Artificial intelligence predicts immune and inflammatory gene signatures directly from hepatocellular carcinoma histology

Qinghe Zeng, Christophe Klein, Stefano Caruso, Pascale Maille, Narmin Ghaffari Laleh, Daniele Sommacale, Alexis Laurent, Giuliana Amaddeo, David Gentien, Audrey Rapinat, et al.

Artificial intelligence predicts immune and inflammatory gene signatures directly from hepatocellular carcinoma histology

Qinghe Zeng[1,2,3*], Christophe Klein[1,2*], Stefano Caruso[4], Pascale Maille[5,6,7], Narmin Ghaffari Laleh[8], Daniele Sommacale[9], Alexis Laurent[9], Giuliana Amaddeo[10], David Gentien[11], Audrey Rapinat[11], Hélène Regnault[10], Cécile Charpy[5], Cong Trung Nguyen[6,7], Christophe Tournigand[12], Raffaele Brustia[9], Jean Michel Pawlotsky[6,7], Jakob Nikolas Kather[8], Maria Chiara Maiuri[1,2], Nicolas Loménie[3#], Julien Calderaro[5,6,7#].

*Co-first authors

#Co-last authors

1. Center of Cellular Imaging and Cytometry, Institut National de la Santé et de la Recherche Médicale (INSERM) UMRS 1138, Cordeliers Research Center, Paris, France.
2.Sorbonne University, UMRS 1138, Cordeliers Research Center, Paris, France.
3. LIPADE, Université de Paris, Paris, France.
4. INSERM UMR-1162, Functional Genomics of Solid Tumors, Paris, France.
5. Assistance Publique-Hôpitaux de Paris, Henri Mondor-Albert Chenevier University Hospital, Department of Pathology, Créteil, France.
6.Université Paris Est Créteil, INSERM, IMRB, F-94010 Créteil, France.
7.INSERM, Unit U955, Team 18, Créteil, France.
8. Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany; Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany.
9.Assistance Publique-Hôpitaux de Paris, Henri Mondor-Albert Chenevier University Hospital, Department of Digestive and Hepatobiliary Surgery, Créteil, France.
10.Assistance Publique-Hôpitaux de Paris, Henri Mondor-Albert Chenevier University Hospital, Department of Hepatology, Créteil, France.
11.Institut Curie, PSL Research University, Translational Research Department, Genomics platform, Paris, F-75248 France.
12. Assistance Publique-Hôpitaux de Paris, Henri Mondor-Albert Chenevier University Hospital, Department of Medical Oncology, Créteil, France.
13. Assistance Publique-Hôpitaux de Paris, Henri Mondor-Albert Chenevier University Hospital, Department of Bacteriology and Virology, Créteil, France.

Authors contributions
Study conception and design: JC, NL, CK
Data collection: DS, AL, GA, CT, RB, HR, DG, AR, PM, CTN
Data analysis: QZ, CK, DG, AR, SC, NGL, JNK, CTN, JMP, MCM
Drafting the manuscript: QZ, CK, JC, NL, SC
Obtained funding: JC
Critical revision of the manuscript: All authors

Conflicts of Interest: JC consults for Crosscope, Keen Eye, and has received research funding from Fondation Bristol Myers Squibb pour la Recherche en Immuno-Oncologie.

Keywords: artificial intelligence, artificial intelligence, deep learning, pathology

Lay summary

Immune and inflammatory gene signatures may be associated with an increased sensitivity to immunotherapy in patients with advanced hepatocellular carcinoma. We show in the present study that the use of artificial intelligence based pathology allows to predict the activation of these signatures directly from histology.

Abstract

Background and aims
Patients with hepatocellular carcinoma (HCC) displaying overexpression of immune gene signatures are likely to be more sensitive to immunotherapy, however the use of such signatures in clinical settings remains challenging. We thus aimed, using artificial intelligence (AI) on whole-slide digital histological Images (WSI), to develop models able to predict the activation of 6 immune gene signatures.

Methods
AI Models were trained and validated in 2 different series of patients with HCC treated by surgical resection. Gene expression was investigated using RNA sequencing or NanoString technology. Three deep-learning approaches were investigated: patch-based, classic MIL and CLAM. Pathological reviewing of the most predictive tissue areas was performed for all gene signatures..

Results
The CLAM model showed the best overall performance in the discovery series. Its best-fold areas under the curve (AUC) for the prediction of tumors with upregulation of the immune gene signatures ranged from 0.74 to 0.87. The different models generalized well in the validation dataset with AUCs ranging from 0.81 to 0.91. Pathological analysis of highly predictive tissue areas showed enrichment in inflammatory cells, plasma cells, and neutrophils.

Conclusion
We have developed and validated AI based pathology models able to predict the activation of several immune and inflammatory genes signatures. Our approach also provides insights on the morphological features that impact the model predictions. This proof-of-concept study shows that AI based pathology could represent a novel type of biomarker that will ease the translation of our biological knowledge of HCC into clinical practice.

Introduction

Hepatocellular carcinoma (HCC) remains one of the leading causes of cancer related deaths worldwide and a global health challenge.[1] The vast majority (80–90%) of cases develop in patients with chronic liver disease or cirrhosis and the main risk factors are viral hepatitis, alcohol intake or metabolic syndrome.[1] Particular treatments, such as surgical resection, percutaneous ablation or liver transplantation, offer a chance of cure. However, more than two thirds of patients present with advanced disease and are therefore not eligible for these strategies.[2]

The landscape of therapeutic options is however rapidly evolving, and several drugs (including lenvatinib, cabozantinib, regaorafenib or ramucirumab) have recently been approved. [3–6] Immunotherapy also holds great promise to improve clinical outcomes. The combination of atezolizumab, an anti-PDL1 (Programmed death Ligand 1) monoclonal antibody, and bevacizumab, an anti-angiogenic agent, was shown to be superior to sorafenib and is now the standard of care for patients with advanced disease.[7] Other clinical trials testing immunomodulating agents, such as nivolumab, have turned out to be negative.[8] Durable responses are however observed with these drugs, and there are several clinical trials underway investigating their use in other clinical settings than advanced disease.[1] A fast and accurate identification of patients likely to respond is therefore a critical issue. In this line, by investigating HCC samples of advanced patients treated by nivolumab, Sangro et al. showed that several immune-related gene signatures were associated with survival.[9] These findings have also been very recently confirmed by Haber et al. who showed that upregulation in Interferon gamma signaling and antigen presentation was able to predict the overall response rate of patients with HCC treated by nivolumab or pembrolizumab.[10] These signatures noticeably include immune checkpoint inhibitors (*LAG3, CD274-PDL1*), cytolytic effectors (*GZMA, PRF1*), and molecules involved in antigen processing/presentation (*HLA-DMA, HLA-DMB*) and immune cells recruitment (*CXCL9, CXCL10, IFNG*) .

These results are consistent with data from other human solid cancers and suggest that such molecular biomarkers may help to 1) identify the potential responders to immunomodulating approaches and 2) spare side effects for patients that are not likely to benefit from these strategies.[11] Their use in a clinical setting is however challenging, as they require access to molecular biology platforms for nucleic acid extraction and their processing/sequencing. They also are highly dependent on the quality of samples, and prone to standardization issues.

On the other hand, histological slides are easily available from pathology departments. It is well-established that they contain an extensive amount of information that allow definitive diagnosis and clinical outcome prediction. The advent of digital pathology and artificial intelligence (AI) also allows to standardize their analysis and extract meaningful morphological features that are not easily accessible to the human eye.[12–15]

We thus aimed, using AI-based pathology, to develop deep learning models able to predict the activation of 6 immune gene signatures associated with response to immunotherapy in patients with HCC.

Material and methods

Patients and samples

For our discovery series, we used The Cancer Genome Atlas public data set (TCGA-LIHC). It consists of patients with primary HCC treated by surgical resection in more than 20 different clinical centers.[16] The inclusion criteria were as follows: 1) unequivocal morphological features of HCC (all slides were reviewed by a pathologist specializing in liver disease (JC), and cases with features suggestive of combined hepatocellular-cholangiocarcinoma were excluded), 2) one or more available digital histological slides from formalin fixed-paraffin embedded (FFPE) material and 3) available gene expression profiling (obtained by RNA sequencing). Data and slides were accessed and downloaded in March 2020.

The validation series consisted of primary HCC samples developed in patients treated by surgical resection in Henri Mondor University Hospital (Créteil, France). Inclusion criteria were : 1) surgical resection performed between 2010-2019, 2) histological diagnosis of HCC confirmed by a liver pathologist (JC) and 3) available slides and tissue blocks for gene expression experiments. For seven patients, pre-operative biopsies were available and included to test the models on this type of samples. The study was performed according to the declaration of Helsinki and was approved by an institutional review board (CPP Ile de France V). All necessary written informed consents were obtained from the patients.

The overall flow-chart of the study is presented on **Figure 1**.

RNA sequencing data processing and clustering analysis in the discovery series

The Fragments Per Kilobase Million (FPKM) counts selected from the TCGA-LIHC project were processed for normalization due to non linear bias induced by genomic screening. Log 2 transformation was then applied after adding 1 to the FPKM matrix. The Z-score approach was chosen for gene-wise standardization.

We further aimed to investigate the 6 immune gene signatures that were previously shown to be associated, in patients with advanced HCC, with improved response and survival rates after treatment by the anti-PD 1 monoclonal antibody nivolumab : “6-Gene Interferon Gamma” (6G IFNg) (*CXCL10, CXCL9, HLA-DRA, IDO1, IFNG, STAT1*) [17], “Gajewski 13-Gene Inflammatory” (Gajewski 13G) (CC*L2, CCL3, CCL4, CD8A, CXCL10, CXCL9, GZMK,*

HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, ICOS, IRF1, *CCL3* was not included in the Nanostring Panel so we also removed it from the discovery series) [18], "Inflammatory" (CD274 / PD-L1, CD8A, LAG3, STAT1) [9], "Interferon Gamma Biology" (IFNg biology) (*CCL5, CD27, CXCL9, CXCR6, IDO1, STAT1*) [17], "Ribas 10-Gene Interferon Gamma" (Ribas 10G) (*CCR5, CXCL10, CXCL11, CXCL9, GZMA, HLA-DRA, IDO1, IFNG, PRF1, STAT1*) [17] and "T-cell Exhaustion" (*CD274/PD-L1, CD276, CD8A, LAG3, PDCD1LG2, TIGIT*) [17] signatures. For each gene signature, we performed hierarchical clustering of samples using Ward2 algorithm[19] (implemented as Ward.D2 in R stats package) and Euclidean distance. [19][20] We flattened the dendrogram into 3 clusters, namely Cluster High, Cluster Median and Cluster Low, and then merged the latter two clusters as Cluster Median/Low. Heatmaps with clustered dendrograms were used for visual validation.

mRNA extraction and gene expression analysis in the validation series

For each HCC sample, 5 µm-thick sections were cut from FFPE blocks. Tumor tissue was then macro-dissected and total RNAs were further isolated using the Recover All™ Total Nucleic Acid Isolation Kit (Invitrogen, Thermo Fisher Scientific), according to manufacturer's instructions. They were monitored to ensure their quality, purity and integrity using an Agilent 2100 Bioanalyzer device with the Pico assay (Agilent, Santa Clara, CA, USA). Samples with a DV200 (percentage of RNA fragments above 200 nucleotides) >30% were further considered adequate for analysis.

Gene expression analysis in the validation series

Gene expression was analyzed using the NanoString Immuno-Oncology 360 panel (NanoString Technologies, Seattle, USA) that includes a set of more than 700 genes involved in the main biological pathways of human immunity. These experiments were performed by the Genomics platform of Institut Curie (Paris, France). Total RNAs were used as templates. A human Universal Reference RNA including a no-template control (water) and 10 cell lines were also hybridized in parallel with the HCC samples. NanoString positive and negative controls were also added to samples as spikes in controls. After an overnight hybridization at 65°C, samples were processed on the NanoString nCounter preparation station (NanoString Technologies) to immobilize biotinylated hybrids and remove probes in excess. The nCounter Digital Analyzer was used to scan the cartridges at maximum resolution (fields of view n=555), count the fluorescent barcodes and quantify RNA molecules. Normalization was performed against the geometric mean of 20 housekeeping

genes in combination with a positive control normalization which uses the geometric mean of six synthetic positive targets.

As performed for the discovery series, gene expression values were log2-transformed and the Z-Score approach was further applied for gene-wise standardization. For each gene signature, hierarchical clustering with Ward2 algorithm and Euclidean distance was then performed. HCC samples were then labelled as Cluster High or Cluster Median/Low.

Slides preprocessing and tessellation, color normalization and data augmentation

Slides from the discovery set were stained with hematein-eosin and encoded in svs format while slides from the validation series were stained with hematein-eosin-saffron and encoded in ndpi format. Tissue regions were then exhaustively split into patches of 256×256 pixels at 20X using the OpenSlide library in Python. Tumor regions were annotated by an expert pathologist in polygonal Regions of Interest (ROIs) using the open source QuPath software.[21]. Patches were extracted from the intersection area of the tissue regions and the tumor regions. Staining conversion, color normalization and data augmentation protocols were assessed (**Supplemental Material and Methods**).

Deep-learning models

Baseline model: patch based workflow with ShuffleNet

A detailed description of all steps involved in the models development is provided in the **Supplemental Material and Methods**, and our code is publicly available (https://github.com/qinghezeng/Histo2GeneSignatures). For our baseline model, we re-implemented a patch-based strategy in Python (**Figure 2**) (https://github.com/jnkather/DeepHistology/tree/v0.2). For the training and cross validation, 500 patches were randomly collected from each slide, with each patch inheriting the label of the WSI. Cluster Median/Low training patches were downsampled to match the same number of Cluster High training patches. We then trained a ShuffleNet, which was pre-trained on ImageNet. With Cluster High as the positive class, optimal patch-level thresholds were computed using the receiver operating characteristic (ROC) curves on test patches. For inference, all patches extracted from a WSI were predicted by our trained ShuffleNet and classified as Cluster High or Low/Median using the optimal threshold determined on the discovery series. The slide level score was calculated by dividing the number of Cluster High patches by the total number of patches in that slide.

Slide based models: MIL and CLAM

Apart from the patch-based approach, we also investigated multiple-instance learning (MIL) approaches (**Figure 3**). MIL is a weakly-supervised learning paradigm in which data is arranged in bags of instances. In the binary MIL assumption, a bag is labelled as 1) negative if all the instances inside are negative or 2) positive if it contains one or more positive instances. Based on this assumption, if there are one or more Cluster High patches in the slide, the tumor slide will be classified as Cluster High. Provided with the WSI-level label (and tumor ROIs in the experiments with annotations), the MIL models have the ability to predict labels for unseen slides by taking account of the most predictive patches. We investigated two MIL approaches, namely classic MIL and Clustering-constrained Attention Multiple Instance Learning (CLAM).[22]

For both strategies, the first stage consists in feature extraction using a modified ResNet50 model pretrained on ImageNet (**Figure 3**). Each of the N_i patches from a given WSI was encoded as a 1024-dimensional feature, allowing the possibility to load all the patches of the WSI into the GPU memory simultaneously. For the classic MIL approach, we used the implementation in the CLAM work (code available at https://github.com/mahmoodlab/CLAM). A first fully-connected (FC) layer further reduced the features to 512 dimensions, and the second FC layer was used as a classifier to generate 2-class score for each patch. A max-pooling function was then applied on the Cluster High class to select the top-1 patch and normalize its scores to WSI-level probabilities by softmax.

CLAM is a recently reported more sophisticated MIL approach specifically designed for digital pathology (code available at https://github.com/mahmoodlab/CLAM). [22] Its attention mechanism helps the model to focus on representative patches automatically. The same ResNet50 network was used for the patch encoding, and, as performed for the classic MIL workflow, the feature vectors were reduced to 512 dimensions for each slide. A FC layer with softmax activation function was used as a classifier to generate 2-class WSI-level probabilities.

For all models investigated, training was performed using a 10-fold Monte Carlo cross-validation strategy. For each fold, the discovery series was randomly partitioned into training (60% of cases) / validation (20%) / test (20%) sets. The whole validation series was used for external validation. Performance was further assessed using ROC curves. For each gene signature, we determined the optimal thresholds by selecting the cutoff with the highest Youden index. In case of multiple optimal thresholds, the one closer to the median was selected.

Attention maps generation and pathological reviewing

CLAM is able to produce interpretable heat maps that allow users to visualize, within each WSI, the relative contribution of every tissue area to the model's predictions.[22,23] These heat maps thus allow pathologists to determine which histological and cytological features are associated with a high predictive value.

For each gene signature, an attention score was learned by CLAM for each patch and converted into percentiles. For each WSI, the percentiles were then normalized to [0, 1] with 1 being the most predictive and 0 being the most non-informative. The normalized scores were represented with a colormap (red for 1 and blue for 0), and reconstructed into a heatmap according to the spatial location of the corresponding patches. We extracted, for each tumor correctly classified as Cluster High and each gene signature, the top 8 image patches classified as highly predictive and the top 8 patches considered non-informative by the model. They were further reviewed by two pathologists (JC and CTN). The following histological and cytological features were systematically recorded: tumor cells, blood, vessels, immune cells (lymphocytes, neutrophils, eosinophils, plasma cells), steatosis, clear cells, atypia, fibrosis, eosinophilic inclusions, cholestasis, hyperchromasia, sarcomatoid changes, multinucleated cells, necrosis, steatohepatitic pattern, vascular spaces, microtrabecular, macrotrabecular, compact and pseudoglandular growth patterns (**Supplemental Table 1**).

Differences between highly predictive and non-informative image tiles were analyzed using Fisher's exact test. The degree of inter-observer agreement was assessed with Cohen Kappa statistics.

Results

Unsupervised hierarchical clustering of samples from the discovery series

A total of 336 cases from the TCGA dataset met our inclusion criteria (349 slides from 336 tumors). The main clinical, biological and pathological features of the patients and tumors are presented in **Supplemental Table 2**. They were common for a series of patients with HCC treated by surgical resection. Median age at surgery was 61 years and a male predominance was observed (sex ratio 68%, 228/336). The main risk factors were alcohol consumption (35%, 111/318) and hepatitis B (32%, 101/318). Microvascular invasion was identified in 29% of the tumors (83/285). As expected in a series of patients treated by surgery, a low rate of significantly fibrotic livers was observed (Ishak score 5-6, 37%, 74/198).

We first performed unsupervised clustering analysis on RNA sequencing data for all 6 gene signatures. Three distinct sample clusters (High, Median and Low) were observed for each gene signature (**Figure 4**). We further aimed to identify samples belonging to the cluster displaying the highest expression of the gene signatures investigated, as they are likely to constitute the subset of HCC that are the most sensitive to immunotherapy.[9] Tumors belonging to Cluster "High" for 6G Interferon Gamma, Gajewski 13-Gene Inflammatory, Inflammatory, Interferon Gamma Biology, Ribas 10G Interferon Gamma, and T-cell Exhaustion signatures represented 13% (44/336), 14% (48/336), 12% (41/336), 11% (36/336), 12% (40/336) and 11% (36/336) of the cases, respectively (**Supplemental Table 3**). A significant overlap in samples classified as Cluster High was observed (25 cases belonged to Cluster High for all signatures investigated). We investigated the associations between clusters and clinical and pathological features of the patients and tumors (**Supplemental Table 4**). We observed significant associations ($p<0.05$) between Cluster High and: higher AFP serum levels (all 6 gene signatures), Hepatitis C Virus infection (all signatures except Interferon Gamma Biology), non-tumor liver fibrosis (Ishak score 5-6) (Ribas 10G) and G4 histological grade (Inflammatory).

Development of deep-learning models for the prediction of HCC with activation of immune gene signatures

The 349 WSIs from the TCGA series were downloaded and we first extracted ~1,000,000 and 4,982,872 tiles (256 x 256 pixels) for the patch-based and classic MIL/CLAM models, respectively. The WSIs were processed as is, meaning that all patches from both the tumor

and the adjacent non-tumor parenchyma were analyzed. These image tiles were further fed, along with their corresponding immune cluster labels, into our 3 different models.

Training was performed using a 10-fold Monte Carlo strategy and, for each model and gene signature, each fold AUC was computed. We observed that the 3 different models showed a relatively weak overall performance (**Supplemental Table 5**). The mean AUC ranged from 0.490 to 0.666, 0.516 to 0.577, and from 0.555 to 0.632 for the patch based approach, MIL and CLAM, respectively (**Supplemental Table 5**). The highest performance was achieved for the Gajewski 13G signature (mean AUC 0.577 and 0.632 for MIL and CLAM, respectively). We hypothesized that these suboptimal results were explained, at least in part, by the existence of irrelevant noise/patterns in the non-tumor tissue included in the WSIs.

We thus modified our strategy, and an expert pathologist (JC) annotated the tumor areas for each of the 349 TCGA WSIs available. The models were re-trained on patches extracted from the manually delineated tumor areas (total of ~600,000 and 2,926,135 patches for the patch-based and classic MIL/CLAM, respectively). This approach yielded better performances with best-fold AUCs ranging from 0.661 to 0.783, 0.677 to 0.893 and 0.780 to 0.914 for the patch-based approach, the classic MIL and the CLAM models, respectively (**Table 1**). AUCs were higher for Gajewski 13G and IFNg biology signatures. Among our 3 different approaches, the CLAM model showed an overall superior performance (**Table 1**). The results (ROC curves and confusion matrix for the optimal threshold) for the best fold models obtained with CLAM on the 6 gene signatures are displayed on **Figure 4**.

External validation of the models

Deep learning systems are prone to overfit to the data they are trained on and external validation is critical. We thus aimed to validate our models in a completely independent series of samples with different 1) gene expression profiling technology (NanoString Panel vs RNA sequencing), 2) staining protocols (hematein-eosin-saffron vs hematein-eosin), and 3) WSI image format (.ndpi vs .svs). The fields of view were also slightly different (~128x128 µm², ~0.5 µm/pixel vs ~115x115 µm², ~0.45 µm/pixel). A total of 139 patients treated by surgical resection in Henri Mondor University Hospital were included (**Figure 1**). The most frequent risk factors for liver disease were alcohol intake (31%, 43/138), HBV (hepatitis C virus) (24%, 34/138) and HCV (hepatitis C virus) (24%, 33/138) infection (**Supplemental Table 6**). Disease stage, according to the Barcelona Clinic of Liver Cancer (BCLC) system,

was 0/A in 78% of the patients and B/C in 22% of the patients. As observed in the discovery series, the frequency of cirrhosis was also low (35%, 34/97).

After mRNA extraction and gene expression analysis using the NanoString Pan Cancer Immuno-Oncology 360 panel, we performed unsupervised hierarchical clustering on all samples. Tumors belonging to Cluster “High” for 6G IFNg, Gajewski 13G, Inflammatory, IFNg Biology, Ribas 10G, and T-cell Exhaustion signatures represented 14% (20/139), 30% (42/139), 9% (13/139), 16% (22/139), 12% (17/139) and 6% (8/139) of the cases, respectively. (**Supplemental Table 7**). As observed in the discovery series, an overlap among cases classified as Cluster High was identified (8 cases belonged to Cluster High for all signatures investigated). Cases classified as Cluster High were associated with the following clinical, biological and pathological features : poor differentiation (except Gajewski 13G), HCV infection (Gajewski 13G p=0.03), higher age at surgery (6G IFNg p=0.006, Ribas 10G p=0.03), higher AFP serum levels (Inflammatory p=0.02, T cell exhaustion p=0.002) and lower tumor size (Gajewski 13G p=0.002) (**Supplemental Table 8**).

As performed for the discovery series, all tumor areas from the 139 WSIs were annotated by a pathologist. A total of 1555984 tiles were thus extracted and fed to the trained CLAM models developed in the TCGA dataset. Samples were classified as Cluster High or Median/low using the optimal thresholds identified on the TCGA test splits.

We were able to validate their performance with AUCs of 0.817, 0.810, 0.850, 0.823, 0.810 and 0.921 for 6G IFNg, Gajewski 13G, Inflammatory, IFNg biology, Ribas 10G and T cell exhaustion gene signatures, respectively. **(Table 2)** The ROC curves and confusion matrices are displayed on **Figure 5**. We also investigated 3 different techniques that may increase the overall performance of our models: 1) conversion of our hematein-eosin-saffron slides to hematein-eosin (staining used in the discovery series), 2) color normalization and 3) data augmentation techniques (during training) (**Supplemental material and methods**). We did not observe any significant improvement (**Supplemental Table 9-11**).

Finally, we were able to analyze pre-operative biopsies from 7 cases. Slides were processed using our best-fold CLAM models and we observed that tumor immune cluster were accurately predicted in 38 out of the 42 cases (7 cases X 6 gene signatures) (**Supplemental Figure 1**).

Generation of heatmaps and pathological review of areas with high predictive value

In order to have a better understanding of the morphological and biological features involved in the classification process, we extracted, for each gene signature, the top 8 most predictive and the top 8 non informative patches from all tumors accurately classified as Cluster High (examples in **Figure 6**). A total of 1264 image tiles were thus reviewed and 23 histological and cytological features were recorded for each tile. Inter-observer agreement between the two pathologists was substantial with Cohen Kappa values ranging from 0.69 to 0.88. We observed highly significant differences between highly predictive and non-informative areas for the following signatures:

-6G IFNg: presence of lymphocytes ($p<0.0001$), neutrophils ($p<0.0001$) and plasma cells ($p=0.0007$) and lack of blood ($p=<0.0001$), tumor cells ($p<0.0001$), steatosis ($p=0.03$), fibrosis ($p=0.007$), and macrotrabecular ($p=0.003$) and compact ($p=0.03$) tumor growth patterns.

-IFNg: presence of lymphocytes ($p<0.0001$), neutrophils ($p<0.0001$) and plasma cells ($p=0.007$) and lack of tumor cells ($p<0.0001$), steatosis ($p<0.0001$), fibrosis ($p<0.0001$) and compact growth pattern ($p<0.0001$).

-Ribas 10G: presence of lymphocytes ($p<0.0001$), neutrophils ($p=0.003$) and plasma cells ($p=0.001$) and lack of tumor cells ($p<0.0001$), steatosis ($p=0.006$), fibrosis ($p<0.0001$) and compact growth pattern ($p<0.0001$) and atypia ($p=0.01$).

No significant associations were observed for the remaining 3 gene signatures: Gajewski 13G, Inflammatory and T cell exhaustion (**Supplemental Table 12**).

We finally compared the frequency of the 23 pathological features on the most predictive patches between the 6 signatures investigated and observed several significant differences (**Supplemental Table 13**). Altogether, these findings shows that the morphological characteristics captured by the models are, at least in part, different.

Discussion

Numerous gene signatures or transcriptomic subgroups have been proposed to better select patients that may benefit from particular targeted therapies.[11] The use of such biomarkers however require molecular biology and bioinformatics expertise, and very few are used in clinical practice.

We show, in this proof-of-concept study, that deep learning applied to digital histological slides has the ability to predict several immune gene signatures related to response to immunotherapy.[9] Although the relevance of these signatures remain to be prospectively validated, our results demonstrate that AI-based pathology is a promising approach to extract clinically and biologically significant information from WSIs of HCC.[15,24–28]

We investigated 3 different deep-learning approaches, and showed that the CLAM network yielded the best overall performance. In addition to the patch based approach and the classic MIL, CLAM uses an attention-based learning process to automatically identify particular areas of high diagnostic value within the WSI.[22,23] It involves an adaptive weighting for each image area, and the model can thus make more granular and flexible predictions. This feature allows to automatically lower the impact of irrelevant tissue patches and may contribute, at least in part, to the higher performance of this method. Although this type of model is designed to reduce the need for manual annotations, we show that tumor areas delineation by a pathologist results in a significant improvement for all gene signatures investigated. We may hypothesize that the negative impact of irrelevant noise/patterns from the adjacent non-tumor parenchyma was avoided by the use of expert driven annotations. These findings underscore the importance of human-machine interactions for the development of highly efficient AI-based pathology.

The processing of surgical and biopsy samples is complex with several critical steps (embedding, cutting, staining, and scanning) that may impact the overall quality of the slide. Models can easily overfit and thus learn noise or fluctuations that are specific to the training dataset, explaining that they most often do not generalize well on external datasets. Technical protocols for gene expression analyses also include multiple steps that are impacted by experimental conditions (RNA extraction, reverse transcription, and amplification, for example) and may introduce nonlinear effects. The identification of robust tumor subgroups across different technological platforms thus remains challenging. These issues explain why, although numerous studies using deep-learning on digital slides have been recently published, the majority of the proposed models were not validated on true independent series of samples processed by different histological and molecular protocols.

We thus believe that one of the strengths of our work is the validation to which the various models investigated have been subjected. We were indeed able to validate our different models on a completely independent data set that included 1) patients treated in a different center, 2) slides stained with different protocols and encoded in a different format and 3) gene expression experiments performed using a different technology. The performance of our classifiers, as assessed by the AUC, was also in most cases >0.80, which is usually considered “excellent”.[29]

By changing the classification thresholds, sensitivity and specificity of our AI-based pathology assays can also be modulated according to particular clinical settings/needs. Deep-learning models with thresholds allowing high sensitivity can indeed be used to pre-screen patients and further trigger confirmatory gene expression profiling in case of a positive prediction. Even with suboptimal specificity, such assays may have the ability to speed up diagnostic workflows.

The next step for the implementation of such models will be their broad validation on HCC biopsies (the only type of samples available for patients with advanced disease). Our preliminary results are encouraging, however the investigation of larger series will be mandatory to determine if such samples are suitable for this type of AI-driven analyses. It may be challenging as they are rarely performed due to the existence of diagnostic non-invasive criteria. There is however a renewed interest in biopsy, in particular in the context of clinical trials, and several studies have shown that their analysis can provide meaningful molecular and prognostic information.[30] We may thus be able to perform a large scale validation of our models on this type of samples within the foreseeable future. Trials investigating in neodjuvant immunotherapy may also encourage us to reconsider the practice of biopsy even in the context of small, resectable tumors. Several drugs, including immunomodulatory molecules, are currently being tested in the adjuvant setting (after surgical resection) and our models may help to identify the patients most likely to benefit from these approaches.[31–34]

In order for AI-based algorithms to be used in daily practice, they must be prospectively validated in ‘real life” clinical workflows to show that they are able to provide useful information in a timely manner. Indeed, model development is the first step of an overall process that includes several technical (image acquisition/storage, reproducibility and robustness), regulatory (quality control framework, data privacy), and clinical (demonstration of improved clinical outcomes) barriers. Improvements in the error/misclassification rates are also critical for the adoption of model such as ours.

Convolutional neural networks consist of multiple layers of complex mathematical computation, and deep-learning models are thus often considered as black boxes. We took advantage of a particular feature of the CLAM method to generate attention maps and analyze the areas with a high predictive value. It does not provide true explainability but highlights tiles that may contain relevant information. For 3 signatures related to interferon gamma signaling, we were able to show that classification as Cluster High relied on tiles containing lymphocytes, neutrophils and plasma cells. These observations are consistent with the function of the genes included, as molecules encoded by genes such as *IFNG, CXCL10* or *CXCL11* are known to promote the recruitment of various immune cells subsets. We did not identify any histological features associated with the 3 remaining signatures, suggesting that the models are also able to capture morphological characteristics that are not easily accessible to the human eye. We believe that the use of explainable models is key to provide the required trust of pathologists and physicians, and thus ease their implementation in clinical practice. The analysis of tissue areas associated with particular predictions may also lead to new scientific discoveries.

In conclusion, we have developed and validated deep-learning based models able to predict the activation of several immune gene signatures that may be associated with improved response to immunotherapy in patients with HCC. These approaches could represent a novel type of biomarker that will ease the translation of our biological knowledge of HCC into clinical practice.

Financial support: Fondation Bristol Myers Squibb pour la Recherche en Immuno-Oncologie, Fondation de l'Avenir, CARPEM, and China Scholarship Council.

Data availability statement : All our code is avaiable at the following website : https://github.com/qinghezeng/Histo2GeneSignatures

References

[1] Llovet JM, Kelley RK, Villanueva A, Singal AG, Pikarsky E, Roayaie S, et al. Hepatocellular carcinoma. Nat Rev Dis Prim 2021;7:6. https://doi.org/10.1038/s41572-020-00240-3.

[2] Llovet JM, Montal R, Sia D, Finn RS. Molecular therapies and precision medicine for hepatocellular carcinoma. Nat Rev Clin Oncol 2018;15:599–616. https://doi.org/10.1038/s41571-018-0073-4.

[3] Kudo M, Finn RS, Qin S, Han KH, Ikeda K, Piscaglia F, et al. Lenvatinib versus sorafenib in first-line treatment of patients with unresectable hepatocellular carcinoma: a randomised phase 3 non-inferiority trial. Lancet 2018;391:1163–73. https://doi.org/10.1016/S0140-6736(18)30207-1.

[4] Abou-Alfa GK, Meyer T, Cheng A-L, El-Khoueiry AB, Rimassa L, Ryoo B-Y, et al. Cabozantinib in Patients with Advanced and Progressing Hepatocellular Carcinoma. N Engl J Med 2018;379:54–63. https://doi.org/10.1056/NEJMoa1717002.

[5] Zhu AX, Kang YK, Yen CJ, Finn RS, Galle PR, Llovet JM, et al. Ramucirumab after sorafenib in patients with advanced hepatocellular carcinoma and increased α-fetoprotein concentrations (REACH-2): a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet Oncol 2019;20:282–96. https://doi.org/10.1016/S1470-2045(18)30937-9.

[6] Bruix J, Qin S, Merle P, Granito A, Huang Y-H, Bodoky G, et al. Regorafenib for patients with hepatocellular carcinoma who progressed on sorafenib treatment (RESORCE): a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet 2017;389:56–66. https://doi.org/10.1016/S0140-6736(16)32453-9.

[7] Finn RS, Qin S, Ikeda M, Galle PR, Ducreux M, Kim T-Y, et al. Atezolizumab plus Bevacizumab in Unresectable Hepatocellular Carcinoma. N Engl J Med 2020;382:1894–905. https://doi.org/10.1056/NEJMoa1915745.

[8] Sangro B, Sarobe P, Hervás-Stubbs S, Melero I. Advances in immunotherapy for hepatocellular carcinoma. Nat Rev Gastroenterol Hepatol 2021:1–19. https://doi.org/10.1038/s41575-021-00438-0.

[9] Sangro B, Melero I, Wadhawan S, Finn RS, Abou-Alfa GK, Cheng A-L, et al. Association of inflammatory biomarkers with clinical outcomes in nivolumab-treated patients with advanced hepatocellular carcinoma. J Hepatol 2020;73:1460–9. https://doi.org/10.1016/j.jhep.2020.07.026.

[10] Haber PK, Torres-Martin M, Dufour J-F, Verslype C, Marquardt J, Galle PR, et al. Molecular markers of response to anti-PD1 therapy in advanced hepatocellular carcinoma. J Clin Oncol 2021;39:4100–4100. https://doi.org/10.1200/JCO.2021.39.15_suppl.4100.

[11] Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat Rev Cancer 2019;19:133–50. https://doi.org/10.1038/s41568-019-0116-x.

[12] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559–67. https://doi.org/10.1038/s41591-018-0177-5.

[13] Echle A, Grabsch HI, Quirke P, van den Brandt PA, West NP, Hutchins GGA, et al. Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. Gastroenterology 2020;159:1406-1416.e11. https://doi.org/10.1053/j.gastro.2020.06.021.

[14] Kather JN, Calderaro J. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. Nat Rev Gastroenterol Hepatol 2020;17:591–2. https://doi.org/10.1038/s41575-020-0343-3.

[15] Calderaro J, Kather JN. Artificial intelligence-based pathology for gastrointestinal and hepatobiliary cancers. Gut 2021;70:1183–93. https://doi.org/10.1136/gutjnl-2020-322880.

[16] Cancer Genome Atlas Research Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell 2017;169:1327-1341.e23. https://doi.org/10.1016/j.cell.2017.05.046.

[17] Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, et al. IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. J Clin Invest 2017;127:2930–40. https://doi.org/10.1172/JCI91190.

[18] Spranger S, Bao R, Gajewski TF. Melanoma-intrinsic β-catenin signalling prevents anti-tumour immunity. Nature 2015;523:231–5. https://doi.org/10.1038/nature14404.

[19] Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif 2014;31:274–95. https://doi.org/10.1007/s00357-014-9161-z.

[20] Deza MM, Deza E. Encyclopedia of distances. Berlin, Heidelberg: Springer; 2009.

[21] Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: Open source software for digital pathology image analysis. Sci Rep 2017;7:16878. https://doi.org/10.1038/s41598-017-17204-5.

[22] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021;5:555–70. https://doi.org/10.1038/s41551-020-00682-w.

[23] Lu MY, Chen TY, Williamson DFK, Zhao M, Shady M, Lipkova J, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature 2021;594:106–10. https://doi.org/10.1038/s41586-021-03512-4.

[24] Saillard C, Schmauch B, Laifa O, Moarii M, Toldo S, Zaslavskiy M, et al. Predicting survival after hepatocellular carcinoma resection using deep‐learning on histological slides. Hepatology 2020:hep.31207. https://doi.org/10.1002/hep.31207.

[25] Calderaro J, Ziol M, Paradis V, Zucman-Rossi J. Molecular and histological correlations in liver cancer. J Hepatol 2019;71:616–30. https://doi.org/10.1016/j.jhep.2019.06.001.

[26] Ziol M, Poté N, Amaddeo G, Laurent A, Nault J-C, Oberti F, et al. Macrotrabecular-massive hepatocellular carcinoma: A distinctive histological subtype with clinical relevance. Hepatology 2018;68:103–12. https://doi.org/10.1002/hep.29762.

[27] Calderaro J, Couchy G, Imbeaud S, Amaddeo G, Letouzé E, Blanc J-F, et al. Histological subtypes of hepatocellular carcinoma are related to gene mutations and molecular tumour classification. J Hepatol 2017;67:727–38. https://doi.org/10.1016/j.jhep.2017.05.014.

[28] Calderaro J, Rousseau B, Amaddeo G, Mercey M, Charpy C, Costentin C, et al. Programmed death ligand 1 expression in hepatocellular carcinoma: Relationship With clinical and pathological features. Hepatology 2016;64:2038–46. https://doi.org/10.1002/hep.28710.

[29] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Assessing the Fit of the Model. Appl. Logist. Regression., vol. 398. 2nd ed., New York: John Wiley & Sons; 2013, p. 160–4. https://doi.org/10.2307/2532419.

[30] Nault J, Martin Y, Caruso S, Hirsch TZ, Bayard Q, Calderaro J, et al. Clinical Impact of Genomic Diversity From Early to Advanced Hepatocellular Carcinoma. Hepatology 2020;71:164–82. https://doi.org/10.1002/hep.30811.

[31] Haber PK, Puigvehí M, Castet F, Lourdusamy V, Montal R, Tabrizian P, et al. Evidence-based management of HCC: Systematic review and meta-analysis of randomized controlled trials (2002-2020). Gastroenterology 2021. https://doi.org/10.1053/j.gastro.2021.06.008.

[32] Hack SP, Spahn J, Chen M, Cheng A-L, Kaseb A, Kudo M, et al. IMbrave 050: a Phase III trial of atezolizumab plus bevacizumab in high-risk hepatocellular carcinoma after curative resection or ablation. Futur Oncol 2020;16:975–89. https://doi.org/10.2217/fon-2020-0162.

[33] Pan Q-Z, Liu Q, Zhou Y-Q, Zhao J-J, Wang Q-J, Li Y-Q, et al. CIK cell cytotoxicity is a predictive biomarker for CIK cell immunotherapy in postoperative patients with hepatocellular carcinoma. Cancer Immunol Immunother 2020;69:825–34. https://doi.org/10.1007/s00262-020-02486-y.

[34] Su Y-Y, Li C-C, Lin Y-J, Hsu C. Adjuvant versus Neoadjuvant Immunotherapy for Hepatocellular Carcinoma: Clinical and Immunologic Perspectives. Semin Liver Dis 2021. https://doi.org/10.1055/s-0041-1730949.

Tables

Table 1. Performances of the models in the discovery series, using annotated WSIs (best-fold and mean AUCs).
AUC: Area Under the ROC Curve, sd: standard deviation.

	Patch-based approach		***Classic MIL***		***CLAM***	
Gene signature	***Best fold***	***Mean ± sd***	***Best fold***	***Mean ± sd***	***Best fold***	***Mean ± sd***
6G Interferon Gamma	0.661	0.560 ± 0.067	0.758	0.630 ± 0.078	0.780	0.635 ± 0.097
Gajewski 13G Inflammatory	0.809	0.688 ± 0.062	0.893	0.694 ± 0.125	0.914	0.728 ± 0.096
Inflammatory	0.706	0.580 ± 0.077	0.806	0.641 ± 0.123	0.796	0.665 ± 0.081
Interferon Gamma biology	0.783	0.561 ± 0.119	0.677	0.610 ± 0.051	0.822	0.674 ± 0.102
Ribas 10G Inflammatory	0.727	0.640 ± 0.074	0.726	0.618 ± 0.065	0.806	0.669 ± 0.067
T cell exhaustion	0.661	0.543 ± 0.073	0.788	0.606 ± 0.086	0.788	0.577 ± 0.092

Table 2. Performances (Area Under the ROC Curve) of the best-fold models in the validation series, using annotated WSI.

Gene signature	***CLAM AUC***
6G Interferon Gamma	*0.817*
Gajewski 13G Inflammatory	*0.810*
Inflammatory	*0.850*
Interferon Gamma biology	*0.823*
Ribas 10G Inflammatory	*0.810*
T cell exhaustion	*0.921*

Figure Legends

Figure 1. Flow-chart of the study. Expression of genes included in the 6 signatures was investigated in 336 HCC samples from the discovery series (TCGA) using RNA sequencing data and unsupervised clustering. Cases were labelled as Cluster High or Median/Low and the different approaches were trained using the available 349 WSIs and the associated immune labels. The best models were further validated in 139 tumors developed in patients treated in Henri Mondor University Hospital. For these samples, the gene expression was investigated using the NanoString Pan Cancer IO360 panel.

Figure 2. Workflow for patch-based strategy. During the training, 500 patches are randomly sampled from each WSI. The Cluster Median/Low training set is then randomly downsampled to match the amount of Cluster High patches. The equalized training sets are used to finetune a ShuffleNet pre-trained on ImageNet, with an optimal patch-level threshold calculated from the receiver operating characteristic (ROC) curve. For the inference, the class of all the patches extracted from the annotated tissue regions are predicted by the trained ShuffleNet and further binarized using the previous threshold. The percentage of Cluster High patches is the probability of this WSI to be Cluster High.

Figure 3. Workflow for multiple instance learning strategy. CLAM and classic MIL share the same preprocessing steps of tessellation and feature extraction. A 1024-dimensional feature embedding is extracted from each patch by a ResNet50 trained on ImageNet. The output of the deep learning network is a WSI-level probability that is thereafter binarized into Cluster High or Cluster Median/Low. N_i: patch number in a WSI. FC: fully connected layer. ⊗: attention-based pooling, composed of element-wise multiplication and sum along the first axis.

Figure 4. Development of deep-learning models in the discovery series dataset (TCGA). For the 6 each gene signatures, Clustering heatmaps, receiver operating characteristic (ROC) curves and confusion matrices on the test set using the best-fold model are provided (optimal threshold: 0.149, 0.288, 0.169, 0.277, 0.152, 0.187, respectively).

Figure 5. Validation of best-fold deep-learning models in the validation series (Mondor). For the 6 gene signatures, Clustering heatmaps, receiver operating characteristic

(ROC) curves and confusion matrices are provided (best-fold models using the optimal thresholds determined on the TCGA series).

Figure 6. Pathological reviewing of highly predictive tiles. Microscopic examination of highly predictive tiles showed enrichment of particular immune-related features. Examples for 6G Interferon Gamma and Interferon Gamma Biology signatures are provided in panels A and B, respectively. Tiles associated with 6G Interferon Gamma included lymphocytes (yellow arrows) and plasma cells (white arrows) We also identified enrichment in neutrophils on tiles associated with Interferon Gamma Biology (red arrows).

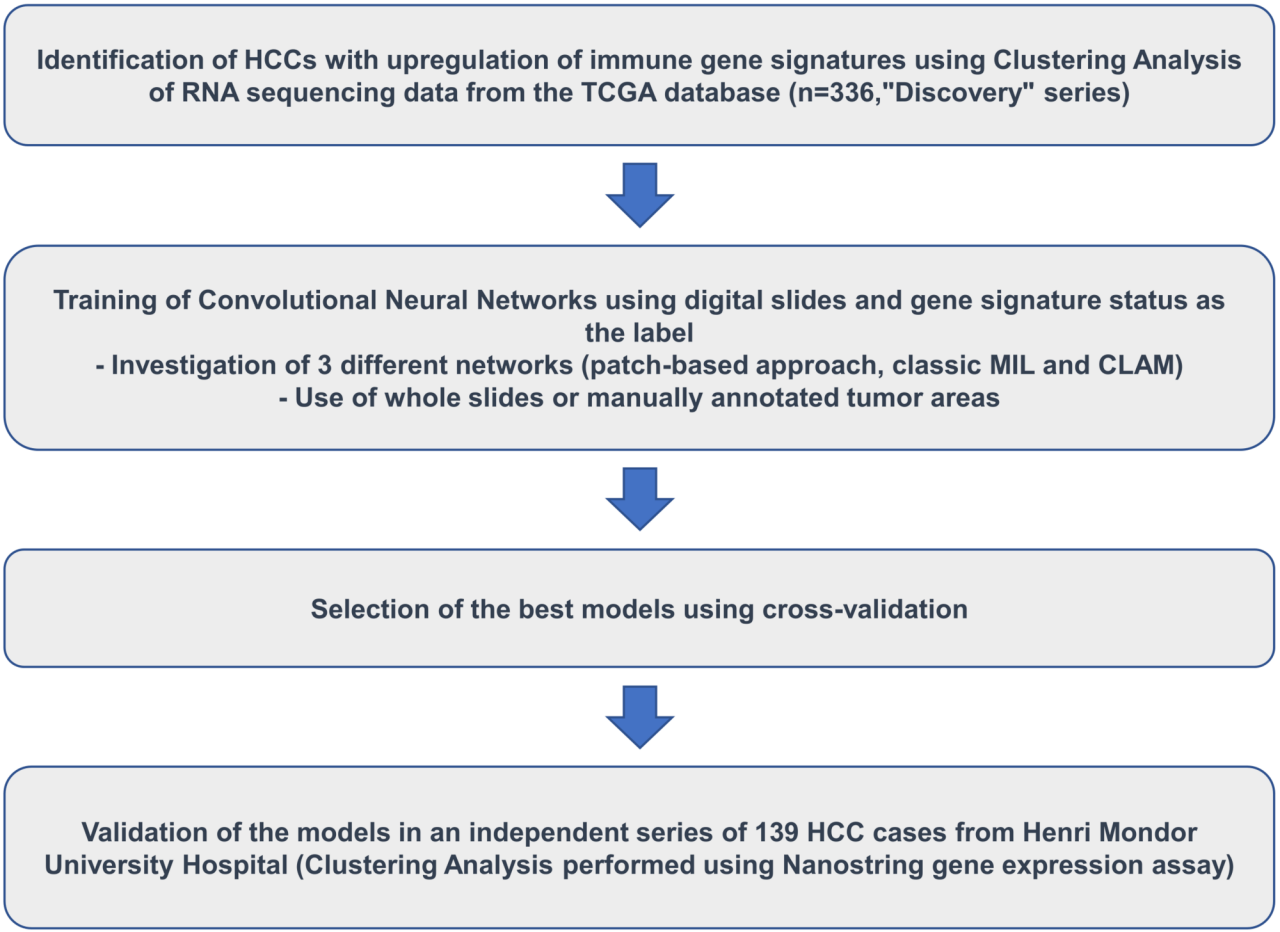


Figure 1. Flow-chart of the study. Expression of genes included in the 6 signatures was investigated in 336 HCC samples from the discovery series (TCGA) using RNA sequencing data and unsupervised clustering. Cases were labelled as Cluster High or Medial/Low and the different approaches were trained using the available 349 WSIs and the associated immune labels. The best models were further validated in 139 tumors developed in patients treated in Henri Mondor University Hospital. For these samples, the gene expression was investigated using the Nanostring Pan Cancer IO360 panel.

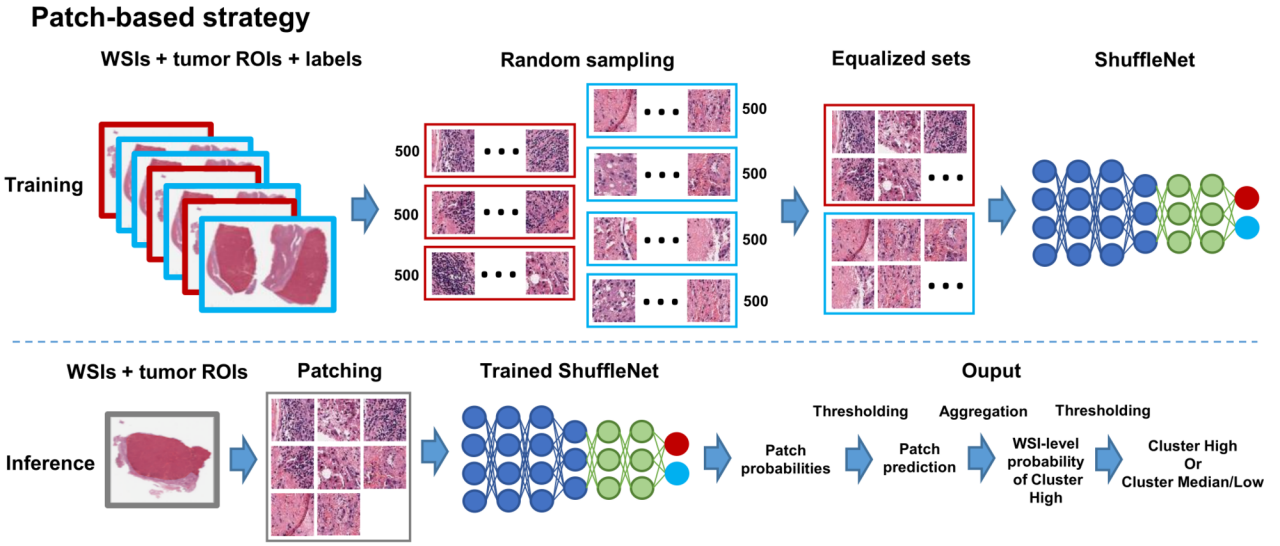


Figure 2. Workflow for patch-based strategy. During training, 500 patches are randomly sampled from each WSI. The Cluster Median/Low training set is then randomly downsampled to match the amount of Cluster High patches. The equalized training sets are used to finetune a ShuffleNet pre-trained on ImageNet, with an optimal patch-level threshold calculated from the receiver operating characteristic (ROC) curve. For inference, all patches extracted from the annotated tissue regions are predicted by the trained ShuffleNet and further binarized with the threshold. The percentage of Cluster High patches is the probability of this WSI being Cluster High.

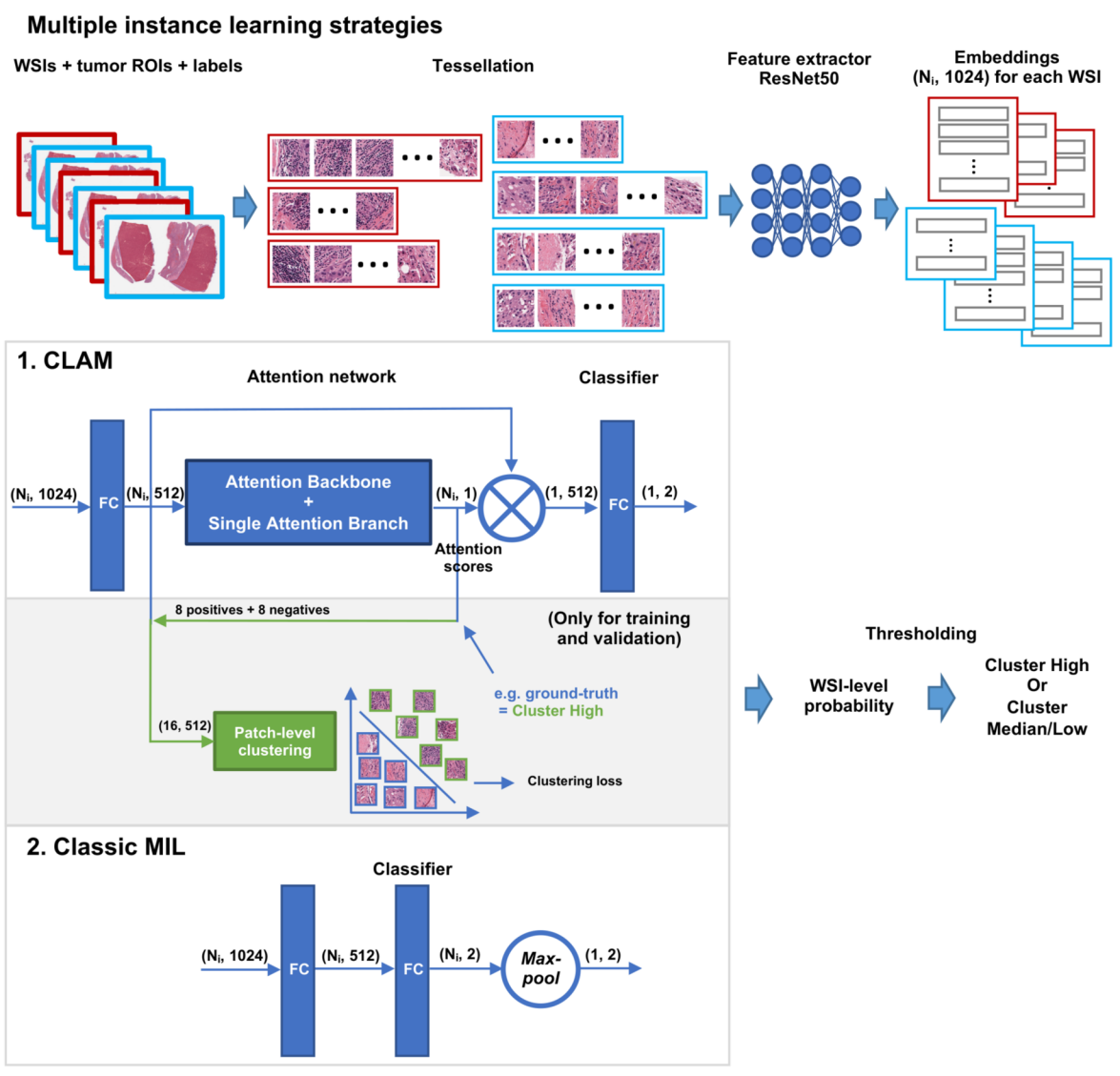


Figure 3. Workflow for multiple instance learning strategy. CLAM and classic MIL share the same preprocessing steps of tessellation and feature extraction. A 1024-dimensional feature embedding is extracted from each patch by a ResNet50 trained on ImageNet. The output of the deep learning network is a WSI-level probability that was further thresholded into Cluster High or Cluster Median/Low. N_i: patch number in a WSI. FC: fully connected layer. ⊗: attention-based pooling, composed of element-wise multiplication and sum along the first axis.

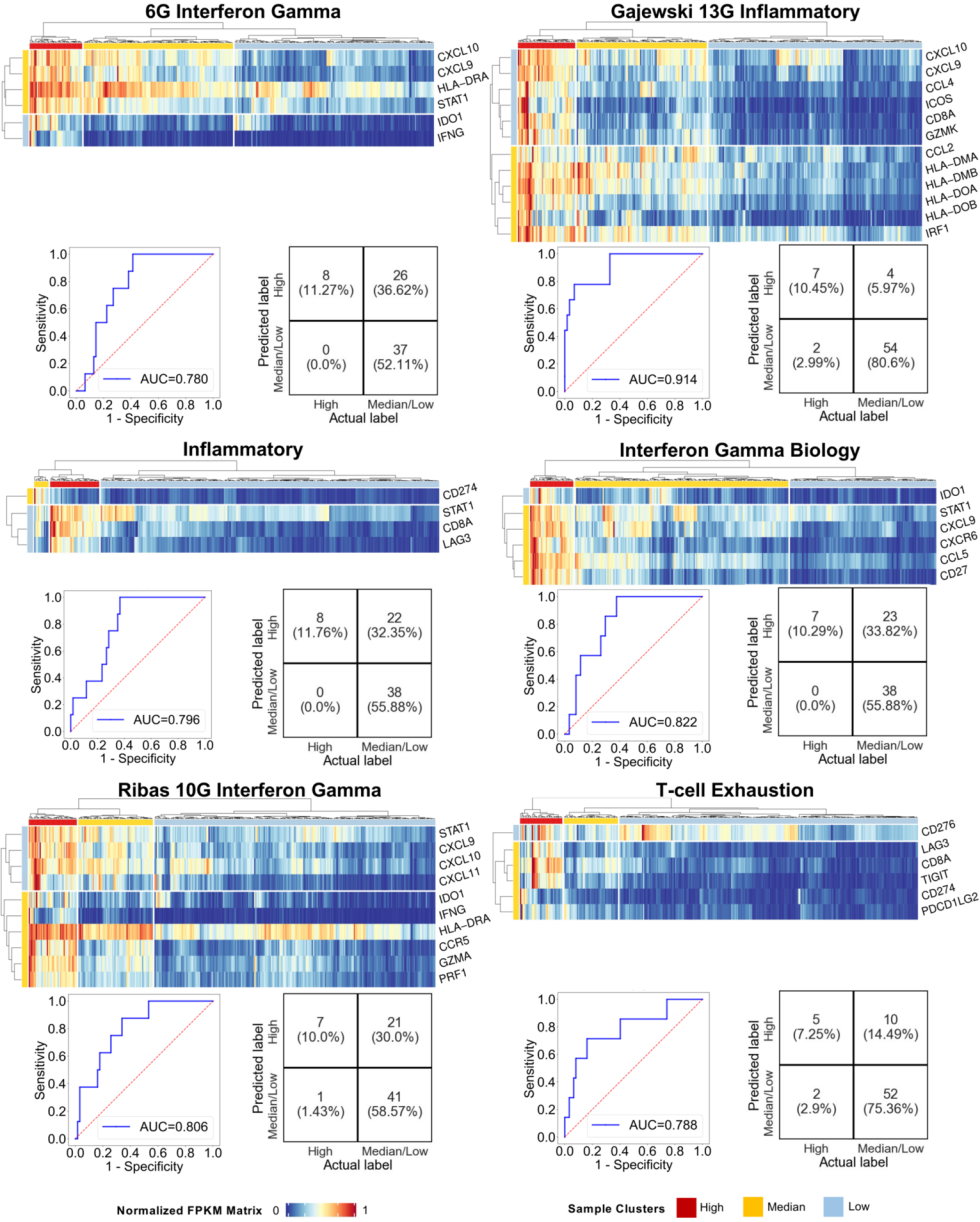


Figure 4. Development of deep-learning models in the discovery series dataset (TCGA). For the 6 each gene signatures, Clustering heatmaps, receiver operating characteristic (ROC) curves and confusion matrices on the test set using the best-fold model are provided (optimal threshold: 0.149, 0.288, 0.169, 0.277, 0.152, 0.187, respectively).

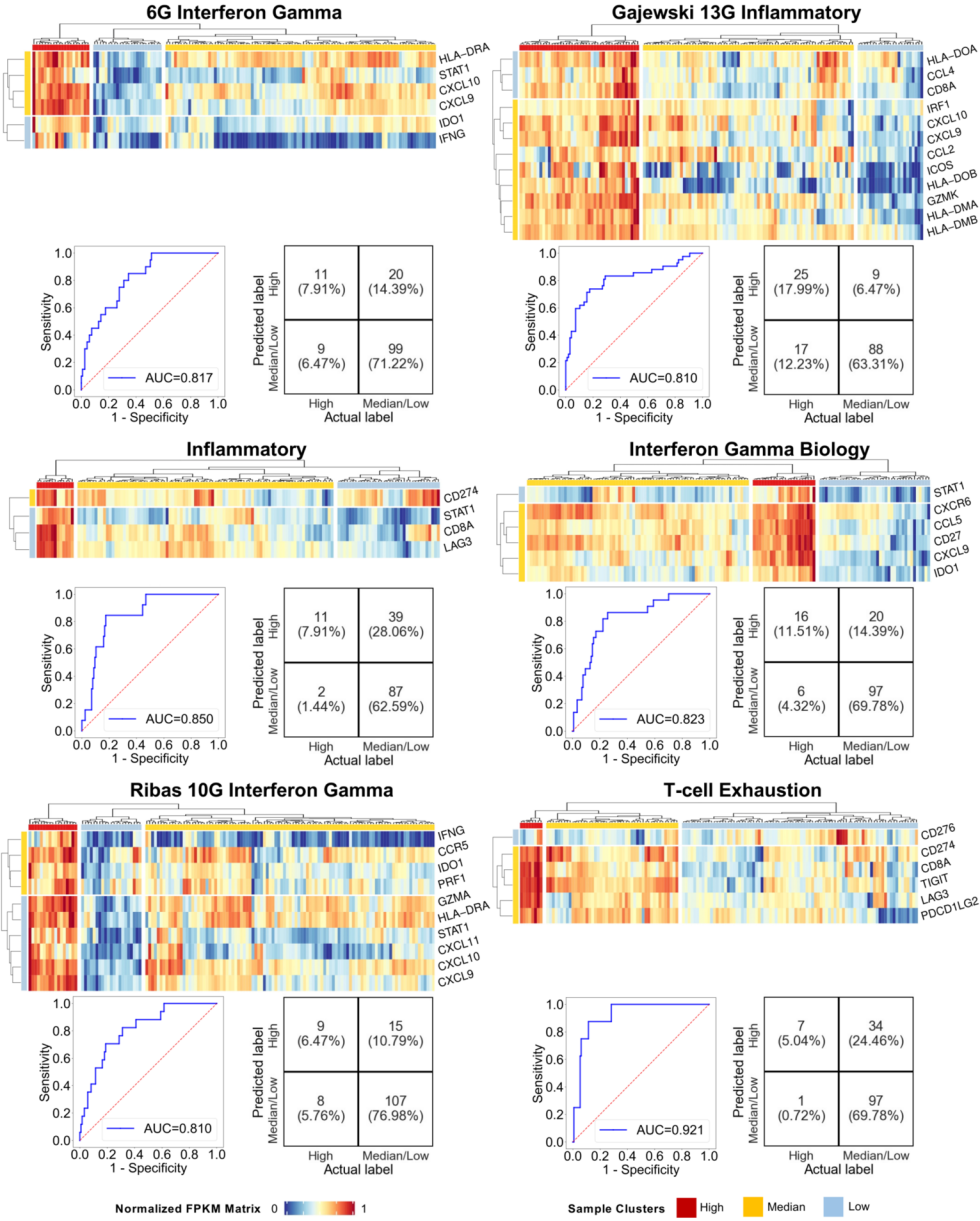


Figure 5. Validation of best-fold deep-learning models in the validation series (Mondor). For the 6 gene signatures, Clustering heatmaps, receiver operating characteristic (ROC) curves and confusion matrices are provided (best-fold models using the optimal thresholds determined on the TCGA series: 0.149, 0.288, 0.169, 0.277, 0.152, 0.187, respectively).

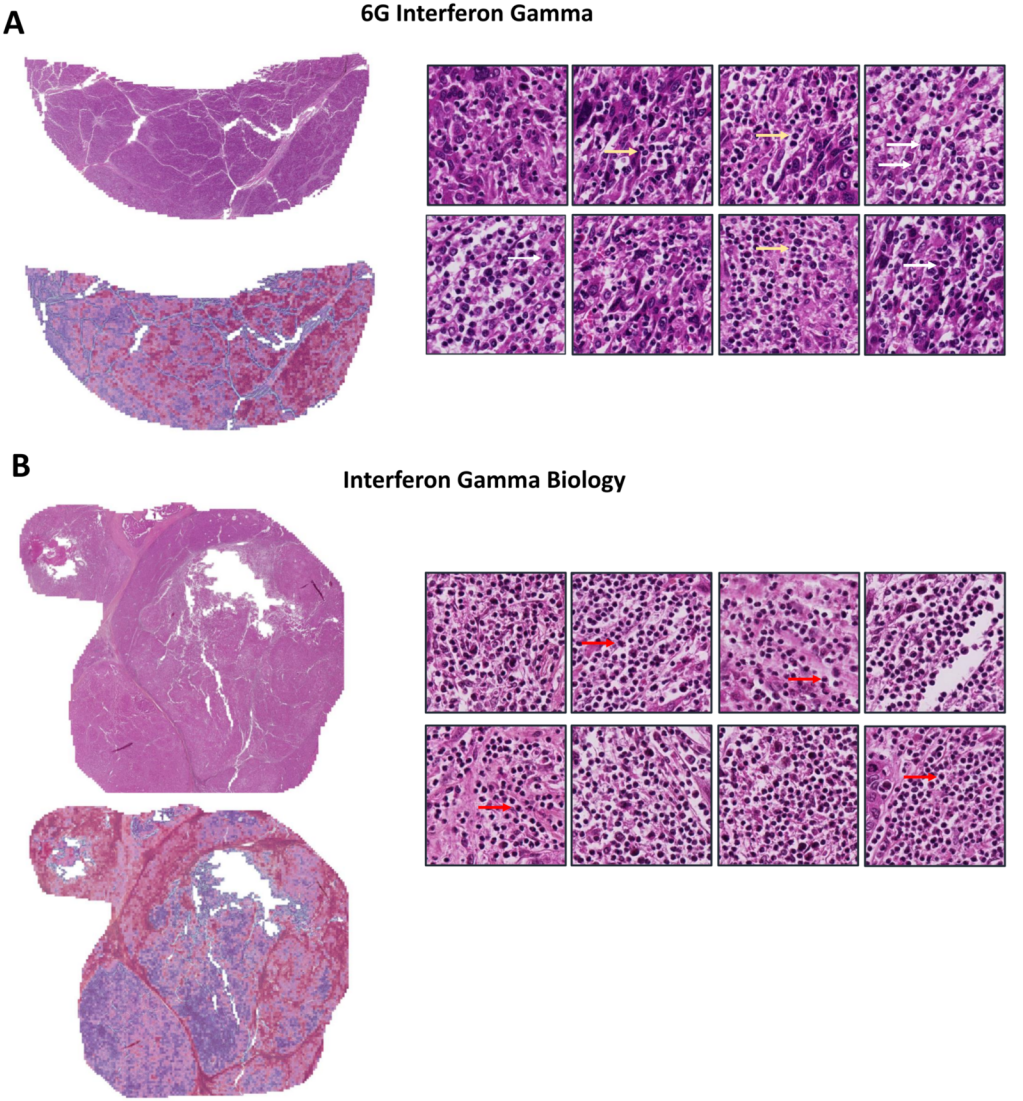
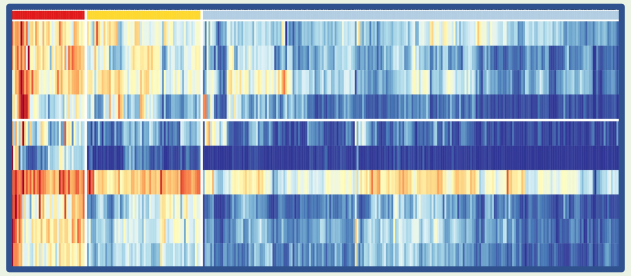
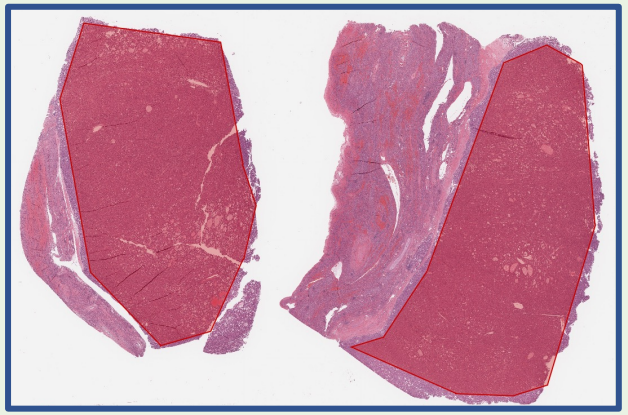


Figure 6. Pathological reviewing of highly predictive tiles. Microscopic examination of highly predictive tiles showed enrichment of particular immune-related features. Examples for 6G Interferon Gamma and Interferon Gamma Biology signatures are provided in panels A and B, respectively. Tiles associated with 6G Interferon Gamma included lymphocytes (yellow arrows) and plasma cells (white arrows) We also identified enrichment in neutrophils on tiles associated with Interferon Gamma Biology.

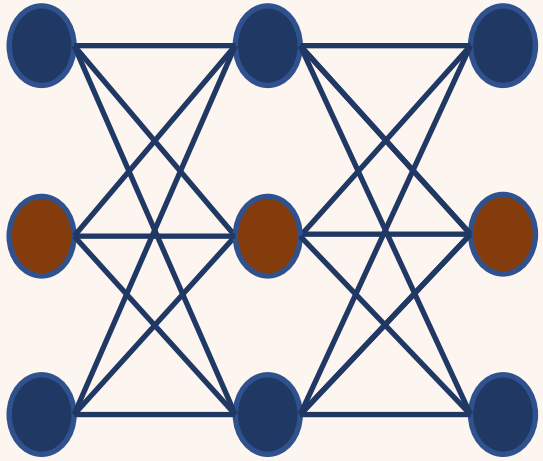
Discovery series (TCGA)

Status (labels) of 6 Immune gene signatures
+
Histological digital slides





Development of deep-learning models (Patch-based, Classic MIL, CLAM)



Validation series (Henri Mondor)

External validation
Prediction of Immune Status
directly from slides