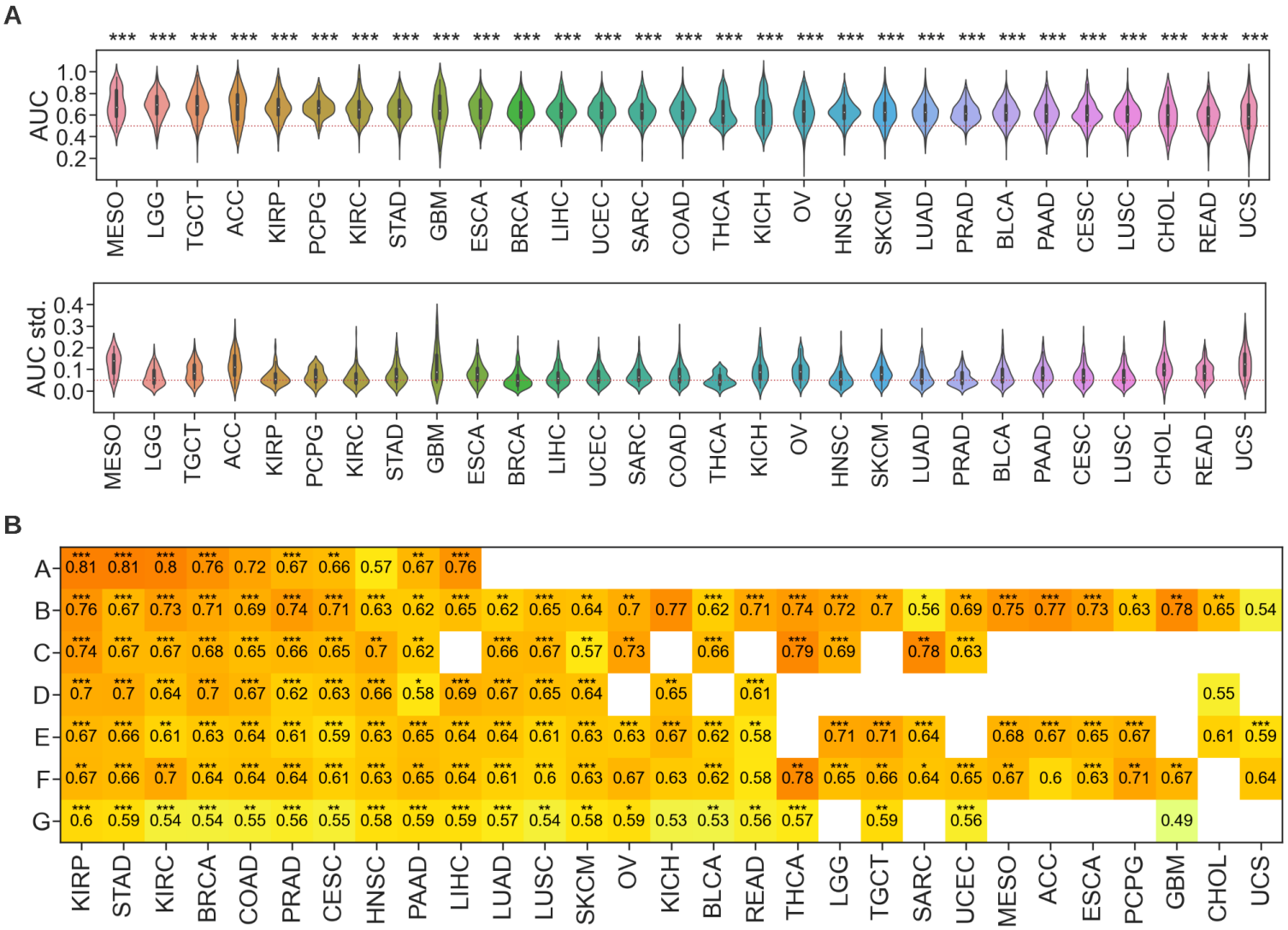


Figure 3: A - (*Top*) Violin plots showing the AUC distribution per cancer type. Plots are sorted by average intra-study AUC. DLBC, UVM, and THYM were excluded from this analysis due to only constituting one to seven valid targets across all biomarker types. (*Bottom*) Violin plots showing the standard deviation distribution of model performance across different folds of each biomarker (in the same order as in A). **B -** Average performance (AUC) across all cancer studies, confounded by biomarker type. More red colours indicate higher AUC. Asterixis atop each violin plot and within heatmap cells indicate the statistical significance of the performance difference between AUC values of a subgroup against randomly sampled values of the same underlying distribution (no asterix: not significant (n.s.), *: $p < 0.05$, **: $p < 0.01$, *** $p <$ 1e-05). The coding introduced in **Figure 2** is used to shorten the names of biomarker types.

Feasibility of predicting genetic alterations from histology: Various research studies, including the recent pan-cancer studies, have shown that mutations can be detected from histomorphological features with deep learning [12], [15]. In our study, we extend the previous work on detecting mutations to a total of 1,950 genes across all tested cancer types. We focused on driver genes that are associated with disease-specific therapies approved by the FDA or are known to be relevant for specific treatments based on evidence from clinical guidelines or well-powered studies with consensus from experts in the field [35]. In our experiments, we used the genomic profiles available

from the TCGA project (**Extended Methods: Biomarker acquisition**). A case was considered "mutated" if it contained at least one SNV mutation; otherwise, it was considered "wild-type".

Genetic alterations were significantly predictable across most of the investigated cancer types (**Figure 3**), with a mean AUC of 0.636 (±0.117). Scatter plots showing the performance of all models trained to predict genetic alterations in all cancer types are provided in **Extended Data Figure 1**. The performance of models across a selected set of malignancies can be seen in **Figure 4**. More than 40% of the mutations were detectable with an AUC of at least 0.65, and considering the highest performing mutations in each cancer type, almost all major malignancies had at least 10 mutations that were predictable with an AUC of 0.70 or above. Among them, endometrial carcinoma had the highest number of predictable mutations (112 genes). It was followed by colon cancer (62 genes), gastric adenocarcinoma (58 genes), skin melanoma (29 genes), lung adenocarcinoma (28 genes), and breast cancer (26 genes). The performance for the top 10 mutations ranged from 0.800-0.896 in endometrial carcinoma, 0.793-0.847 in gastric adenocarcinoma, 0.802-0.916 in renal clear cell carcinoma, 0.786-0.840 in colon adenocarcinoma, 0.784-0.858 in melanoma, 0.758-0.898 in lung adenocarcinoma, 0.761-852 in bladder urothelial carcinoma, and 0.754-0.891 in breast cancer. Among all the tested mutations, the top-performing ones were *NUMA1* and *JAK1* in kidney renal clear cell carcinoma, *PDGFRB* and *BCL6* in lung cancer, *IRS2* in endometrial carcinoma, and *GNAS* in breast cancer, each with an AUC of at least 0.89. A large number of genes were highly predictable across multiple cancer types. There were 117 genes with identifiable mutation status in at least two malignancies with a minimum AUC of 0.7. Notable ones that were predictable in multiple cancers are shown in Figure 10. Alterations in *TP53* were detectable in almost all cancer types, with 7 of them having an AUC of at least 0.7 and 14 of them showing AUCs greater than 0.65, reaching up to 0.841 for brain lower grade glioma, up to 0.785 for breast cancer and up to 0.771 for endometrial cancer. Other genes with a cross-cancer AUC of at least 0.7 were *BAP1* (mutations predictable in eight cancers), *PRDM16* and *JAK1* (mutations predictable seven cancers), and *CDH1* (mutation predictable five cancers). *CDK12*, *RB1*, *MTOR*, *NOTCH2, UBR5,* and *KMT2A,* are also worth mentioning with their mutations being detectable in at least ten different cancers with a mean AUC of 0.65 (**Figure 5**).

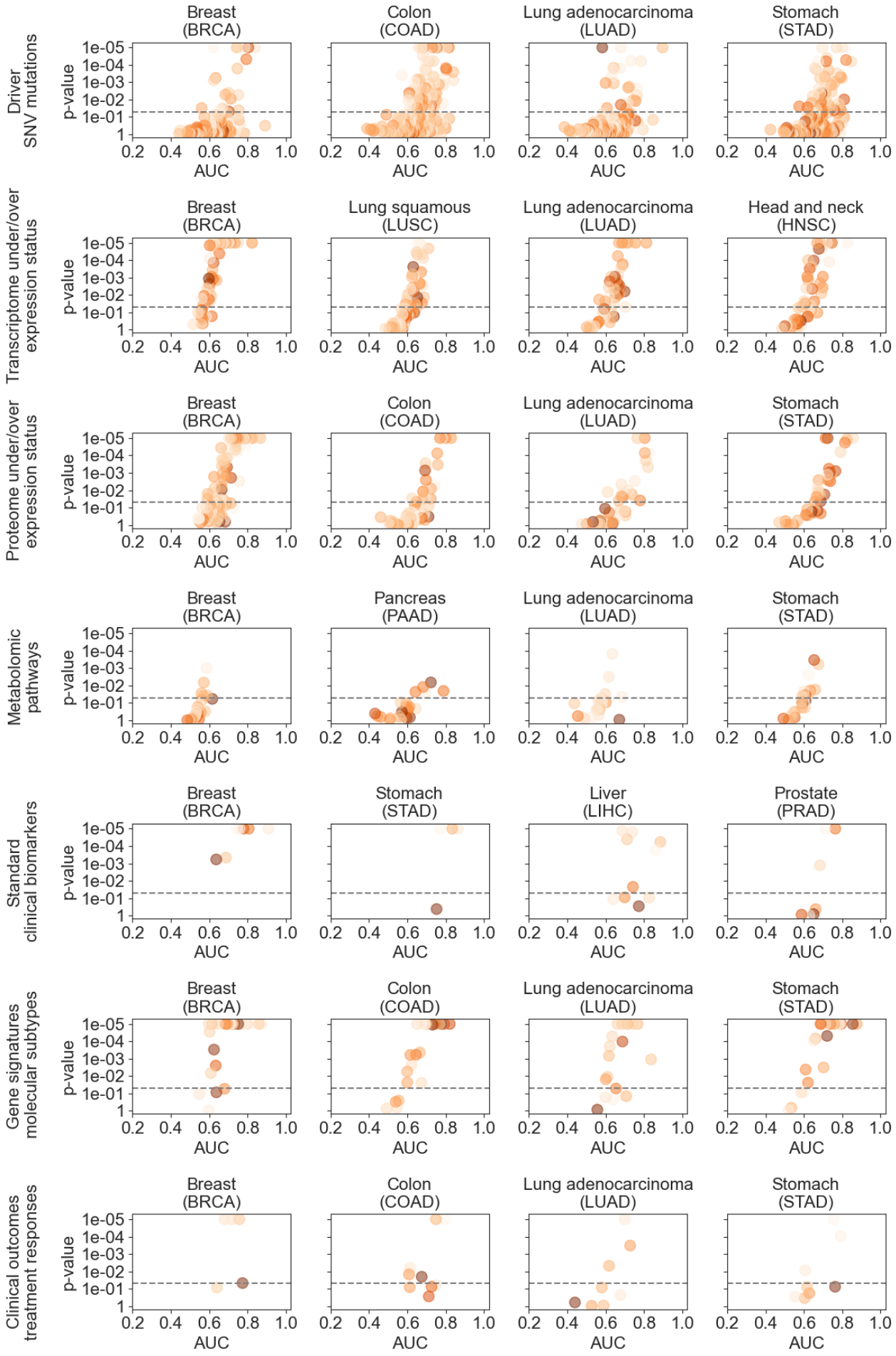


Figure 4: Scatter plots showing the performance of each model trained to predict biomarkers of different types/omics across selected cancer types. Each row corresponds to a single biomarker type, namely from top to bottom, *driver SNV mutations*, *under/over-expression of transcriptomic genes*, *protein under/over-expression status*, *metabolomic pathways*, *standard clinical biomarkers*, *gene signatures and subtypes*, and finally, *clinical outcomes and treatment responses*. Whilst only 4 malignancies per biomarker type were shown here, **Extended Data Figures 1-7** show scatter plots showing the individual model performance across all cancer types for each biomarker type/omic. A two-sided t-test was applied to the prediction scores of each model to assess the statistical significance, and the corresponding p-values were corrected for false discovery rate (FDR). The p values (y-axis) of each plot were inverted log-transformed for visualisation purposes. P values smaller than 1e-05 were set to 1e-05 to avoid numerical errors during transformation. The statistical significance threshold of 0.05 is marked with a dashed line. The x and y axes are scaled to the same range in all plots. Marker shading indicates the intra-biomarker standard deviation, where lighter colours correspond to a deviation close to 0.

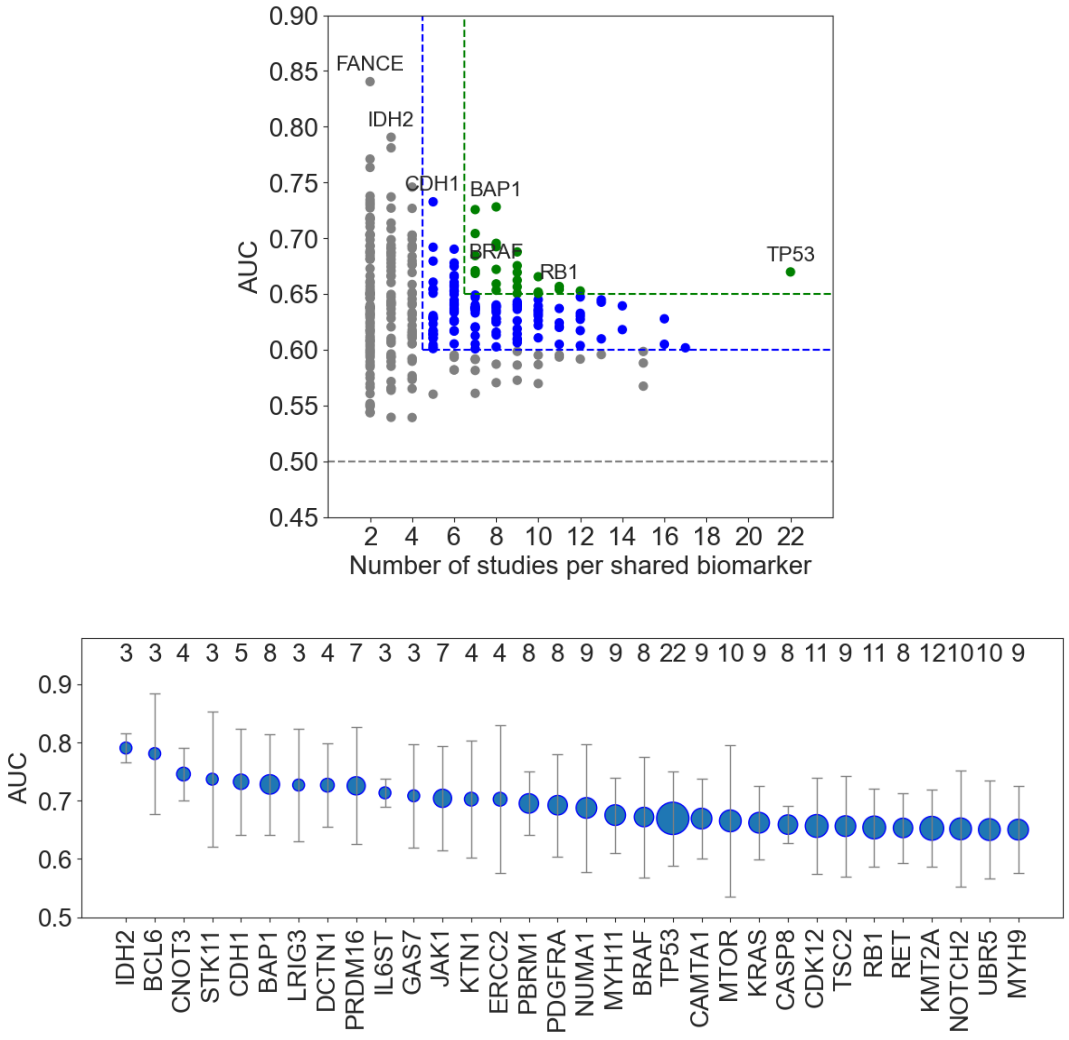


Figure 5: **Top**: Scatter plot showing the AUC values of predicted genetic alterations that appear in at least two cancer types. Areas outlined with blue and green lines mark the zones with high predictability and frequency of appearance. The green zone corresponds to the biomarkers with an AUC of at least 0.65 and a frequency of 6 and above. For blue, AUC and frequency are limited at 0.6 and 4, respectively. **Bottom**: List of most predictable targets that appear in at least 3 and 7 studies with an average inter-study AUC of 0.7 or 0.65, respectively. Whiskers show the variability of AUC across studies. The size of a marker represents the frequency of a biomarker (independent of its AUC). The exact number of appearances of a biomarker across studies is further provided in the secondary x-axis.

Feasibility of inferring over/under-expression of transcriptomes from diagnostic histology slides: Over or under-expression of a driver gene can cause altered phenotypes, which is likely to affect cancer development in multiple ways. Analysis of gene expressions is a key to better understanding the underlying cancer mechanisms and potentially help with improving cancer diagnosis and drug discovery [36]. While it is already known that genomic alterations could potentially be detected from histomorphology with deep learning, studies to understand the extent of predictability at transcriptome and proteome levels have been rather limited. Recently, Schmauch et al. showed that RNA-Seq profiles are correlated with histo-morphological features detected via deep learning in an annotation-free setup [16]. Our study took a more direct and comprehensive approach and trained deep learning models to predict the under- and/or over-expression status in driver genes, using the transcriptomic profiles available from the TCGA project (see **Extended Methods: Biomarker acquisition** for details).

A total of 1030 genes had enough samples to study the predictability of under and over-expression at the transcriptomic level. 97 and 933 of the genes had either under or over-expression status available, respectively. The gene expression status was predictable across most of the investigated cancer types (**Figure 3**) with a mean AUC of 0.637 (±0.108). The average performance was slightly lower for the under-expressed genes (mean AUC 0.633 ±0.115). Scatter plots showing the performance of all models trained to predict the over/under-expression of transcriptomic genes in all cancer types are provided in **Extended Data Figure 2** and the performance of models across a selected set of malignancies can be seen in **Figure 4**. Expression status in at least 40% of the genes was detectable with an AUC of 0.65 or above. Esophageal carcinoma and testis cancer were the malignancies with the highest number of genes with a predictable expression status (a total of 28 for both) considering an AUC level of 0.7. It was followed by ovarian cancer (18 genes) and adrenocortical carcinoma (16 genes). Several cancer types had at least ten genes with transcriptome expression status predictable at an AUC level of 0.7. The average AUC for the top-performing genes ranged from 0.812 to 0.911 in testicular germ cell tumours, 0.766-0.846 in esophageal carcinoma, 0.741-0.908 in adrenocortical carcinoma, 0.729-0.788 in ovarian cancer, 0.709-0.808 in melanoma, and 0.707-0.802 in kidney renal papillary cell carcinoma.

Almost all of the top-performing markers were gene over-expressions, with *PMS2* in thymoma; *CARD11*, *LASP1*, *STIL, POLE, KMT2C*, and *CLIP1* in testis cancer; *ERC1, WRN, OLIG2, FANCC,* and *ACSL6* in adrenocortical carcinoma; and *SOX2* and *NDRG1* in esophageal carcinoma leading in performance with AUCs ranging from 0.832 to 0.911. Considering the predictability of under-expressed genes, the most notable ones were *RHOA* in thymoma (AUC 0.908 ±0.05), *LSM14A*, *THRAP3*, and *MTOR* in the brain lower-grade glioma (AUCs ranging from 0.785 to 0.818) and *BAP1* in mesothelioma (AUC 0.818 ±0.084). The expression status of certain genes was highly predictable across multiple cancer types. There were 41 genes with the expression status being detectable in at least two cancers with an AUC of 0.7 or above. Among them, over-expression of *KMT2C* had a consistently high prediction rate with AUCs ranging from 0.733-0.837 in kidney renal papillary cell carcinoma, ovarian serous cystadenocarcinoma, and testis cancer. Other notable genes that were detectable across multiple cancer types at over-expression levels were *ERC1*, *CRTC3*, *LASP1*, *CDK4*, *SOX2*, *ERCC5*, *BCL6 and* the under-expression status of *RABEP1*, each being predicted in three or more different cancers with AUCs reaching up to 0.908.

Feasibility of predicting proteome expression level status with deep learning: As the next step in our multi-omics pan-cancer study, we assessed the ability of deep learning for detecting histomorphological changes that might be associated with the alterations in the expression of proteins. Towards this end, we used the proteomic profiles available from the TCGA project (**see Extended Methods: Biomarker acquisition**) and trained models to predict the protein over and/or under-expression status associated with driver genes. It is worth noting that association with a gene in this context refers to the encoding of a protein by that gene and we use "associated with/encoded by" interchangeably throughout the paper.

A total of 576 genes were qualified to evaluate the predictability of their corresponding proteomic expression status. In contrast to its transcriptomics counterpart, the distributions of the under and over-expressed proteomes were more balanced, with 309 and 267 of the genes being associated with either over or under expression of proteomes, respectively. We achieved an average AUC of 0.666 (±0.107), with the under-expression status being slightly less predictable on average (mean AUC 0.662, ±0.105) compared to its over-expressed counterpart (mean AUC 0.669 ±0.109). All the investigated cancer types had a statistically significant performance (**Figure 3**). Scatter plots showing the performance of all models trained to predict the over/under-expression status of proteomes encoded by driver genes in all cancer types are provided in **Extended Data Figure 3** and the performance of models across a selected set of malignancies can be seen in **Figure 4**.

The expression status prediction of almost all genes performed above random, with more than half of them being detectable with an AUC of at least 0.65 and over 30% of them further achieving an AUC above 0.7. Breast invasive carcinoma had the highest predictability rate, where the expression status of 37 genes was detectable with an AUC of at least 0.7. It was followed by kidney renal clear cell carcinoma and low-grade brain glioma (25 genes). Several cancers had at least ten genes with proteome expression status predictable at an AUC level of 0.7. The average AUC for these genes ranged from 0.787–0.866 in breast invasive cancers, from 0.769-0.952 in brain lower grade glioma, from 0.733-0.800 in kidney renal clear cell carcinoma, from 0.727-0.855 in stomach adenocarcinoma, and 0.721-0.827 in colon cancer.

The expression level status of a large number of proteins encoded by driver genes is highly predictable, with *TFRC*, *ATM*, and *PIK3CA* in low-grade brain glioma; *NRAS*, *FOXO3*, *MYC*, and *TP53* in renal papillary cell carcinoma; *CDKN1B* in head and neck squamous cell carcinoma; and *MYC* in sarcoma, exhibiting the top performance with AUCs ranging from 0.835 to 0.974. Multiple under-expressed proteins were also predictable to a great extent, the top-performing ones being *CASP8*, *MET*, *BCL2*, and *SETD2* in breast cancer; *AR* and *TFRC* in gastric adenocarcinoma; and *VHL* in lung cancer with AUCs ranging from 0.814 to 0.866. It was possible to predict the expression status of certain genes at a consistently high performance in multiple malignancies. There were 55 genes with the expression status being predictable in at least two cancers with a minimum AUC of 0.7. Among them, over-expression of *MYH11* was detected in 7 out of 11 cancer types with AUCs ranging from 0.705 to 0.809. Other notable genes associated with protein over-expression were *TFRC*, *TP53*, *MSH2*, *MSH6*, *ARID1A*, *RAF1*, and *NRG1*, showing a high-predictability in at least 5 malignancies, with AUCs reaching 0.942. Under-expression of proteomes encoded by *MYH1, NF2, VHL, and AR* were also predictable in at least 4 cancers, with AUCs reaching 0.856.

One gene that is worth noting is *TP53*, whose molecular alterations were already shown to be detectable in almost all malignancies. *TP53* is associated with the expression of p53 protein, which is known to cause poor prognosis and is correlated with advanced stage and high-grade cancer in kidney renal papillary [37]. We found that the *p53* protein over-expression was consistently predictable in six of the eight tested cancers, including renal cell carcinomas, lower-grade brain glioma, and endometrial cancer, with AUCs ranging from 0.672 to 0.835. *TP53* under-expression data was only available for breast cancer and lower grade brain glioma, and we acquired AUCs of 0.739 and 0771 for those malignancies, respectively.

Overall, the predictability of the protein expression status (mean AUC 0.666) appears to be significantly better (p-value < 1e-05) than that of transcriptome (mean AUC 0.637). However, both omic types had a comparable number of highly-predictable genes, i.e. 204 and 198 genes have an AUC of 0.7 or above, in the transcriptomic and proteomic cohorts, respectively. This might indicate that the overall lower performance of transcriptome expressions may be associated with this omic type having almost twice as many genes compared to its proteomic counterpart, and most of these genes have a relatively low predictability performance. This trend can also be observed in the distribution plots (Figure 2), where most of the values in both omic types are clustered around a similar median but the distribution of the transcriptomic cohort exhibits a longer tail that stretches well below the random performance of 0.5.

Feasibility of inferring metabolomic pathway characteristics from histology: We assessed the capability of our deep learning method for inferring metabolic pathway under-/over-representation directly from histopathology as our final target in the omic landscape (see **Extended Methods: Biomarker acquisition**)**.** Overall, metabolomic characteristics were less predictable than those of the other omic types, yielding an average AUC of 0.564 (± 0.081). Despite 87% of the pathways having a non-random performance, only 9% of them were likely to be predictable among a total of 450 tested targets (i.e. based on an AUC > 0.65). Considering the individual performance of each metabolomic pathway, we found that *nucleotide metabolism* in testicular germ cell tumours and pancreatic cancer; *nuclear transport* in colon adenocarcinoma, skin cutaneous melanoma, and lung cancer; *mitochondrial transport* and *histidine metabolism* in ovarian cancer were highly predictable, with AUCs ranging from 0.720 to 0.847. Similarly, it was possible to predict certain pathways involved in *beta-oxidation of fatty acids* directly from histomorphology to a certain degree (AUC > 0.65) for another four cancers, including thyroid carcinoma, liver carcinoma, ovarian cancer, as well as adenocarcinomas of pancreas and stomach. Scatter plots showing the performance of all models trained to predict metabolomic pathways in all cancer types are provided in **Extended Data Figure 4** and the performance of models across a selected set of malignancies can be seen in **Figure 4**.

Feasibility of predicting standard clinical biomarkers with deep learning: We tested the feasibility of deep learning to predict the established biomarkers that are routinely used in clinical management, as described in a related study by Kather et al. [12]. Towards this end, a set of standard of care features was compiled by following the biomarker acquisition approach in [12], including data for tumour grade, MSI status (in colorectal and gastric cancer), histological subtypes, hormone receptor status (in breast cancer), and Gleason sum (refer to [12] for an extensive list of all biomarkers and see **Extended Methods: Biomarker acquisition** for details).

Among all investigated omics and features, standard pathology biomarkers showed the highest predictability performance with an average AUC of 0.742 (±0.120). None of the biomarkers had a performance worse than random (i.e. all AUCs > 0.5) and almost 30% of them could be inferred with an AUC of above 0.8, a sign of very high predictability. Scatter plots showing the performance of standard clinical biomarkers in all cancers are provided in **Extended Data Figure 5** and the performance of models across a selected set of malignancies can be seen in **Figure 4**. Histological subtypes were in general highly predictable for breast cancer, renal cell carcinomas, liver cancer and gastric cancer. Clear cell and chromophobe subtypes of renal cell carcinoma had the most remarkable performance, reaching up to an AUC of 0.999. Two main histological types of breast cancer, i.e. invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC) were also highly detectable from whole slide images, with AUCs ranging from 0.759 to 0.908. Our models were also able to predict hormonal receptor status in breast cancer, with AUCs of 0.806 and 0.744, for ER and PR, respectively. Notably, multiple clinical biomarkers important for liver cancer could also be accurately inferred from histology, including growth patterns (AUC up to 0.862) and the status of non-alcoholic fatty liver disease (NAFLD, AUC 0.826 ±0.054). Another highly predictable biomarker was MSI status, which was detectable in both colon and gastric cancer with an AUC of up to 0.773.

Feasibility of inferring molecular subtypes and gene expression signatures from routine images: To evaluate the extent of our deep learning model to detect well-established molecular subtypes and gene expression signatures of cancer from whole slide images, we compiled a set of well-known features that have clinical and/or biological significance, by closely following the experimental details in [12]. This includes features that are highly relevant for prognosis and targeted therapies, including molecular subtypes and clusters, immune-related gene expressions, homologous recombination defects, cell proliferation, interferon-γ signalling and macrophage regulation and hypermethylation/mutation [38]–[40] (please refer to [12] for an extensive list of all features and (see **Extended Methods: Biomarker acquisition** for details).

Overall, molecular subtypes and gene signatures were more predictable than single-gene mutations and alterations at transcriptome and metabolome levels, with an average AUC of 0.653 (± 0.097). Almost half of the features were detectable at an AUC level greater than 0.65. Scatter plots showing the performance of all models trained to predict molecular subtypes and gene expression signatures in all cancer types are provided in **Extended Data Figure 6** and the performance of models across a selected set of malignancies can be seen in **Figure 4**. The most predictable features were observed in breast cancer (18) and the adenocarcinomas of the stomach (16) and colon (14). Our method was capable of inferring TCGA molecular subtypes in multiple cancer types, including kidney renal papillary cell carcinoma (AUC up to 0.884 ±0.085), gastric cancer (AUC up to 0.875 ±0.048), lung squamous cell carcinoma (AUC up to 0.861 ± 0.015), breast cancer (AUC up to 0.859 ± 0.028), kidney chromophobe (AUC up to 0.852 ± 0.053), lung adenocarcinoma (AUC up to 0.836 ± 0.043), and colon cancer (AUC up to 0.821 ± 0.084). Notably, the average predictability for PAM50 subtypes in breast cancer (i.e. *Basal, LuminalA, LuminalB, Her2, Normal*) was 0.752 (± 0.080), reaching up to an AUC of 0.871 (±0.015) for the *Basal* subtype. Consensus molecular subtypes in colon cancer (i.e. CMS1, CMS2, CMS3, CMS4) were also highly detectable with an inter-subtype average AUC of 0.763 (±0.068), reaching up to 0.821 (±0.083) for *CMS1*. Cell proliferation and hyper-methylation were also among well-predicted features, especially in cancers of breast, stomach, colon, and lung, with AUCs reaching up to 0.854. Immune subtypes were also consistently predictable across different

malignancies, including liver carcinoma, breast cancer, head and neck squamous cell cancer, gastric adenocarcinoma, colon cancer, and lung adenocarcinoma, with AUCs reaching up to 0.799.

Feasibility of inferring clinical outcomes and treatment response from diagnostic histology slides: Prognostic information refers to the information related to patient survival and outcome. It is highly vital for day-to-day patient management operations. Previous work has focused on developing prognostic models from routine clinical data, the standard of care features, histopathological assessment, molecular profiling, and more recently, morphological features acquired via deep learning [30], [31], [41]–[43]. In our study, we explored the end-to-end predictability of clinical outcomes directly from whole slide images across multiple cancer types by treating the clinical outcome endpoints such as overall survival (OS), disease-specific-survival (DSS), disease-free-interval (DFI), and progression-free interval (PFI) as potential prognostic biomarkers. We further expanded our analysis towards detecting treatment responses directly from H&E images to assess whether deep learning models can identify histo-morphological fingerprints that are correlated with the outcome of a therapy or drug. Each target of interest was systematically binarised into actionable biomarkers by taking multiple clinical and prognostic features into account [44] (see **Extended Methods: Biomarker acquisition** for details).

Overall, the predictive performance of clinical outcomes and treatment responses was relatively high compared to other biomarker types and omics, with a mean AUC of 0.671 (±0.120). Almost 40% of the tested targets were predictable at an AUC level of 0.7 or above. The majority of the investigated cancer types had a statistically significant performance (**Figure 3**). We acquired the best overall performance in glioblastoma multiforme, adrenocortical carcinoma, and kidney chromophobe with mean AUCs of 0.77. They were followed by renal papillary cell carcinoma, mesothelioma, thyroid carcinoma, prostate cancer, renal clear cell carcinoma, and esophageal carcinoma, with overall AUCs ranging from 0.731 to 0.76. *Residual tumour status* was the most consistently predicted biomarker, with AUCs reaching up to 0.821 in endometrial carcinoma and 0.795 in colon cancer**.** *DSS* was another highly predictable clinical outcome, which was detected in 12 different cancers with an AUC of at least 0.7. *DFI* and *OS* were the other notable biomarkers that were predictable across 7 to 10 diverse cancers, respectively.

Scatter plots showing the test of the performance of all models trained to predict clinical outcomes and treatment responses in all cancer studies are provided in **Extended Data Figure 7** and the performance of models across a selected set of malignancies can be seen in **Figure 4**. *OS* in kidney chromophobe, glioblastoma multiforme, thyroid carcinoma, and adrenocortical carcinoma; *DFI* in esophageal carcinoma and renal clear cell carcinoma; *DSS* in glioblastoma multiforme; *residual tumour* status in endometrial carcinoma, and *treatment response* in renal papillary cell carcinoma were among the top-performing targets with AUCs ranging from 0.815 to 0.924. Despite drug responses being somewhat less predictable compared to clinical outcomes, *cisplatin* in cervical, testis and gastric cancers; *temozolomide* in lower-grade glioma; and *paclitaxel* in breast cancer had relatively high performance of AUCs greater than 0.75.

Discussion

This paper assessed the general feasibility of predicting a plethora of biomarkers from a pan-cancer multi-omic perspective, using deep learning and histo-morphological features extracted from H&E stained routine diagnostic slides. The histo-morphological characteristics were shown to be predictable in inferring the biomarkers across the omics spectrum from genomics (genetic alteration), transcriptomics (mRNA over- and under-expression), proteomics (protein over- and under-expression), metabolomics (pathway over- and underrepresentation), as well as clinically-relevant outcomes including prognostic, standard of care and phenotype markers, and response to treatment (**Results: Multi-omic biomarkers and clinically-relevant features can be predicted directly from histo-morphology**). In addition, we observed varying predictability depending on the biomarker type.

Specific to genomic and standard clinical/prognostic markers (e.g. *TP53* and molecular subtyping), we observed high predictability across multiple cancer types (**Discussion: Feasibility of predicting genetic alterations from histology** and **Results: Feasibility of inferring molecular subtypes and gene expression signatures from routine images**). The histo-morphological characteristics were shown to be predictable in inferring the multi-omic biomarkers across major cancer types (**Results: Pan-cancer predictability of multi-omic biomarkers from histology**). We observed varying predictability depending on the cancer type. The prediction of a biomarker appeared to be reproducible, with the predictor performing comparable (as indicated by AUC) (**Results: Multi-omic biomarkers and clinically-relevant features can be predicted directly from histo-morphology** and **Figure 2A**).

The performance of a biomarker did not seem to depend on the putative confounders, such as the size of the dataset and the ratio of positive-to-negative samples (**Extended Data Figure 8**). On the other hand, they still seemed to be rather important for predictability, but not as high as the combined impact of the biomarker types (**Extended Data Figure 9**). Based on the results of the same experiment, the predictability was also somewhat influenced by the cancer types. Comparing various factors: biomarker type, cancer type and the biomarker in question seem to be the strongest factors contributing to the predictability of the biomarker.

Clinically-relevant genomic markers that are significantly predictable and widespread: Identifying patients with certain mutations are highly important for clinical management and developing targeted therapies [12]. In our studies, we found that detecting alterations from histology was mostly feasible for the majority of genes tested. Among others, *TP53* mutations are known to be one of the most commonly identified alterations in multiple cancer types ranging from 10% in haematological malignancies to almost close to 100% in high grade serous ovarian carcinomas [45]. Its germline variant can cause Li-Fraumeni syndrome, which is associated with a wide range of cancers, including soft tissue sarcoma, breast cancer and melanoma [35]. Tumours with TP53 mutations are likely to be poorly differentiated and can be linked to higher grade cell changes [15]. Alterations in *EP300*, while can be seen in diverse cancers including colon, lung, breast, stomach, are associated with increased tumour burden and antitumour immune activity and could potentially be a predictive biomarker for immunotherapy response in certain cancers [46]. *BAP1* is a known tumour suppressor and is associated with improved prognosis in certain cancers, such as mesothelioma [47]. Recent studies show that detection of *BAP1* mutation can potentially be useful for the development of targeted

treatment strategies in clear-cell renal cell carcinoma [48]. *MTOR* is involved in the mechanistic target of rapamycin (mTOR), a protein kinase that promotes tumour growth and metastasis when activated [49]. *MTOR* mutations can serve as biomarkers for predicting tumour responses to mTOR inhibitors, which are already being used to treat human cancer [50]. One of the highly predicted genes, *GNAS,* is known to promote cell proliferation and migration in breast cancer when expressed at high levels, and thus, can potentially be used as a therapeutic target [51].

Clinically-relevant metabolomic pathways that are significantly predictable: Upstream alterations affecting the expression of metabolic enzymes or their regulatory proteins can lead to cell metabolism changes that promote oncogenic activities including malignant cell growth and proliferation [52]. There is growing evidence indicating that the genetic aberrations in signalling pathways lead to metabolic dysregulation that promotes oncogenesis and disease progression [53] (also see [52] for an extensive list of transcriptomic and metabolomic studies conducted in various cancer tumours for this endeavour). Identification of metabolic changes in tumours and how they differentiate from normal cells can help uncover the complex mechanisms that control the anticancer drug response and potentially yield better prediction of therapeutic outcomes [52], [54]–[56]. Despite many studies targeting tumour metabolism, the detectability of metabolic activities from WSIs has been highly limited. One recent study used a deep-learning model to predict the activities of ten canonical biological pathways in breast cancer and found that the p53, PI3K, and cell cycle pathways are predictable to a certain degree (AUC > 0.65). In our study, metabolomic biomarkers were overall less predictable than the other omics of the molecular landscape. We found that the *nuclear transport pathway* can be detected in the adenocarcinomas of the lung and colon and in skin melanoma. This pathway is a critical part of normal cell activity, and abnormalities in this mechanism have been identified in a variety of tumours [57]. Another highly predictable pathway was *nucleotide metabolism*, which is known to support the uncontrolled growth of tumours when increased [58]. *Fatty acid beta-oxidation* is a well-known metabolic pathway that is dysregulated in tumours [59]. We found that in several different cancer types, it was possible to predict this pathway directly from histology.

Predictability of standard of care markers and comparison to the relevant studies: Our findings indicated that deep learning can potentially detect the footprints of well-established standard of care clinical biomarkers in histopathological images. This marks another important step in pursuit of achieving end-to-end detection systems from the histology in the current standard practice, which can potentially accelerate clinicians in patient management and accelerate diagnosis. Our results mostly agree with those obtained in the recent pan-cancer study by Kather and colleagues [12] and we observed a performance improvement in 71% of the biomarkers investigated in both studies (considering average AUCs for each feature-cancer pair).

Predictability of molecular subtypes and gene expression signature; and comparison to the relevant studies: We found that the differences in the tumour microenvironment and regional architecture can visually be determined with deep learning, enabling the prediction of standard molecular subtypes directly from H&E-stained whole slide images. Considering the importance of these features in cancer treatment, accurate predictability from histology can help develop more patient-centric treatments. Similar findings have been reported in the recent pan-cancer study performed by Kather and colleagues [12], where they observed that high-level molecular features

are more detectable than genetic alterations, with breast cancer, lung adenocarcinomas, gastric cancer, and colon cancer exhibiting the highest predictability. The molecular changes associated with these features can potentially have a much greater impact on the cellular morphology as compared to the alterations at the gene level, which in turn can be associated with the improved detection using deep learning [12]. Given the similarities between the two studies, our results can confirm their findings and provide more comprehensive evidence for the feasibility of detecting molecular subtypes from histology. In addition, our models overall showed much greater predictability for 72% of the biological processes investigated in both studies, notable ones being TCGA molecular subtypes, immune subtypes, macrophage regulation, dendritic cell activation, PAM50 subtypes of breast cancer, consensus colorectal subtypes of colon cancer, and proliferation.

Clinical outcomes and treatment responses that are significantly predictable: Classification of residual tumours is a critical stage for the course of treatment and is considered an important prognostic biomarker [60]. In our analysis, we cast this classification task as the binary prediction of the existence of residual tumour (positive class). We found that deep learning could detect the (no-)occurrence of residual tumours in multiple cancers. This predictability might indicate that some visual clues correlated with complete remission after treatment are already present in histomorphology at the time of diagnosis. Other clinical outcomes, such as overall survival (OS) and disease-specific survival (DSS), were also predictable to a certain extent. However, it is worth noting that the accurate definition of clinical outcomes is based on multiple events known at different times during treatment. As a result, their usability should be carefully assessed, especially for the cancer types that need longer follow-up times, have a small cohort size or have a limited number of events [44]. We also found that responses to some drugs such as *cisplatin*, *temozolomide* and *paclitaxel* were highly detectable from histology. Overall, our findings show the potential of deep learning to predict specific prognostic outcomes and treatment responses at diagnosis, which opens up many opportunities for investigating the correlation between disease outcome and histomorphological features.

Conclusion

The histo-morphological characteristics were shown to be informative for inferring the biomarkers across the omics spectrum from genomics (genetic alteration), transcriptomics (mRNA over- and under-expression), proteomics (protein over- and under-expression), metabolomics (pathway over- and underrepresentation), as well as well-established clinically-relevant biomarkers including prognostic, standard of care and phenotype markers, and response to treatment. The multi-omic biomarkers were predictable across all cancer types with varying degrees of performance. The predictability was also repeatable within-marker. Confounding factors such as positivity ratio and sample size surprisingly did not influence the predictability a lot, indicating that there may exist inherent detectability associated with histomorphology. Comparing various factors, biomarker type, cancer type, and the biomarker itself in question seem to be the most substantial factors contributing to the predictability.

Going forward, the questions around the specific mechanism of biomarker detectability, such as the predictive pattern and how it is conserved across the population(s), merit further research. While

this study has elucidated early observations on the factors determining a biomarker's predictability, further understanding would be necessary before the mainstream adoption of such a technique in day-to-day clinical settings.

Having explained the mechanism as well as the risk and the opportunity associated with multi-omic biomarker profiling from standard tissue imaging, the potential of applying such approaches in clinical management is enormous.

Extended Methods

Dataset: We conducted our experiments on the data provided from The Cancer Genome Atlas (TCGA) project, which was retrieved via the Genomic Data Commons (GDC) Portal (https://portal.gdc.cancer.gov/). The TCGA dataset consisted of 10,954 hematoxylin and eosin (H&E)-stained, formalin-fixed, and paraffin-embedded (FFPE) whole slides images of 8,890 patients, acquired from the following studies: breast invasive carcinoma (BRCA), cervical squamous cell carcinoma (CESC), kidney renal papillary cell carcinoma (KIRP), kidney renal clear cell carcinoma (KIRC), kidney chromophobe (KICH), skin cutaneous melanoma (SKCM), sarcoma (SARC), pancreatic adenocarcinoma (PAAD), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), bladder urothelial carcinoma (BLCA), esophageal carcinoma (ESCA), thyroid carcinoma (THCA), lymphoid neoplasm diffuse large B-cell Lymphoma (DLBC), brain lower grade glioma (LGG), thymoma (THYM), head and neck squamous cell carcinoma (HNSC), uterine corpus endometrial carcinoma (UCEC), glioblastoma multiforme (GBM), cholangiocarcinoma (CHOL), liver hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), lung adenocarcinoma (LUAD), and lung squamous cell carcinoma (LUSC), colon adenocarcinoma (COAD), rectum adenocarcinoma (READ), adrenocortical carcinoma (ACC), mesothelioma (MESO), pheochromocytoma and paraganglioma (PCPG), testicular germ cell tumours (TGCT), uterine carcinosarcoma (UCS) and uveal melanoma (UVM). Only images scanned at 0.5 microns per pixel (MPP) (corresponding to a magnification of 20X) were kept and images with no MPP information were automatically discarded. The number of the images and patients included in the TCGA cohort are provided in **Extended Data Table 2**. DLBC, UVM, and THYM were excluded from certain results due to having less than 7 valid targets considering all biomarker types (see Extended Methods: Molecular biomarker acquisition for more details on the biomarker inclusion criteria for each omic/biomarker type).

Biomarker acquisition:

Acquisition of actionable driver genes: Clinically-relevant driver genes were retrieved from https://cancervariants.org [61]. We only considered driver genes that are known to be associated with 1) FDA-approved disease-specific therapies and 2) response or resistance to therapies as shown in professional guidelines and/or based on well-powered studies with consensus from experts in the field based on evidence provided in [35]. Driver mutation and drug-associated data were downloaded from https://drive.google.com/drive/folders/1ZY6o3uaLOZSjOQPpXSMsnMbXmFWpb58d with the following sources being taken into account: BRCA exchange, the Cancer Genome Interpreter Cancer Biomarkers Database (CGI), Clinical Interpretation of Variants in Cancer (CIViC), Jackson Laboratory

Clinical Knowledgebase (JAX-CKB), the Precision Medicine Knowledgebase (PMKB). The source files were parsed into an intermedia format and the associations between SNP mutations and phenotypes were created. An expert pathologist mapped these phenotypes to TCGA studies. Finally using this mapping and driver mutation data per phenotype, we created a set of driver genes per TCGA study and subsequently used them to filter actionable biomarkers for transcriptomic, proteomic and genomic data.

Genomic biomarker profiles: Genomic biomarker data was collected using the cBioPortal web API (https://docs.cbioportal.org/6.-web-api-and-clients/api-and-api-clients) and the GDC API (https://gdc.cancer.gov/developers/gdc-application-programming-interface-api). For each TCGA study, we retrieved all samples with associated diagnostic slides. The samples that did not have whole-genome or whole-exome sequencing data were excluded from further consideration, allowing us to assume that all genes of interest were profiled across the remaining ones. While there existed samples without WGS or WXS data with mutations, it was not possible to assume that genes with no mutations were present in their wildtype, as they might simply have not been sequenced. As a next step, for all TCGA studies listed on the cBio portal, we acquired molecular profiles with the *MUTATION_EXTENDED* alteration type and all mutations belonging to these molecular profiles within the collected samples were retrieved and stored in an intermediate format. Finally, we created molecular profiles for all driver genes using this mutation data. A sample was considered positive for a driver gene if it contained at least one SNV mutation for that gene. The resulting profiles were filtered to exclude driver genes that had less than 10 positive samples in a given cancer.

Transcriptomic and proteomic profiles: Transcriptomic and proteomic data for TCGA datasets were retrieved from the cBioPortal API (http://www.cbioportal.org/api). FPKM files containing raw counts of gene expression and the corresponding z-scores were acquired for each coding gene and each sample with an associated tissue slide in the TCGA studies. The transcriptomic z-scores were obtained among samples in which the tumour comprised diploid cells, whereas the proteomic z-scores were calculated among all available samples. The z-scores were binarised for each gene and sample based on thresholds chosen as follows: For each sample, genes with a z-score of less than or equal to *t_under* were considered *underexpressed* and those with a z-score of larger than or equal to *t_over* were considered *overexpressed*. We empirically set {t_under, t_over} to {-2, 2} and {-1.5, 1.5}, for transcriptomic and proteomic data, respectively. These thresholds were then used to generate two under/over-expression profiles for proteomic and transcriptomic genes. In an "under-expression" profile all samples that were considered underexpressed were labelled as positive whereas all other samples were labelled as negative. Similarly, overexpressed samples in an "over-expression" profile were assigned a positive label, while the remaining samples were considered negative. Finally, to reduce the number of target biomarkers, we limited the over-and under-expression profiles to only include the driver genes (see **Acquisition of actionable driver genes**) for each study. Furthermore, profiles that did not contain enough positive samples were excluded. The minimum number of positive samples for proteomic genes was set to 20. For the transcriptomic profiles, only the ones with at least a positive ratio of 10% and having a minimum of 10 positive samples were kept.

Metabolomic pathways: Metabolic data was downloaded from https://www.ebi.ac.uk/biomodels/pdgsmm/index. It contained personalized genome-scale metabolic

models (GMMMs) stored in XML format for 21 of the TCGA studies. We used the pathology atlas of the human cancer transcriptome data [62] to acquire the generalised base network used to create the GSMMs. For each pathway in these networks, we derived a biomarker by comparing the genes present in the pathway for a sample, to all genes present in the generalised underlying network for the given pathway. We created sets of genes present for each pair of pathways/samples and performed t-tests comparing them to the set of genes for the same pathways in the generalised network. The resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure [63] with an FDR rate of 0.05. All samples with an adjusted p-value of less than 0.05 were considered significantly associated with the pathway. We limited ourselves to a set of "pathways of interest" which contained 32 pathways selected based on the entropy of their positive/negative distributions over all studies, as well as a few hand-picked pathways with known connections to cancer phenotypes. We created biomarker profiles for the 21 TCGA projects, by considering each pathway as a biomarker, where a sample was labelled as positive if it was significantly associated with a pathway. Pathways constituting less than 20 positive samples in each study were excluded from the resulting profiles.

Standard of care features, gene expression signatures and molecular subtypes: A publicly available dataset provided as part of a recent pan-cancer study on detecting clinically actionable molecular alterations was used to acquire the biomarker profiles for gene expression signatures, molecular subtypes and standard clinical biomarkers [12]. The dataset was originally curated from the results of systematic studies using the TCGA data (https://portal.gdc.cancer.gov/) [38]–[40] and contained profiles for 17 TCGA datasets (please refer to [12] for the description of biomarkers and other details regarding the acquisition protocol). For certain biomarkers, we used the consensus opinion to map the molecular status to binary labels. For instance, considering microsatellite instability (MSI), all patients defined as MSI-H were included in the positive class, while microsatellite stable (MSS) and MSI-L patients were labelled as negative. Profiles with multiple categorical values were binarised with one-hot-encoding, where a profile was created for each category with only the samples of that category being set to positive. Non-categorical profiles with continuous values were binarised at mean after eliminating NaN values.

Clinical outcomes and treatment responses: Survival data was acquired from the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR), a publicly available dataset that provides four major clinical outcome endpoints [44], namely overall survival (OS), disease-specific-survival (DSS), disease-free-interval (DFI) and progression-free interval (PFI). These endpoints were systematically binarized into actionable events by considering multiple clinical and prognostic features acquired from TCGA's routinely-collected clinical data (https://portal.gdc.cancer.gov/) such as vital status, tumour status, cause of death, new tumour events, local recurrence and distract metastasis. The details of the integration of the clinical data into actionable survival outcomes are given in [44]. Additionally, we added the residual tumour status acquired from the TCGA clinical files as another prognostic target. Patients with microscopic or macroscopic residual tumours (R1 or R2) were classified as positive whereas those with no residual tumour (R0) were included in the negative class [60]. Since TCGA-BRCA did not have residual tumour information, we used margin status as per the guidance in [55]. Similarly, `treatment_outcome_first_course` was used to create binary targets representing the treatment response. Towards this end, any patient with "Complete Remission/Response" was included in the positive class whereas "Stable Disease", "Partial Response"

and “Progressive Disease” were considered negative. Finally, clinical drug files in the TCGA datasets were used to identify drug responses. This was achieved by first unifying drug names based on the data provided in [64] and then identifying drug-study pairs with enough samples. Finally, the `treatment_best_response` attribute was used to map the drug responses into binary categories, with “Complete Response” constituting the positive class and the others being negative.

Experimental setup: We assessed the predictive performance of each biomarker in a 3-fold cross-validation setting, where the cases with a valid biomarker status in each dataset were split into three random partitions (folds), each having approximately the same proportion of positive samples. We trained and tested three models per biomarker, each time keeping aside a different fold for validation and using the remaining ones for training. This setting ensured that a test prediction could be acquired for each patient in a biomarker cohort and allowed us to assess the variability of model performance. The images were partitioned at the patient level and a biomarker profile with less than 10 positive patients was discarded from the study.

Pre-processing pipeline and training details: A proprietary convolutional neural network (CNN) was used for predicting molecular profiles from H&E images as illustrated in Figure 1. A single CNN was end-to-end trained from scratch for each biomarker and fold, yielding a total of 13,443 unique models that were used to obtain the results presented in this study. Each model was trained on a set of 256x256 tiles acquired from whole slide images (WSI) stained with H&E. A standard deviation filter was used to eliminate the tiles which do not contain any relevant information, allowing us to extract the tissue from the rest of the image. A slide was discarded from analysis if it contained fewer than 10 tiles after the filtering process. Macenko colour and brightness normalization [65] was applied to the remaining tiles before they were assigned with a ground-truth molecular profile (see **Biomarker acquisition**).

A CNN consisted of a feature extractor (encoder), a decoder and a classification module. The encoder can capture the tissue properties within tiles throughout a set of convolutional filters applied to tiles at various layers of depth, effectively encoding the high-level visual features into a d-dimensional feature vector, where d depends on the architecture of the CNN. These vectors are regarded as the fingerprints of the tiles and are submitted to both the decoder and the classification module. The decoder module takes a *d*-dimensional embedding as input and returns an output of the same shape as the original tile that the embedding represents. It consists of a series of transposed convolutional and upsampling layers, which resembles an inverted copy of the CNN used for feature extraction, known as the encoder. Its purpose is to reconstruct the original tile from the latent vector to achieve better representations of each tile that do not contain irrelevant features. In parallel to the decoder, the feature vector representative of each tile is also submitted to the classification module, which consists of a fully connected layer. The output classification score is then compared to the label of the tile's parent WSI. All modules are trained end-to-end and each tile is given a score, i.e. confidence of it being positive. Finally, the scores of all the sampled tiles from each WSI are aggregated via mean pooling to produce the final slide-level scores.

Model hyperparameters and the CNN architecture were determined based on a relevant benchmark analysis from the clinical validation study of a deep learning model developed for molecular profiling

of breast cancer [self-ref]. We adopted the best-performing model's feature extractor network (based on a "resnet34" architecture [66]) and hyper-parameters to configure a CNN for each biomarker in the current pan-cancer study. Each model was trained for 10 epochs using the Adam optimiser with a learning rate of 0.0001. 200 tiles were randomly sampled from each of the training slides and oversampling was applied to the tiles from the underrepresented class to tackle the class in-balance problem. During validation, predictions across all tiles were averaged to determine a slide-level prediction. Validation AUC was monitored as the target metric to select the final model during training.

Performance characteristics and statistical procedures: Performance of a model was measured with the area under receiver operating characteristic curve (AUC), which plots the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) across different predictive thresholds. An AUC of 0.5 denotes a random model, while a perfect model that can predict all samples correctly yields an AUC of 1. For each biomarker, we reported the performance as the average AUC across the three models (unless otherwise specified), accompanied with the standard deviation (denoted with ± where appropriate).

Statistical significance of results were determined with a two-sided t-test. For group-level analysis, the AUC values were compared against randomly sampled values of the same underlying distribution. To determine the significance of the predictability at biomarker level, we applied the test on the prediction scores obtained from the true negative and positive cases. The reported p values were corrected for false discovery rate (FDR).

Acknowledgement

The results shown in this study are partially based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga).

The authors would also like to thank Oscar Maiques Carlos and Jakob Kather for the discussion prior, during and after the writing of this publication.

Conflict of Interest

None reported.

References

[1] J. C. Smith and J. M. Sheltzer, 'Systematic identification of mutations and copy number alterations associated with cancer patient prognosis', *eLife*, vol. 7, p. e39217, Dec. 2018, doi: 10.7554/eLife.39217.

[2] E. R. Malone, M. Oliva, P. J. B. Sabatini, T. L. Stockley, and L. L. Siu, 'Molecular profiling for precision cancer therapies', *Genome Med.*, vol. 12, no. 1, p. 8, Dec. 2020, doi: 10.1186/s13073-019-0703-1.

[3] The Cancer Genome Atlas Research Network *et al.*, 'The Cancer Genome Atlas Pan-Cancer

analysis project', *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, doi: 10.1038/ng.2764.

[4] J. Wang and B. Xu, 'Targeted therapeutic options and future perspectives for HER2-positive breast cancer', *Signal Transduct. Target. Ther.*, vol. 4, no. 1, p. 34, Dec. 2019, doi: 10.1038/s41392-019-0069-2.

[5] D. Kazandjian, G. M. Blumenthal, W. Yuan, K. He, P. Keegan, and R. Pazdur, 'FDA Approval of Gefitinib for the Treatment of Patients with Metastatic *EGFR* Mutation–Positive Non–Small Cell Lung Cancer', *Clin. Cancer Res.*, vol. 22, no. 6, pp. 1307–1312, Mar. 2016, doi: 10.1158/1078-0432.CCR-15-2266.

[6] E. T. Tanda *et al.*, 'Current State of Target Treatment in BRAF Mutated Melanoma', *Front. Mol. Biosci.*, vol. 7, p. 154, Jul. 2020, doi: 10.3389/fmolb.2020.00154.

[7] O. Evans and R. Manchanda, 'Population-based Genetic Testing for Precision Prevention', *Cancer Prev. Res. (Phila. Pa.)*, vol. 13, no. 8, pp. 643–648, Aug. 2020, doi: 10.1158/1940-6207.CAPR-20-0002.

[8] K. A. Tendl and Z. Bago-Horvath, 'Molecular profiling in breast cancer—ready for clinical routine?', *Memo - Mag. Eur. Med. Oncol.*, vol. 13, no. 4, pp. 445–449, Dec. 2020, doi: 10.1007/s12254-020-00578-0.

[9] M. Rusch *et al.*, 'Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome', *Nat. Commun.*, vol. 9, no. 1, p. 3962, Dec. 2018, doi: 10.1038/s41467-018-06485-7.

[10] N. Naik *et al.*, 'Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains', *Nat. Commun.*, vol. 11, no. 1, p. 5727, Dec. 2020, doi: 10.1038/s41467-020-19334-3.

[11] A. C. Wolff *et al.*, 'Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update', *J. Clin. Oncol.*, vol. 36, no. 20, pp. 2105–2122, Jul. 2018, doi: 10.1200/JCO.2018.77.8738.

[12] J. N. Kather *et al.*, 'Pan-cancer image-based detection of clinically actionable genetic alterations', *Nat. Cancer*, vol. 1, no. 8, pp. 789–799, Aug. 2020, doi: 10.1038/s43018-020-0087-6.

[13] N. Coudray *et al.*, 'Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning', *Nat. Med.*, vol. 24, no. 10, pp. 1559–1567, Oct. 2018, doi: 10.1038/s41591-018-0177-5.

[14] J. N. Kather *et al.*, 'Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer', *Nat. Med.*, vol. 25, no. 7, pp. 1054–1056, Jul. 2019, doi: 10.1038/s41591-019-0462-y.

[15] Y. Fu *et al.*, 'Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis', *Nat. Cancer*, vol. 1, no. 8, pp. 800–810, Aug. 2020, doi: 10.1038/s43018-020-0085-8.

[16] B. Schmauch *et al.*, 'A deep learning model to predict RNA-Seq expression of tumours from whole slide images', *Nat. Commun.*, vol. 11, no. 1, p. 3877, Dec. 2020, doi: 10.1038/s41467-020-17678-4.

[17] H. D. Couture *et al.*, 'Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype', *Npj Breast Cancer*, vol. 4, no. 1, p. 30, Dec. 2018, doi: 10.1038/s41523-018-0079-1.

[18] R. R. Rawat *et al.*, 'Deep learned tissue "fingerprints" classify breast cancers by ER/PR/Her2 status from H&E images', *Sci. Rep.*, vol. 10, no. 1, p. 7275, Dec. 2020, doi: 10.1038/s41598-020-64156-4.

[19] D. Bychkov *et al.*, 'Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy', *Sci. Rep.*, vol. 11, no. 1, p. 4037, Dec. 2021, doi: 10.1038/s41598-021-83102-6.

[20] K. Sirinukunwattana *et al.*, 'Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning', *Gut*, vol. 70, no. 3, pp. 544–554, Mar. 2021, doi: 10.1136/gutjnl-2019-319866.

[21] A. Echle *et al.*, 'Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning', *Gastroenterology*, vol. 159, no. 4, pp. 1406-1416.e11, Oct. 2020, doi: 10.1053/j.gastro.2020.06.021.

[22] L. Sha *et al.*, 'Multi-field-of-view deep learning model predicts nonsmall cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images', *J. Pathol. Inform.*, vol. 10, no. 1, p. 24, 2019, doi: 10.4103/jpi.jpi_24_19.

[23] A. J. Schaumberg, M. A. Rubin, and T. J. Fuchs, 'H&E-stained Whole Slide Image Deep Learning Predicts SPOP Mutation State in Prostate Cancer', Pathology, preprint, Jul. 2016. doi: 10.1101/064279.

[24] H. Zhang *et al.*, 'Predicting Tumor Mutational Burden from Liver Cancer Pathological Images Using Convolutional Neural Network', in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, Nov. 2019, pp. 920–925. doi: 10.1109/BIBM47256.2019.8983139.

[25] R. H. Kim *et al.*, 'A Deep Learning Approach for Rapid Mutational Screening in Melanoma', Pathology, preprint, Apr. 2019. doi: 10.1101/610311.

[26] D. Anand, K. Yashashwi, N. Kumar, S. Rane, P. H. Gann, and A. Sethi, 'Weakly supervised learning on unannotated H&E-stained slides predicts *BRAF* mutation in thyroid cancer with high accuracy', *J. Pathol.*, vol. 255, no. 3, pp. 232–242, Nov. 2021, doi: 10.1002/path.5773.

[27] D. Bychkov *et al.*, 'Deep learning based tissue analysis predicts outcome in colorectal cancer', *Sci. Rep.*, vol. 8, no. 1, p. 3395, Dec. 2018, doi: 10.1038/s41598-018-21758-3.

[28] P. Courtiol *et al.*, 'Deep learning-based classification of mesothelioma improves prediction of patient outcome', *Nat. Med.*, vol. 25, no. 10, pp. 1519–1525, Oct. 2019, doi: 10.1038/s41591-019-0583-3.

[29] N. Harder *et al.*, 'Automatic discovery of image-based signatures for ipilimumab response prediction in malignant melanoma', *Sci. Rep.*, vol. 9, no. 1, p. 7449, Dec. 2019, doi: 10.1038/s41598-019-43525-8.

[30] J. N. Kather *et al.*, 'Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study', *PLOS Med.*, vol. 16, no. 1, p. e1002730, Jan. 2019, doi: 10.1371/journal.pmed.1002730.

[31] P. Mobadersany *et al.*, 'Predicting cancer outcomes from histology and genomics using convolutional networks', *Proc. Natl. Acad. Sci.*, vol. 115, no. 13, pp. E2970–E2979, Mar. 2018, doi: 10.1073/pnas.1717139115.

[32] A. Echle, N. T. Rindtorff, T. J. Brinker, T. Luedde, A. T. Pearson, and J. N. Kather, 'Deep learning in cancer pathology: a new generation of clinical biomarkers', *Br. J. Cancer*, vol. 124, no. 4, pp. 686–696, Feb. 2021, doi: 10.1038/s41416-020-01122-x.

[33] J. A. Diao *et al.*, 'Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes', *Nat. Commun.*, vol. 12, no. 1, p. 1613, Dec. 2021, doi: 10.1038/s41467-021-21896-9.

[34] J. Noorbakhsh *et al.*, 'Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images', *Nat. Commun.*, vol. 11, no. 1, p. 6367, Dec. 2020, doi: 10.1038/s41467-020-20030-5.

[35] M. M. Li *et al.*, 'Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer', *J. Mol. Diagn.*, vol. 19, no. 1, pp. 4–23, Jan. 2017, doi: 10.1016/j.jmoldx.2016.10.002.

[36] E. Segal, N. Friedman, N. Kaminski, A. Regev, and D. Koller, 'From signatures to models: understanding cancer using microarrays', *Nat. Genet.*, vol. 37, no. S6, pp. S38–S45, Jun. 2005, doi: 10.1038/ng1561.

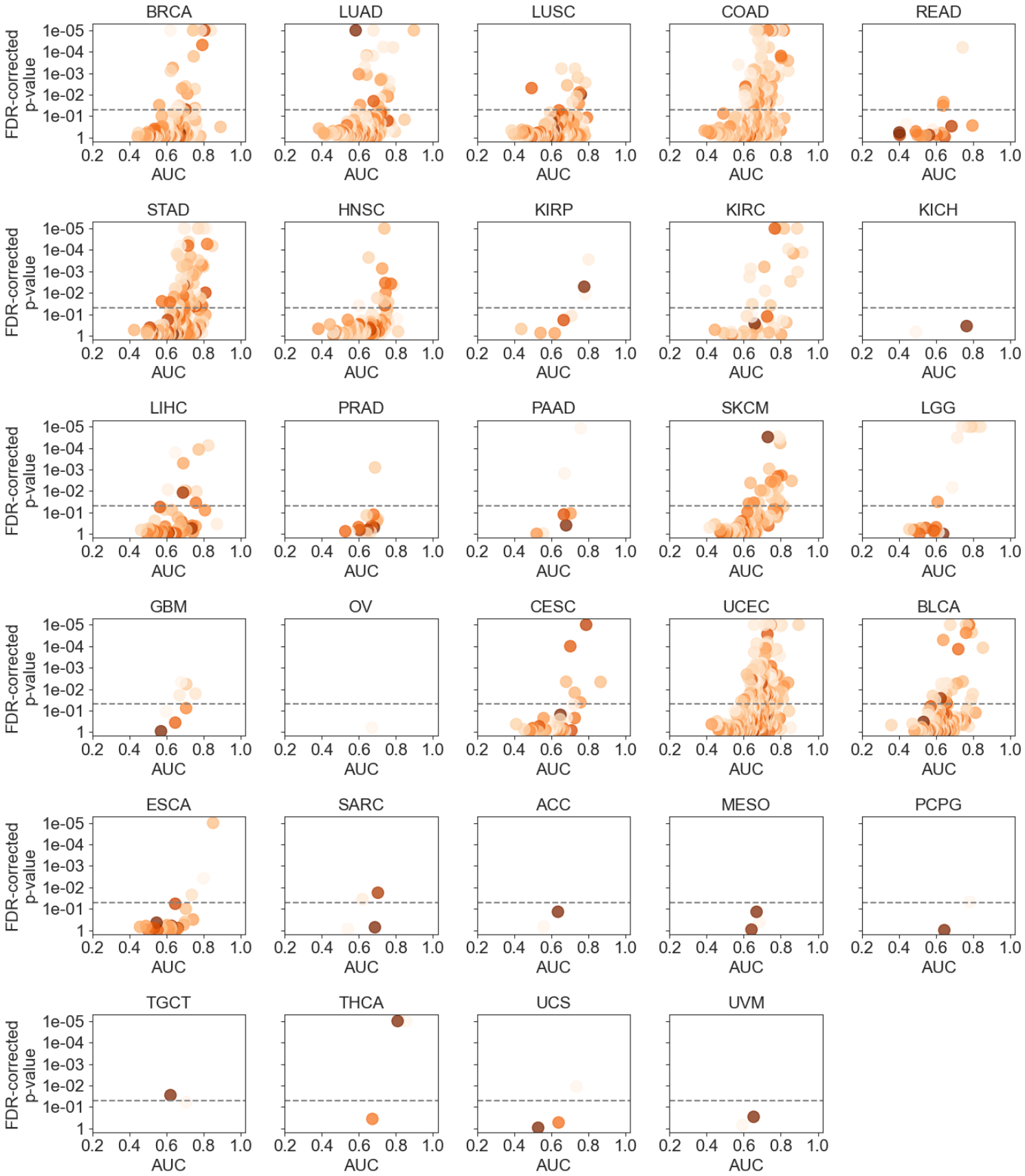
[37] Z. Wang *et al.*, 'Prognostic and clinicopathological value of p53 expression in renal cell carcinoma: a meta-analysis', *Oncotarget*, vol. 8, no. 60, pp. 102361–102370, Nov. 2017, doi: 10.18632/oncotarget.21971.

[38] V. Thorsson *et al.*, 'The Immune Landscape of Cancer', *Immunity*, vol. 48, no. 4, pp. 812-830.e14,

Apr. 2018, doi: 10.1016/j.immuni.2018.03.023.
[39] A. C. Berger *et al.*, 'A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers', *Cancer Cell*, vol. 33, no. 4, pp. 690-705.e9, Apr. 2018, doi: 10.1016/j.ccell.2018.03.014.
[40] Y. Liu *et al.*, 'Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas', *Cancer Cell*, vol. 33, no. 4, pp. 721-735.e8, Apr. 2018, doi: 10.1016/j.ccell.2018.03.010.
[41] E. Wulczyn *et al.*, 'Deep learning-based survival prediction for multiple cancer types using histopathology images', *PLOS ONE*, vol. 15, no. 6, p. e0233678, Jun. 2020, doi: 10.1371/journal.pone.0233678.
[42] O.-J. Skrede *et al.*, 'Deep learning for prediction of colorectal cancer outcome: a discovery and validation study', *The Lancet*, vol. 395, no. 10221, pp. 350–360, Feb. 2020, doi: 10.1016/S0140-6736(19)32998-8.
[43] C. Saillard *et al.*, 'Predicting Survival After Hepatocellular Carcinoma Resection Using Deep Learning on Histological Slides', *Hepatology*, vol. 72, no. 6, pp. 2000–2013, Dec. 2020, doi: 10.1002/hep.31207.
[44] J. Liu *et al.*, 'An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics', *Cell*, vol. 173, no. 2, pp. 400-416.e11, Apr. 2018, doi: 10.1016/j.cell.2018.02.052.
[45] N. Rivlin, R. Brosh, M. Oren, and V. Rotter, 'Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis', *Genes Cancer*, vol. 2, no. 4, pp. 466–474, Apr. 2011, doi: 10.1177/1947601911408889.
[46] Z. Chen, C. Chen, L. Li, T. Zhang, and X. Wang, 'Pan-Cancer Analysis Reveals That E1A Binding Protein p300 Mutations Increase Genome Instability and Antitumor Immunity', *Front. Cell Dev. Biol.*, vol. 9, p. 729927, Sep. 2021, doi: 10.3389/fcell.2021.729927.
[47] M. Farzin *et al.*, 'Loss of expression of BAP1 predicts longer survival in mesothelioma', *Pathology (Phila.)*, vol. 47, no. 4, pp. 302–307, Jun. 2015, doi: 10.1097/PAT.0000000000000250.
[48] G. Tan *et al.*, 'The critical role of BAP1 mutation in the prognosis and treatment selection of kidney renal clear cell carcinoma', *Transl. Androl. Urol.*, vol. 9, no. 4, pp. 1725–1734, Aug. 2020, doi: 10.21037/tau-20-1079.
[49] H. Hua, Q. Kong, H. Zhang, J. Wang, T. Luo, and Y. Jiang, 'Targeting mTOR for cancer therapy', *J. Hematol. Oncol.J Hematol Oncol*, vol. 12, no. 1, p. 71, Dec. 2019, doi: 10.1186/s13045-019-0754-1.
[50] B. C. Grabiner *et al.*, 'A Diverse Array of Cancer-Associated *MTOR* Mutations Are Hyperactivating and Can Predict Rapamycin Sensitivity', *Cancer Discov.*, vol. 4, no. 5, pp. 554–563, May 2014, doi: 10.1158/2159-8290.CD-13-0929.
[51] X. Jin, L. Zhu, Z. Cui, J. Tang, M. Xie, and G. Ren, 'Elevated expression of GNAS promotes breast cancer cell proliferation and migration via the PI3K/AKT/Snail1/E-cadherin axis', *Clin. Transl. Oncol.*, vol. 21, no. 9, pp. 1207–1219, Sep. 2019, doi: 10.1007/s12094-019-02042-w.
[52] M. Sinkala, N. Mulder, and D. Patrick Martin, 'Metabolic gene alterations impact the clinical aggressiveness and drug responses of 32 human cancers', *Commun. Biol.*, vol. 2, no. 1, p. 414, Dec. 2019, doi: 10.1038/s42003-019-0666-1.
[53] F. Sanchez-Vega *et al.*, 'Oncogenic Signaling Pathways in The Cancer Genome Atlas', *Cell*, vol. 173, no. 2, pp. 321-337.e10, Apr. 2018, doi: 10.1016/j.cell.2018.03.035.
[54] U. E. Martinez-Outschoorn, M. Peiris-Pagés, R. G. Pestell, F. Sotgia, and M. P. Lisanti, 'Cancer metabolism: a therapeutic perspective', *Nat. Rev. Clin. Oncol.*, vol. 14, no. 1, pp. 11–31, Jan. 2017, doi: 10.1038/nrclinonc.2016.60.
[55] S. R. Rosario, M. D. Long, H. C. Affronti, A. M. Rowsam, K. H. Eng, and D. J. Smiraglia, 'Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas', *Nat. Commun.*, vol. 9, no. 1, p. 5330, Dec. 2018, doi: 10.1038/s41467-018-07232-8.
[56] H. Zhang, Y. Chen, and F. Li, 'Predicting Anticancer Drug Response With Deep Learning Constrained by Signaling Pathways', *Front. Bioinforma.*, vol. 1, p. 639349, Apr. 2021, doi: 10.3389/fbinf.2021.639349.

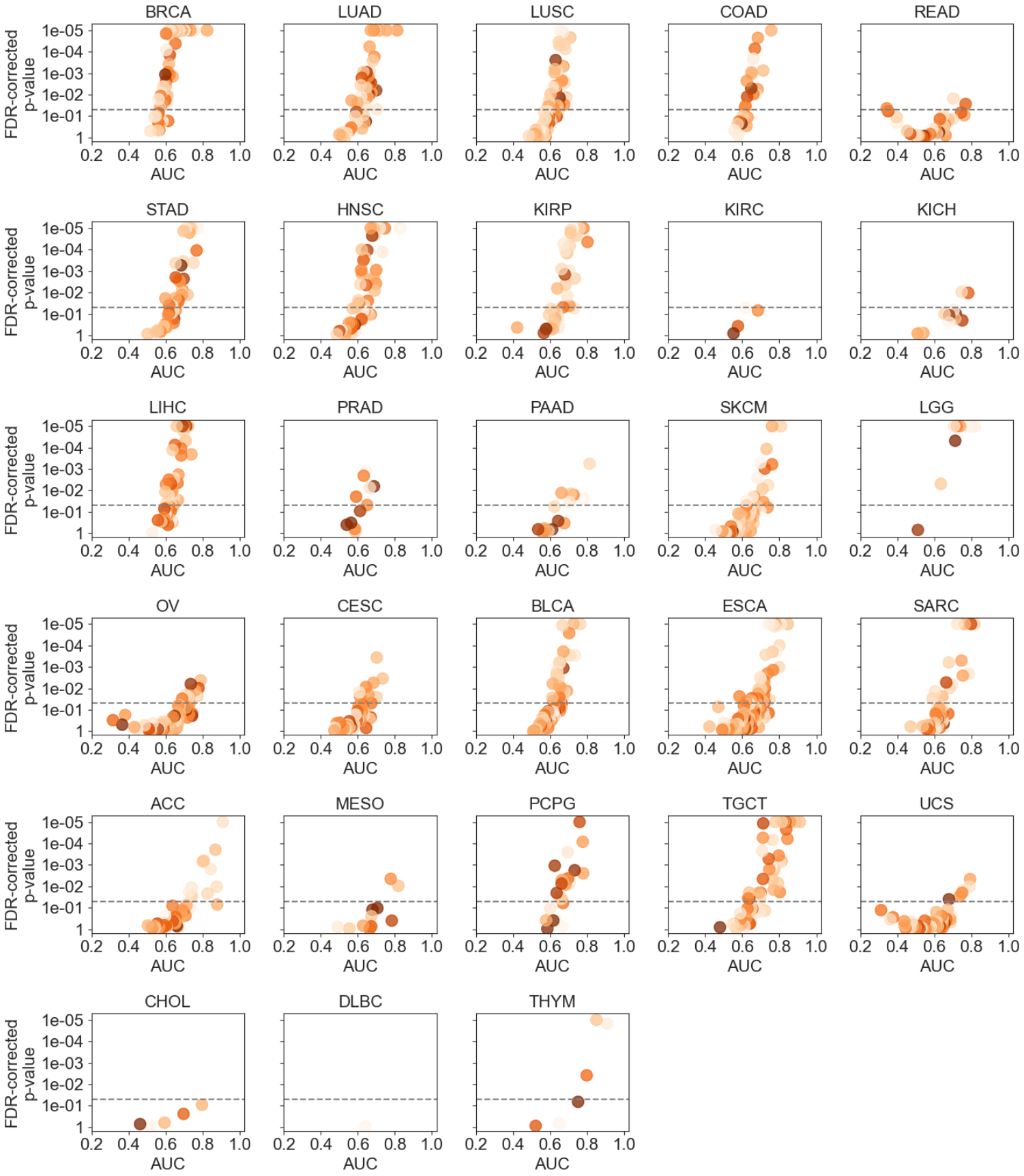
[57] T. R. Kau, J. C. Way, and P. A. Silver, 'Nuclear transport and cancer: from mechanism to intervention', *Nat. Rev. Cancer*, vol. 4, no. 2, pp. 106–117, Feb. 2004, doi: 10.1038/nrc1274.
[58] A. Siddiqui and P. Ceppi, 'A non-proliferative role of pyrimidine metabolism in cancer', *Mol. Metab.*, vol. 35, p. 100962, May 2020, doi: 10.1016/j.molmet.2020.02.005.
[59] Y. Ma *et al.*, 'Fatty acid oxidation: An emerging facet of metabolic transformation in cancer', *Cancer Lett.*, vol. 435, pp. 92–100, Oct. 2018, doi: 10.1016/j.canlet.2018.08.006.
[60] P. Hermanek and C. Wittekind, 'Residual tumor (R) classification and prognosis', *Semin. Surg. Oncol.*, vol. 10, no. 1, pp. 12–20, Jan. 1994, doi: 10.1002/ssu.2980100105.
[61] Variant Interpretation for Cancer Consortium *et al.*, 'A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer', *Nat. Genet.*, vol. 52, no. 4, pp. 448–457, Apr. 2020, doi: 10.1038/s41588-020-0603-8.
[62] M. Uhlen *et al.*, 'A pathology atlas of the human cancer transcriptome', *Science*, vol. 357, no. 6352, p. eaan2507, Aug. 2017, doi: 10.1126/science.aan2507.
[63] Y. Benjamini and Y. Hochberg, 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *J. R. Stat. Soc. Ser. B Methodol.*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.
[64] E. Moiso, 'Manual curation of TCGA treatment data and identification of potential markers of therapy response', Oncology, preprint, May 2021. doi: 10.1101/2021.04.30.21251941.
[65] M. Macenko *et al.*, 'A method for normalizing histology slides for quantitative analysis', in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Boston, MA, USA, Jun. 2009, pp. 1107–1110. doi: 10.1109/ISBI.2009.5193250.
[66] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

Extended Data



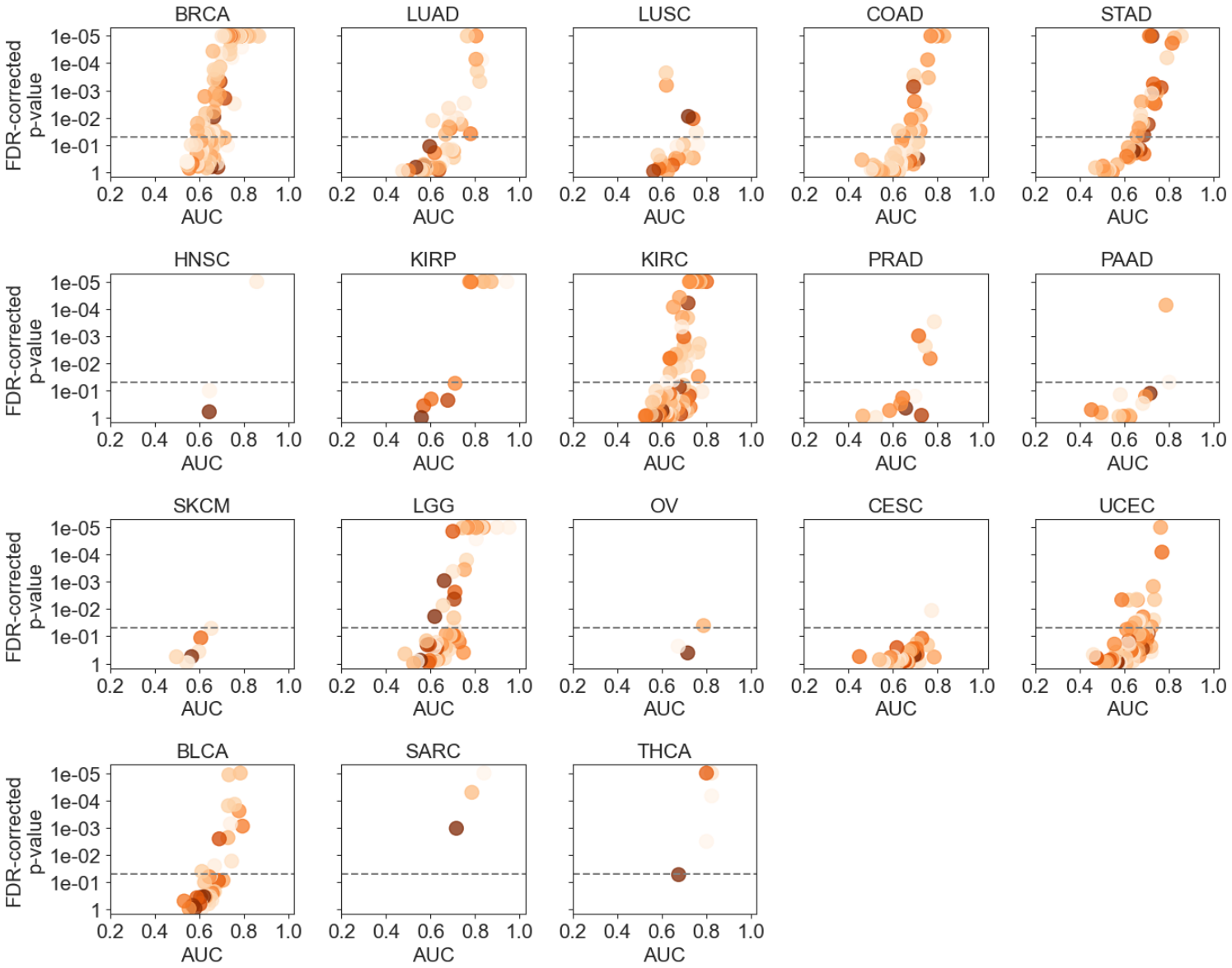
Extended Data Figure 1: Scatter plots showing the average test AUC for each model trained to predict genetic alterations across all tested cancer types. A two-sided t-test was applied to prediction scores of each model to assess the statistical

significance and the corresponding p-values were corrected for false discovery rate (FDR). The y-axis of each plot was inverted and the p-values were log-transformed for visualisation purposes. P values smaller than 1e-05 were set to 1e-05 to avoid numerical errors during transformation. The statistical significance threshold of 0.05 is marked with a dashed line. Marker shading indicates the intra-biomarker standard deviation, where lighter colors correspond to a deviation close to 0.

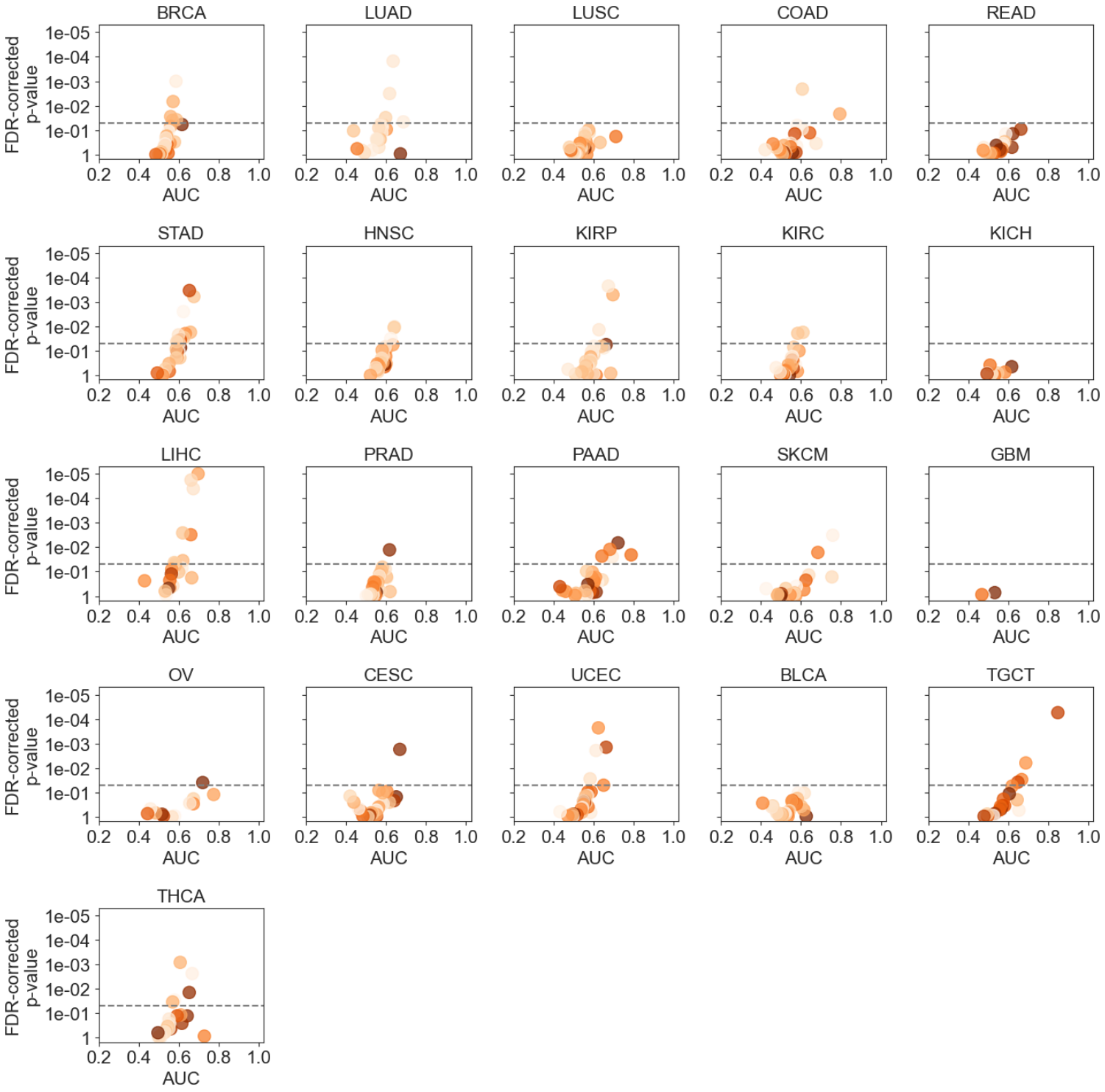


Extended Data Figure 2: Scatter plots showing the average test AUC for each model trained to predict over-/under-expression status for the transcriptomic driver genes across all investigated cancer types. A two-sided t-test was applied to prediction

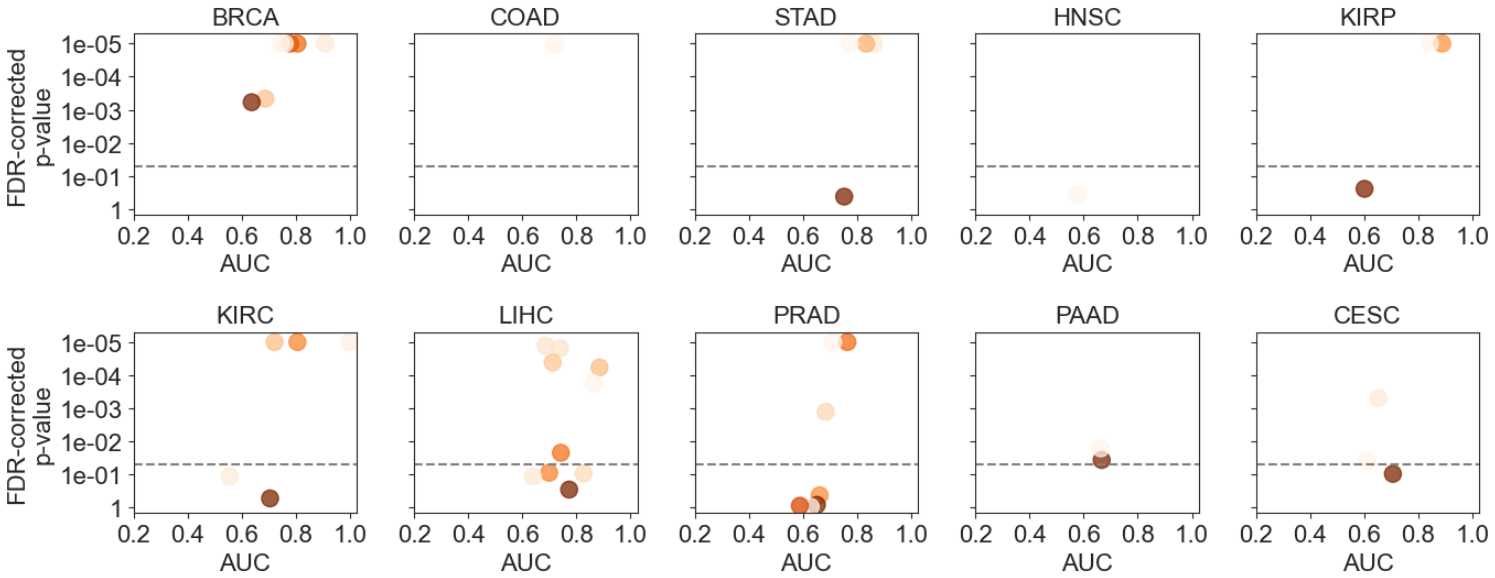
scores of each model to assess the statistical significance and the corresponding p-values were corrected for false discovery rate (FDR). Please refer to the caption of **Extended Data Figure 1** for a detailed explanation of the visualisation.



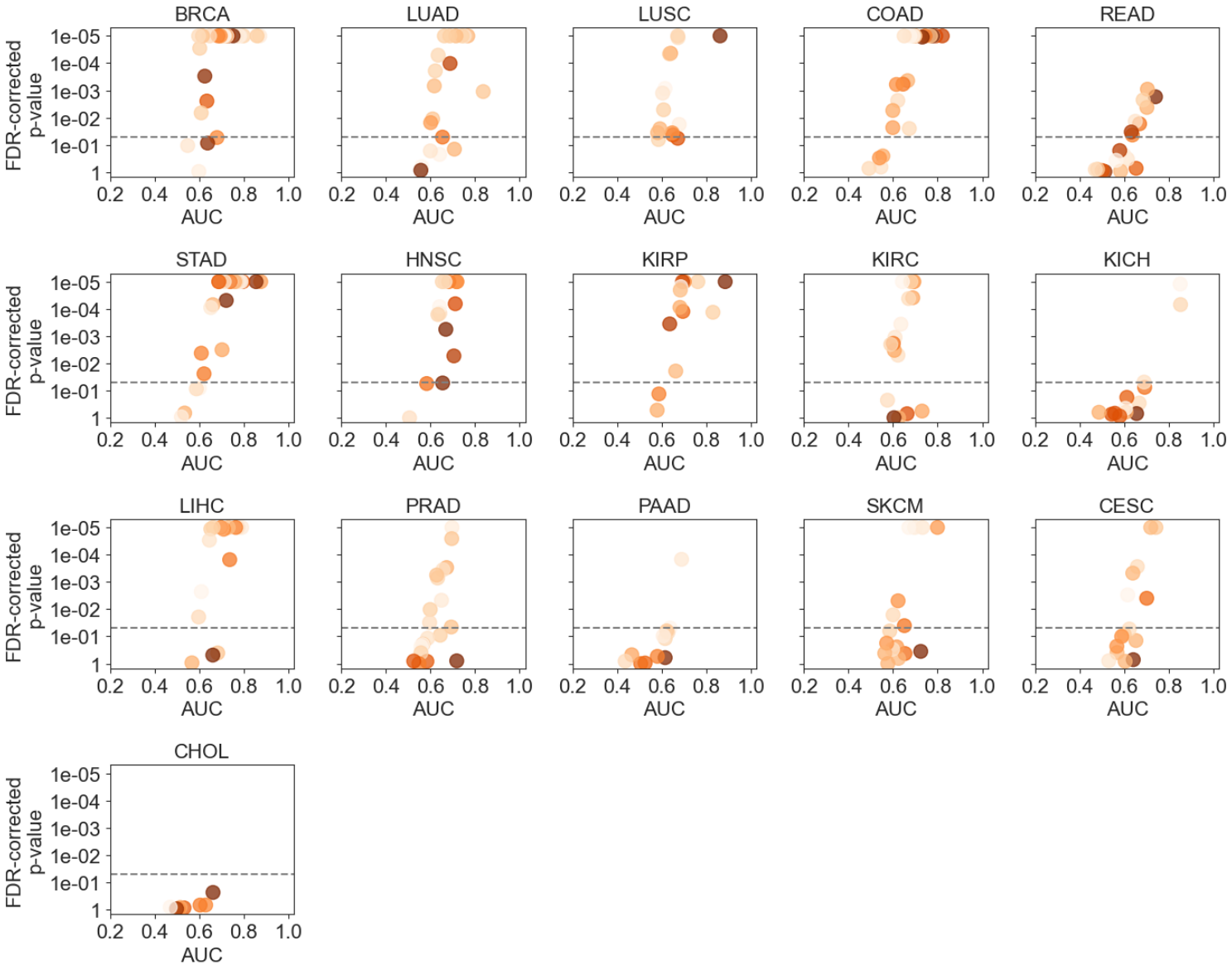
Extended Data Figure 3: Scatter plots showing the average test AUC for each model trained to predict over-/under-expression status of proteomes associated with driver genes across all investigated cancer types. A two-sided t-test was applied to prediction scores of each model to assess the statistical significance and the corresponding p-values were corrected for false discovery rate (FDR). Please refer to the caption of **Extended Data Figure 1** for a detailed explanation of the visualisation.



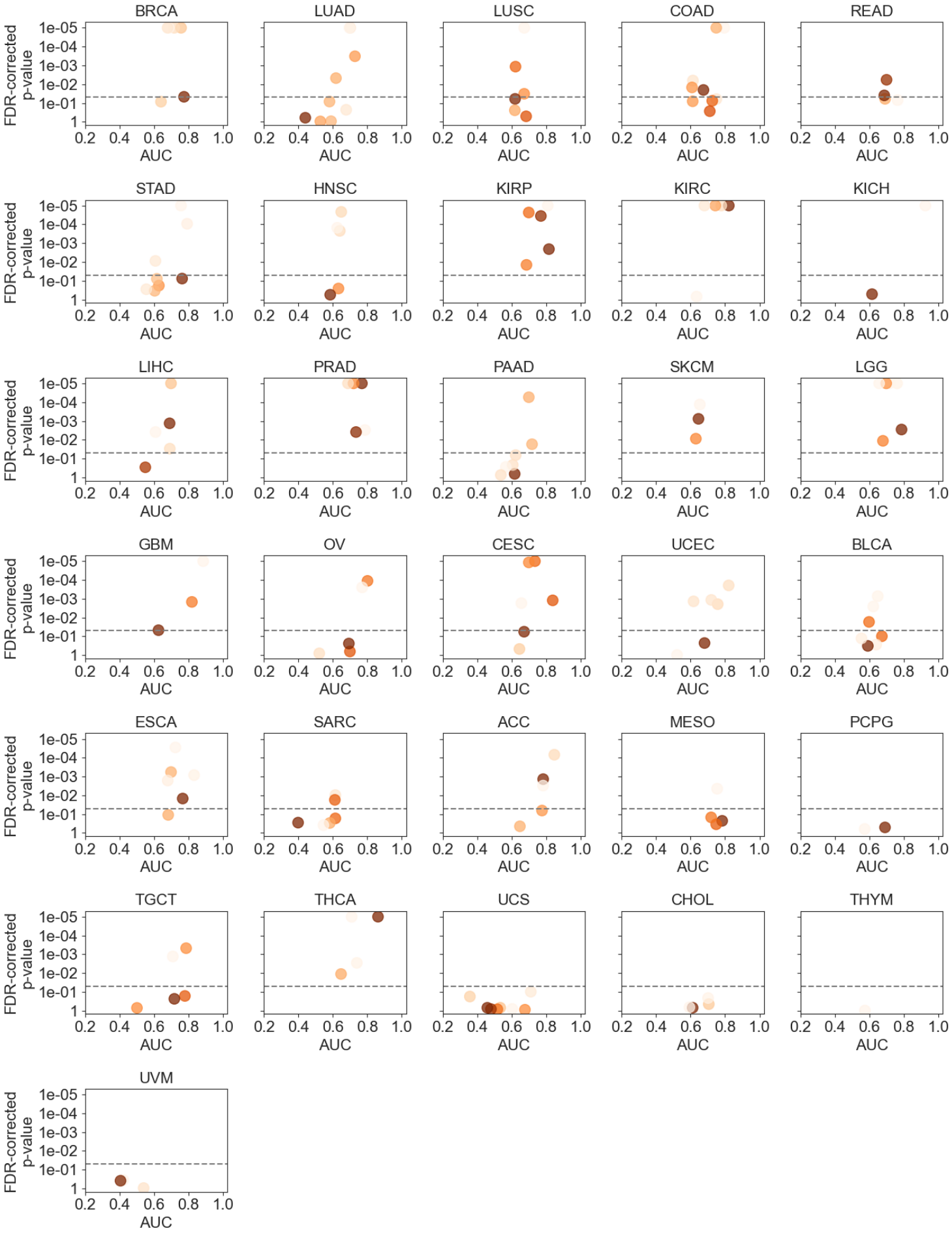
Extended Data Figure 4: Scatter plots showing the average test AUC for each model trained to predict metabolomic pathways across all investigated cancer types. A two-sided t-test was applied to prediction scores of each model to assess the statistical significance and the corresponding p-values were corrected for false discovery rate (FDR). Please refer to the caption of **Extended Data Figure 1** for a detailed explanation of the visualisation.



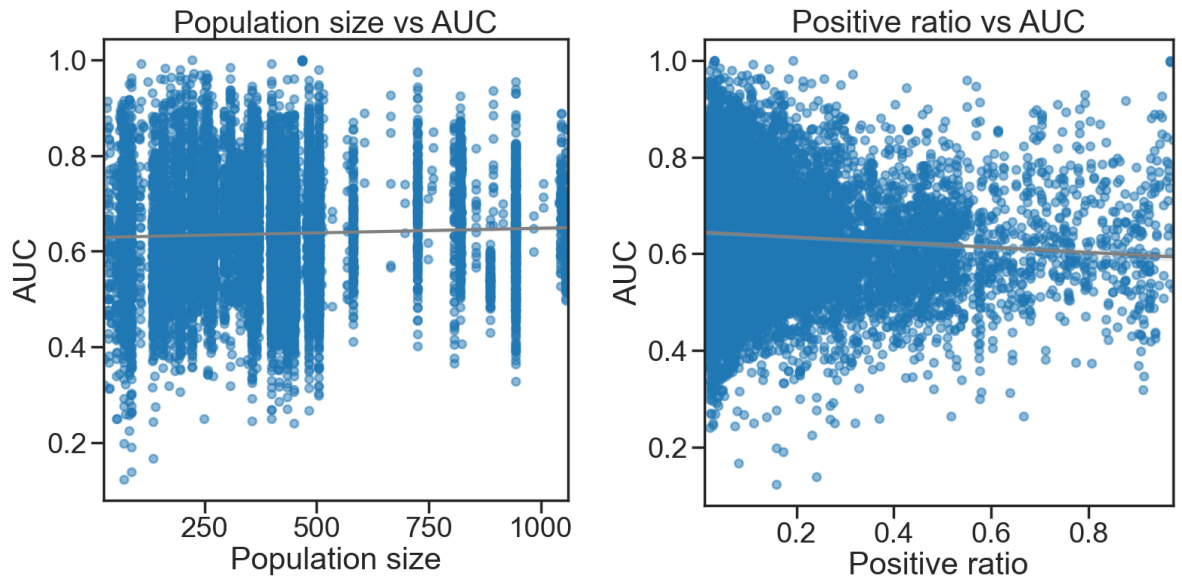
Extended Data Figure 5: Scatter plots showing the average test AUC for each model trained to predict standard clinical biomarkers across all investigated cancer types. A two-sided t-test was applied to prediction scores of each model to assess the statistical significance and the corresponding p-values were corrected for false discovery rate (FDR). Please refer to the caption of **Extended Data Figure 6** for a detailed explanation of the visualisation.



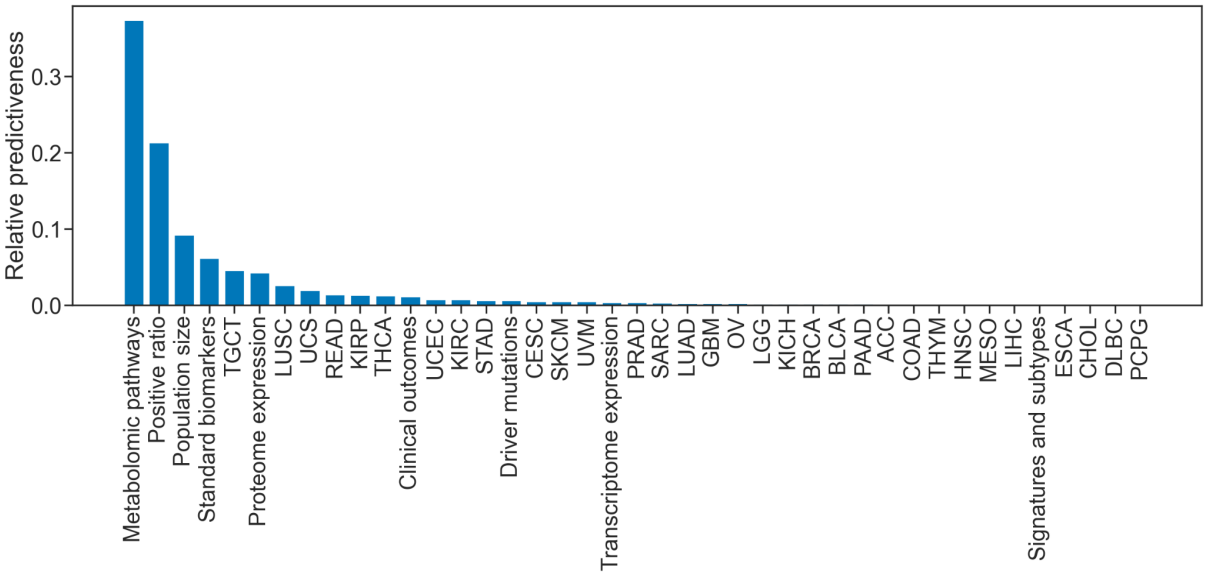
Extended Data Figure 6: Scatter plots showing the average test AUC for each model trained to predict molecular subtypes and gene signatures across all investigated cancer types. A two-sided t-test was applied to prediction scores of each model to assess the statistical significance and the corresponding p-values were corrected for false discovery rate (FDR). Please refer to the caption of **Extended Data Figure 1** for a detailed explanation of the visualisation.



Extended Data Figure 7: Scatter plots showing the average test AUC for each model trained to predict clinical outcomes and treatment responses across all investigated cancer types. A two-sided t-test was applied to prediction scores of each model to assess the statistical significance and the corresponding p-values were corrected for false discovery rate (FDR). Please refer to the caption of **Extended Data Figure 1** for a detailed explanation of the visualisation.



Extended Data Figure 8 | Impact of sample size and positive class ratio on the prediction performance: To assess the impact of sample size and positive class ratio on the prediction performance, we measured their linear relationship to the biomarker AUC values with Pearson correlation analysis. Scatter plots show the correlation of AUC values (i.e. intra-biomarker mean AUC of the three cross-validated models for each biomarker) with sample size (i.e. number of total samples per biomarker, *left*) and positive ratio (i.e. number of positive samples over the size of whole population, *right*) of all biomarkers. Pearson correlation coefficient (PCC) between the sample size and the AUC values was 0.036 (p-value < 1e-05), indicating no linear relationship between the two variables. Similarly, a PCC of -0.084 (p-value < 1e-05) was obtained for the positive ratio, showing almost no impact of prevalence on predictiability.



Extended Data Figure 9 | Importance of variables on biomarker predictability: Considering the very low impact of positive ratio and population size onto the model performance (**Extended Data Figure 8**) we performed a simple experiment to estimate the importance of variables on the biomarker predictability. Alongside sample size and positive ratio, we included omic type (e.g. driver mutations) and cancer type into our analysis. A random forest regression (RFR) model was fitted on the data to predict the AUC values. During training, the RFR model assigns a weight to each feature, which, in turn, can be used to estimate the "predictiveness" of a variable for regressing the AUC values). Overall, the combined impact of the omic types was the largest, with metabolomic pathways being the most important feature for model predictictibility. This was followed by positive ratio and population size, both of which were assigned more importance than most of the omic and cancer types, when comparisons were done on an individual basis. We also observed that prevalence influenced the predictability more than the sample size. The impact of cancer type, on the other hand, was rather limited. On aggregate, however, the impact of cancer type outsized the impact of the prevalence and population size.

Abbr.	Cancer Name	Mean AUC	Std. of AUCs
MESO	Mesothelioma	0.692	0.137
LGG	Brain Lower Grade Glioma	0.684	0.118
TGCT	Testicular Germ Cell Tumors	0.683	0.132
ACC	Adrenocortical carcinoma	0.676	0.151
KIRP	Kidney renal papillary cell carcinoma	0.673	0.117
PCPG	Pheochromocytoma and Paraganglioma	0.670	0.086
KIRC	Kidney renal clear cell carcinoma	0.659	0.115
STAD	Stomach adenocarcinoma	0.657	0.113
GBM	Glioblastoma multiforme	0.653	0.164
ESCA	Esophageal carcinoma	0.653	0.114
BRCA	Breast invasive carcinoma	0.649	0.101
LIHC	Liver hepatocellular carcinoma	0.643	0.103
UCEC	Uterine Corpus Endometrial Carcinoma	0.641	0.101
SARC	Sarcoma	0.640	0.108
COAD	Colon adenocarcinoma	0.639	0.110
THCA	Thyroid carcinoma	0.634	0.123
KICH	Kidney Chromophobe	0.631	0.144
OV	Ovarian serous cystadenocarcinoma	0.629	0.136
HNSC	Head and Neck squamous cell carcinoma	0.628	0.101
SKCM	Skin Cutaneous Melanoma	0.627	0.118
LUAD	Lung adenocarcinoma	0.622	0.109
PRAD	Prostate adenocarcinoma	0.621	0.091
BLCA	Bladder Urothelial Carcinoma	0.620	0.104
PAAD	Pancreatic adenocarcinoma	0.616	0.115
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	0.612	0.109
LUSC	Lung squamous cell carcinoma	0.607	0.102
CHOL	Cholangiocarcinoma	0.594	0.137
READ	Rectum adenocarcinoma	0.586	0.125
UCS	Uterine Carcinosarcoma	0.585	0.158

Extended Data Table 1: Average performance and standard deviation for all cancer types. DLBC, UVM, and THYM were excluded from the table due to only constituting one to seven valid targets across all biomarker types

Abbr.	Num. images	Num. patients
LUSC	453	419
PRAD	449	403
COAD	442	434
SKCM	418	377
THCA	504	492
OV	105	104
BRCA	1061	992
STAD	358	333
DLBC	38	38
BLCA	450	379
CHOL	38	38
PAAD	204	180
UCEC	509	448
KIRP	269	245
LUAD	512	449
CESC	265	255
KIRC	499	493
THYM	179	120
LIHC	369	361
SARC	582	239
HNSC	434	414
READ	158	157
ESCA	157	155
LGG	823	472
KICH	120	108
GBM	666	233
MESO	81	70
PCPG	193	174
TGCT	253	148
UCS	87	53
UVM	53	53
ACC	225	54
Total	10954	8890

Extended Data Table 2: NUmber of whole slide images and patients included in the study across all cancer types. Cancer names corresponding to abbreviations are provided in **Extended Data Table 1.**