Computational biomarker predicts lung ICI response via deep learning-driven hierarchical spatial modelling from H&E

Chensu Xie
Memorial Sloan Kettering Cancer Center
Chad Vanderbilt
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0002-8114-0237
Chao Feng
Memorial Sloan Kettering Cancer Center
David Ho
Memorial Sloan Kettering Cancer Center
Gabrielle Campanella
The Icahn School of Medicine at Mount Sinai
Jacklynn Egger
Memorial Sloan Kettering Cancer Center
Andrew Plodkowski
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0002-3772-370X
Jeffrey Girshman
Memorial Sloan Kettering Cancer Center
Peter Sawan
Memorial Sloan Kettering Cancer Center
Kathryn Arbour
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-5951-8427
Matthew Hellmann
AstraZeneca Oncology R&D
Thomas Fuchs (Thomas.Fuchs.AI@mssm.edu)
The Icahn School of Medicine at Mount Sinai

Article

Keywords:

Posted Date: January 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1251762/v1

Computational biomarker predicts lung ICI response via deep learning-driven hierarchical spatial modelling from H&E

Chensu Xie[*,1,2], Chad M. Vanderbilt[*,1], Chao Feng[1,2], David Ho[1], Gabrielle Campanella[1,2,#], Jacklynn V. Egger[3], Andrew J. Plodkowski[4], Jeffrey Girshman[4], Peter Sawan[4], Kathryn C. Arbour[3], Matthew D. Hellmann[3,+], Thomas J. Fuchs[1,#]

[1]Department of Pathology, [3]Department of Medicine, [4]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[2]Weill Cornell Graduate School of Medical Sciences, Cornell University, New York, USA

+: Current affiliation: AstraZeneca Oncology R&D, New York, USA

#: Current affiliation: The Icahn School of Medicine at Mount Sinai, Department of AI and Human Health, Hasso Plattner Institute for Digital Health at Mount Sinai, New York, USA

Body wordcount: 3109

Competing interests: T.J.F. is the founder, chief scientist and equity holders of Paige.AI Inc.. C.X. and T.J.F. have intellectual property interests relevant to the work that is the subject of this paper. MSK has financial interests in Paige.AI Inc. and intellectual property interests relevant to the work that is the subject of this paper. Matthew D. Hellmann reports grants from BMS; and personal fees from Achilles; Adagene; Adicet; Arcus; AstraZeneca; Blueprint; BMS; DaVolterra; Eli Lilly; Genentech/Roche; Genzyme/Sanofi; Janssen; Immunai; Instil Bio; Mana Therapeutics; Merck; Mirati; Natera; Pact Pharma; Shattuck Labs; and Regeneron; as well as equity options from Factorial, Immunai, Shattuck Labs, Arcus, and Avail Bio. A patent filed by Memorial Sloan Kettering related to the use of tumor mutational burden to predict response to immunotherapy (PCT/US2015/062208) is pending and licensed by PGDx. Subsequent to the completion of this work, M.D.H. began as an employee (and equity holder) at AstraZeneca.

Abstract

Determining which lung cancer patients are likely to respond to immune checkpoint inhibitors (ICI) remains a crucial challenge. Existing FDA-approved biomarkers lack sensitivity and specificity for identifying treatment candidates. To overcome this problem, we present a computational biomarker for predicting ICI response directly from routine H&E stained whole slide images of the initial biopsy. To achieve this, we developed an end-to-end deep learning system (EPL-GNN) that performs hierarchical spatial modeling on whole slide images to learn both spatial and morphological features from 2.1 billion cells and output a response score for each patient. The computational biomarker was trained and evaluated on the largest reported cohort of stage 4 lung cancer patients with ICI treatment response (N=583), resulting in an AUC of 0.69 and sensitivity of 91% on the independent test cohort, which compares favorably to PD-L1 immunohistochemistry (IHC) with an AUC of 0.68 and sensitivity of 57%, and tumor mutation burden (TMB) with an AUC of 0.62. The EPL-GNN model correctly identified 81% of the patients with a negative PD-L1 IHC result as responders. Visualizations of the hierarchical spatial

[*] These authors contributed equally

model revealed potential cellular patterns that correspond to ICI treatment response. In addition to the increased sensitivity achieved by the EPL-GNN model, H&E-based Computational Biomarkers offer a faster, less expensive, more objective and reproducible alternative or adjunct to existing IHC or sequencing based biomarkers.

Main

Immuno-oncology by immune checkpoint blockade

Lung cancer remains the leading cause of cancer death worldwide in 2020 [1]. Immune checkpoint inhibitors (ICI) are increasingly being used in treatment protocols for non-small cell lung cancer (NSCLC). While ICI therapy can lead to dramatic responses and clinical remission, clinical benefit is limited to a subset of all lung cancer patients. Thus, identifying the patients who benefit from ICI therapy is among the largest challenges in clinical practice. Extensive explorations for biomarkers have been undertaken to identify clinical features that correspond to tumor response to ICI, spanning from demographic associations to complex genomic investigations [2, 3, 4]. To date, the Food and Drug Administration (FDA) of the United States has approved two biomarkers which are used as testing modalities for predicting response to ICIs: PD-L1 protein expression by immunohistochemistry (IHC) protocol and tumor mutation burden (TMB) by large panel somatic next-generation sequencing (NGS) assays [5]. While these biomarkers improve response rates relative to unselected populations, they suffer from challenges. PD-L1 IHC suffers from the need to expend tumor tissue for testing, high preanalytical variability, and significant interobserver variability [6, 7, 8]. TMB requires large amounts of tumor tissue with frequent technical failures and the threshold for what is considered “high” TMB has not been standardized across platforms, although efforts in this regard are ongoing [9]. NGS as a methodology is limited by important aspects such as a turnaround time of 10 or more days, logistical complexity of transferring physical material between laboratories, technical challenges, high cost, and frequency of quantity insufficient samples [10]. The challenges mentioned above in performing the currently approved methodologies often prevent patients from being assessed for clinical benefit from ICI.

Emergence of Computational Pathology

Computational pathology is an emerging field of technologies that is rapidly gaining clinical acceptance. The FDA recently authorized the use of a computational model for clinical use in prostate cancer [11]. Computational biomarkers create opportunities to overcome many of the limitations of tissue-based biomarkers as they utilize existing standardized protocols, such as H&E histopathology images, and thus no additional tissue is consumed. In the era of COVID-19, the FDA has cleared digital pathology image viewers for clinical sign out [12]. Like the digitization of radiographic images, the digitization of pathology opens up many additional use cases for digital microscopy images. For example, digital pathology allows for remote viewing by histopathologists and immediate retrieval of archived cases. The adoption of digital pathology opens up the opportunity for deployment of systems that can predict response to therapy, which we refer to as computational biomarkers.

Computational models for predicting response to ICI in lung cancer

The current computational pathology efforts for predicting ICI response in lung cancer fall into two categories: First, computational models that reproduce existing biomarkers, such as PD-L1 IHC and TMB, from histopathology images [13, 14]. Second, computational models that correlate clinical data directly to prognosis [15, 16, 17]. Computational models trained to learn from manual PD-L1 IHC interpretations

remain subjected to the variation of staining protocols between laboratories and the subjective interpretations of the pathologists who provide the labels for training [6, 7, 8] . More importantly, due to the fact that these models are attempting to replicate manual IHC interpretation, their predictive power is logically limited at the upper-bound by the limit of PD-L1 IHC expression. Likewise, models trained to predict TMB suffer from the same limitation. These approaches that predict proxy biomarkers instead of outcome cannot surpass the limitations of the IHC or NGS biomarkers they were trained on. Some published models are trained against overall survival. Unfortunately, overall survival does not isolate clinical benefit from ICIs as additional variables, such as response to other therapies, age, underlying conditions, etc., often have a greater impact on the survival than the response to ICI therapy. To overcome the limitations of previous studies, we have acquired digital H&E histopathology images from patients with ICI response measured by a standard protocol developed to isolate the specific response to therapy Response Evaluation Criteria in Solid Tumors (RECIST) [18] that categorize patients into responder and non-responders via assessment of standardized pre and post treatment CT scans.

AI-based hierarchical spatial modelling for histology-outcome mapping

We present the first computational biomarker that predicts response to ICI therapy for lung cancer, measured by RECIST, directly from standard H&E histopathology slides. To achieve this, we developed a novel deep learning framework named EPL-GNN. The model incorporates deep features of 2.1 billion cells based on self-supervised learning as input, performs hierarchical spatial modelling from cell to tissue level in the context of the entire gigapixel-scale digitized slide, and learns to map image signals to patient response determined by the RECIST criteria in an end-to-end manner. We demonstrate that this trained model surpasses the prediction power of the current FDA-approved biomarkers for lung ICI response prediction in a direct comparison on the largest reported cohort of NSCLC patients (N=583) with ICI therapy and RECIST response measurements reported to date.

Overall workflow and EPL-GNN

Figure 1 presents the high-level workflow of the computational system. As ICI target the interaction between ligands and receptors on lymphocytes and tumor cells, previous studies suggest the correlation between tumor immune microenvironment and the ICI treatment response [19, 20, 21]. Based on this hypothesis, we first consider a computational quantitative analysis of lymphocyte-tumor cell interactions on digital histology slides. A state-of-the-art deep-learning based cell nuclei detector, vector-oriented confidence accumulation (VOCA) [22], is used to determine the spatial location of lymphocytes and tumor cells on all H&E-stained whole slides from the patient cohort. Clusters of lymphocytes and tumor cells are identified using spatial clustering, specifically the DBSCAN algorithm [23], and described by their enclosing polygons. We define the intersections, which represent extensive colocalization of lymphocytes and tumor cells as infiltration hotspots (Figure 2a). Two scores describing slide level infiltration are calculated (c.f. methods Section). We construct cell graphs for each infiltration hotspot tile: 1) each pair of cells within a Euclidian distance of 25 microns is connected by an edge., 2) each cell is described by morphology features that were extracted using a self-supervised deep learning model called Nuc2Vec [24]. In aggregate, EPL-GNN performs hierarchical spatial analysis based on these cell graphs and the slide infiltration scores to generate a probability score for response to ICI.

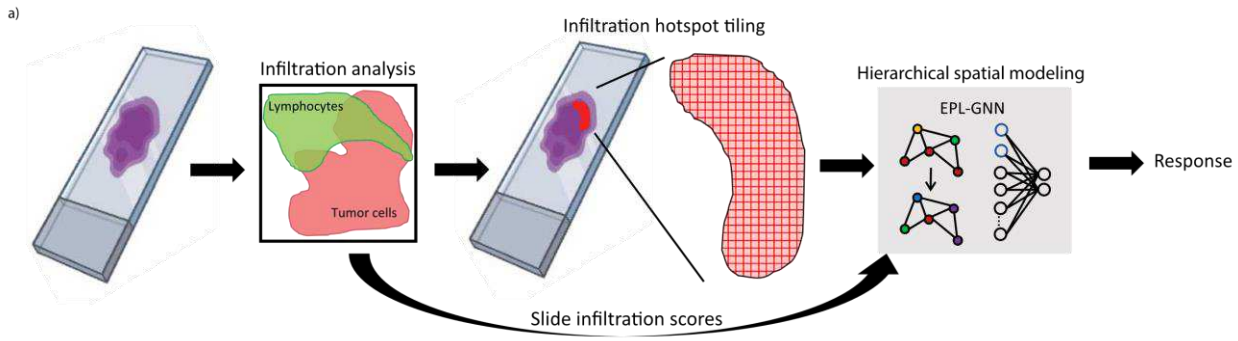


Figure 1. The high-level workflow of the proposed computational system for lung ICI response prediction. A state-of-the-art deep-learning based cell nuclei detector, vector-oriented confidence accumulation (VOCA), is used to determine the spatial location of lymphocytes and tumor cells on all H&E-stained whole slides from the patient cohort. Clusters of lymphocytes and tumor cells are identified using spatial clustering, specifically the DBSCAN algorithm, and described by their enclosing polygons. We define the intersections, which represent extensive colocalization of lymphocytes and tumor cells as infiltration hotspots Two scores describing slide level infiltration are calculated We construct cell graphs for each infiltration hotspot tile: 1) each pair of cells within a Euclidian distance of 25 microns is connected by an edge., 2) each cell is described by morphology features that were extracted using a self-supervised deep learning model called Nuc2Vec. In aggregate, EPL-GNN performs hierarchical spatial analysis based on these cell graphs and the slide infiltration scores to generate a probability score for response to ICI.

EPL-GNN adopted the end-to-end part learning (EPL) framework [25], which is a recent breakthrough for end-to-end WSI-outcome mapping, to learn and model a wide array of features over the entire whole slide image. It has proven to be successful both on traditional tasks like cancer identification [25] as well as prognosis predictions [26]. The EPL-GNN system presented here extends this model into the spatial domain by replacing the encoder of EPL with a graph neural network (GNN). The GNN encoder performs spatial modeling on the cell graph of hotspot tiles through graph convolutional layers and maps the infiltration hotspot tiles to a feature space that represents both cell morphology and cell-cell interactions. Next, EPL-GNN groups feature embeddings of all tiles from a slide into k clusters, each of which represents a subtype of learned cell patterns. For each cluster, the tile nearest to the cluster centroid in feature space was selected by EPL-GNN as representative for the subtype (Figure 2e). The final aggregation layer integrates tile-level subtype feature vectors and the slide-level infiltration scores to output the response probability. During training, the embedding vectors of all tiles belonging to a cluster are pushed towards the corresponding cluster centroid for a compact subtype representations and easier differentiation to other clusters. For model introspection and explainability the locations of tiles belonging to each cluster can be visualized on the infiltration heatmaps and represent distinct spatial patterns (Figure 2f). In summary, EPL-GNN performs a deep hierarchical spatial modeling at both cellular and tissue level, aggregates a variety of information from all scales over the whole slide, and predicts the ICI response. The whole system is trained end-to-end and requires only an H&E slide as input to predict the response for a patient.

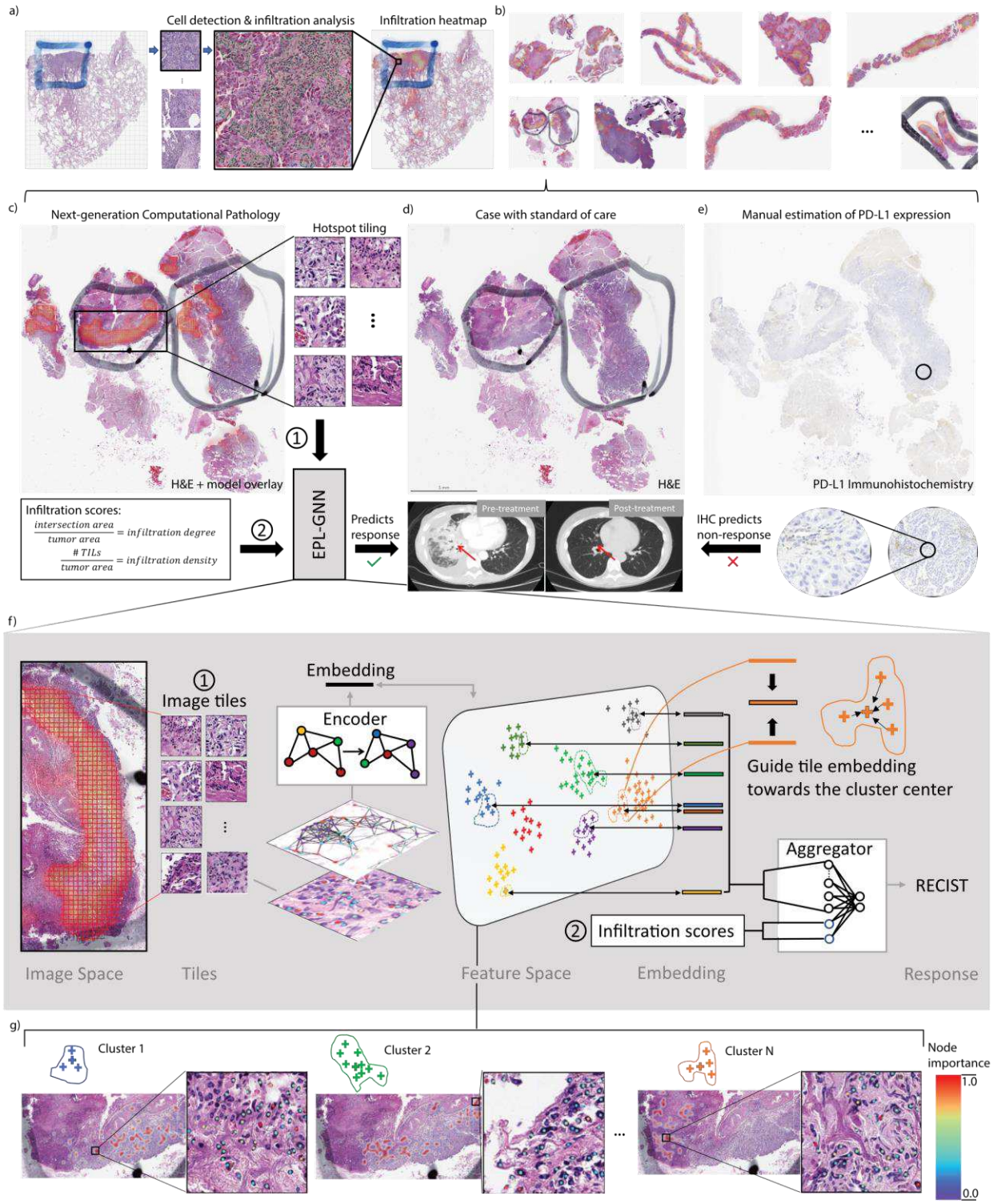


Figure 2. Illustration of the pipeline with individual example cases. a) Slides were tiled as large patches (448x448 microns) for infiltration analysis as described in Figure 1. b) Infiltration hotspots presented as heatmaps of multiple cases of the cohort. c-e) Comparison between computational pathology and manual assessment of PD-L1 expression on an example case. The patient is identified as a non-responder by post-treatment CT scan assessments according to the RECIST criterion. Manual estimation of PD-L1 expression of this patient based on IHC was negative. EPL-GNN took all image tiles at infiltration hotspots along with slide infiltration scores as input, and correctly predicted the patient as a non-responder with high model confidence. f) Illustration of EPL-GNN model. First, GNN encoder performs spatial modeling on the cell graph of hotspot tiles through graph convolutional

layers and maps the infiltration hotspot tiles to a feature space that represents both cell morphology and cell-cell interactions. Next, feature embeddings of all tiles from a slide are grouped into k clusters, each of which represents a subtype of learned cell patterns. For each cluster, the tile nearest to the cluster centroid in feature space was selected as representative for the subtype. The final aggregation layer integrates tile-level subtype feature vectors and the slide-level infiltration scores to output the response probability. During training, the embedding vectors of all tiles belonging to a cluster are pushed towards the corresponding cluster centroid for a compact subtype representations and easier differentiation to other clusters. g) Visualization of the learned subtype of cell patterns. The tiles belonging to each cluster are viewed as a heatmap according to their distances to the corresponding centroid in feature space. Each cluster of tiles show specific spatial patterns. The representative tile that EPL-GNN selected as input for each cluster are also presented with the importance of each single cell nucleus for ICI response prediction.

Results

We report results on the largest outcome dataset to-date, comprising 583 NSCLC patients with pre-treatment H&E slides and response data measured on post-treatment CT scans according to the RECIST criterion. On the 583 digitized whole slides, 2.1 billion cells were detected, and classified into 1.3 billion tumor cells and 0.8 billion lymphocytes. First, a training set (N=417) was compiled and used for cross-validation of the EPL-GNN model. Subsequently, an independent test set (N=147) was obtained for final evaluation. Table 1 compares the characteristics of the training and test set.

Table 1. Data characteristics of the training and independent test patient cohorts. All patients are non-small cell lung cancer (NSCLC) patients. Manual estimation of PD-L1 expression is available on a subset of the cohorts.

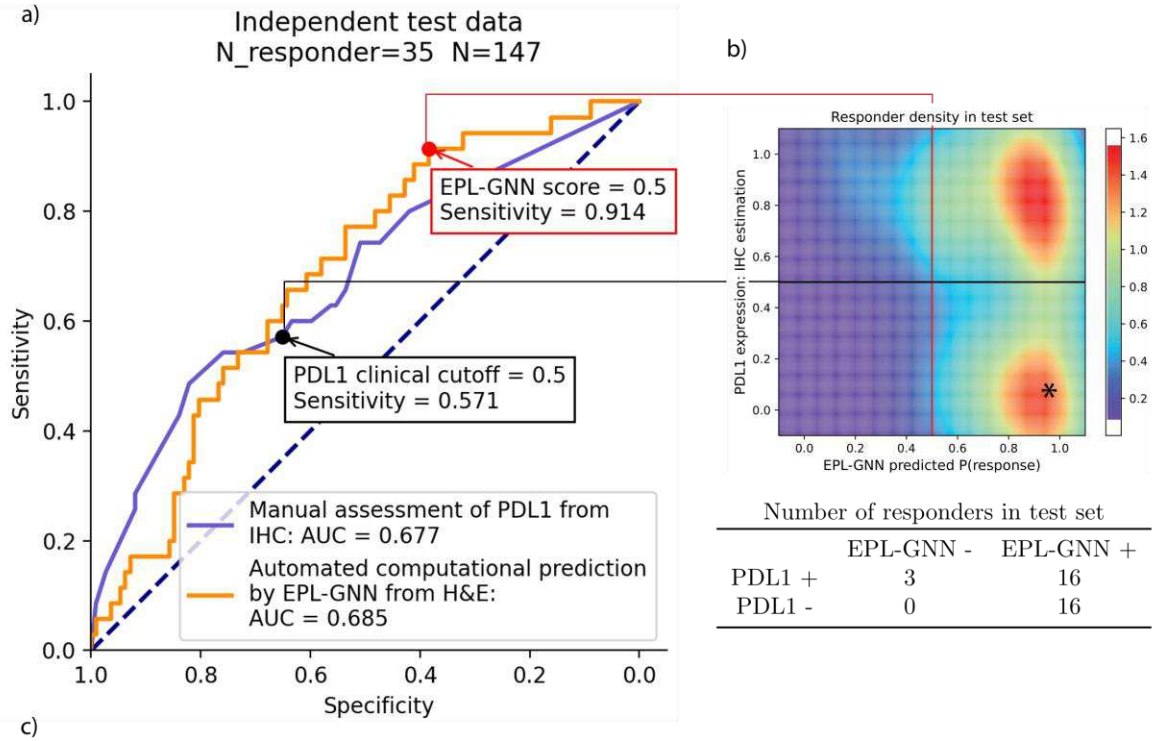
Characteristics	Training (n=417) n (%)	Test (n=166) n (%)
Age, median (range)	67 (30 – 93)	69 (37 – 89)
Sex		
Male	200 (48)	79 (48)
Female	217 (52)	87 (52)
Performance status		
ECOG 0/1	385 (92)	153 (92)
ECOG 2+	32 (8)	13 (8)
Smoking status		
Current/former	363 (87)	139 (84)
Never	54 (13)	27 (16)
Histology		
Adenocarcinoma	312 (75)	109 (66)
Squamous	67 (16)	29 (18)
Adenosquamous	4 (1)	1 (<1)
Large cell	16 (4)	7 (4)
NSCLC, NOS	18 (4)	20 (12)
Line of therapy		
1	163 (39)	71 (43)
2	189 (45)	74 (44)
≥3	65 (16)	21 (13)
Therapy type		
Anti-PD-(L)1 monotherapy	364 (87)	145 (87)
Anti-PD-(L)1+CTLA-4 combination	53 (13)	21 (13)
PD-L1 expression		
0	116/286 (41)	54/147 (37)
1%-49%	53/286 (18)	34/147 (23)
≥50%	117/286 (41)	59/147 (40)
Unknown	131/417 (31)	19/166 (11)
Best overall response		
Complete Response (CR) / Partial Response (PR)	77 (19)	36 (22)
Stable Disease (SD)	109 (26)	37 (22)
Progressive Disease (PD)	231 (55)	93 (56)
Tumor mutation burden		
≥10 mut/Mb	274 (66)	109 (66)
<10 mut/Mb	143 (34)	57 (34)

Computational model rescues responders with low PD-L1 expression in terms of the area under receiver-operator curve (AUROC), EPL-GNN has comparable performance to PD-L1 IHC on the independent test data showing only a slightly higher AUROC (0.685 vs 0.677) (Figure 3a). The AUROC for TMB is 0.623. However, EPL-GNN predicts response likelihood with notably higher sensitivity comparing to PD-L1 IHC. Sensitivity is not calculated for TMB as there is no established clinical threshold. The threshold for determining positivity of PD-L1 IHC has been established at 50%, which is the cutoff used for FDA approval of pembrolizumab and atezolizumab [27, 28, 29]. The 50% cutoff was established based on the ability of multiple observers to achieve agreement. Lowering the cutoff when manually interpreting PD-L1 IHC stains to increase sensitivity of the IHC test is not practical since inter-observer variability significantly goes up when reducing the threshold below 0.5 [8]. At this cutoff, PD-L1 has a sensitivity of 57% on the test set. At inference stage, EPL-GNN is a deterministic system that produces consistent, reproducible outputs. Using the same cutoff as PD-L1 IHC of 0.5 at the softmax score of the neural network, the sensitivity of EPL-GNN is 91%. Importantly, the computational biomarker identifies responders with low PD-L1 expression. Of the 43% of responders with low PD-L1 expression, EPL-GNN rescued 81% of these patients (Figure 3b).

EPL-GNN combined with PD-L1 expression for ICI response

Because EPL-GNN learns cell patterns based on hierarchical spatial modeling, it is independent of PD-L1 IHC expression. There was no correlation between the score of EPL-GNN and the PD-L1 IHC (*r* score = 0.086). The combination of EPL-GNN and PD-L1 IHC using a multi-layer perceptron (MLP) resulted in an AUROC of 0.741(+-) on the test set (Figure 3c). Thus, the incorporation of IHC into a computational biomarker show better performance characteristics than either PD-L1 or EPL-GNN independently.

The technical costs associated with computational biomarkers when deployed in the clinic are limited to slide scanning and minimal computational costs for inference when analyzing new slides. The system can be deployed in the cloud or on premises using a single off-the-shelf graphical processing unit (GPU). Model inference takes only minutes from image capture to response prediction and can be reported well in advance of any manual IHC reads. H&E slides are required for all routine cancer diagnoses and therefore no additional tissue must be expended for more tissue sections or staining. Finally, H&E-based computational biomarker do not suffer from the staining quality variation of PD-L1 IHC nor inconsistent manual interpretation, and generates consistent, reproducible predictions. The technical costs and approximate turnaround times of IHC, sequencing and computational biomarkers are listed in Figure 3c.



Number of responders in test set

	EPL-GNN -	EPL-GNN +
PDL1 +	3	16
PDL1 -	0	16

c)

Comparison between FDA approved biomarkers and EPL-GNN

Biomarker	Technology	Technical cost ($)	Time	Cross Validation ROC AUC		Independent Test ROC AUC		Assessment
				All cases (N=417)	PDL1 estimated cases (N=286)	All cases (N=166)	PDL1 estimated case (N=147)	
PDL1	IHC	~200	~2d	-	0.700	-	0.677	Subjective
TMB	Sequencing	~2000	~5 weeks	0.634	0.634	0.641	0.623	Subjective
EPL-GNN	HE	**<10**	**~10 minutes**	**0.701**	**0.719**	**0.680**	**0.685**	**Objective**
EPL-GNN + PDL1				-	0.749	-	0.741	

Figure 3. Comparison of prediction performance for lung ICI between EPL-GNN and current FDA-approved biomarkers. a) The ROC curves of automated computational prediction by EPL-GNN and manual assessment of PD-L1 from IHC on our independent test set (N=147). The AUC scores of EPL-GNN and PD-L1 IHC are 0.685 and 0.677 respectively. However, the sensitivity of PD-L1 at its clinical cutoff 50% is 57%. EPL-GNN has a sensitivity of 91% using the same cutoff as PD-L1 IHC of 0.5 at the softmax score of the neural network. b) Top: responder density map generated from the scattered plot between EPL-GNN score and PD-L1 expression level of the 35 responders in test set. Bottom: 2x2 table of responders with different PD-L1 and EPL-GNN prediction results. 16 of the 35 responders have PD-L1 expression level <0.5. EPL-GNN correctly identifies 16 of the 19 (81%) responders with a negative PD-L1 result. c) Table comparing FDA approved biomarkers (PD-L1 and TMB) and EPL-GNN. Technical cost is approximate cost of technology in USD. Time is turn-around time from when H&E slide is generated to reporting of results. PD-L1 estimated cases are cases with manual PD-L1 score available.

Interpretable spatial heterogeneity

Investigators have speculated on the relevance of tumor cell and immune cell interactions for patient response to immunotherapies [30, 31]. EPL-GNN can not only be used as a robust, highly sensitive biomarker for lung ICI, but can also shed light on the spatial heterogeneity of cancer at multiple scales by model introspection and visualization.

We first visualize the importance of each cluster estimated by absolute value of the gradients with respect to the slide loss of the features as described by [25]. The model focuses on the feature vector of one representative tile for each cluster in feature space (Figure 2e) and combines the information of all clusters for the final prediction. Figure 4a shows the representative tiles of ten responders and non-

responders with high model confidence. Each column represents one slide/case, and each row stands for a feature cluster. The importance of each cluster is color-coded and mapped back onto the original slide (Figure 4b).

The importance for response prediction of each single cell nucleus can also be calculated from the learned parameters of the GNN encoder (c.f. methods Section). The importance score of each individual cell reflects the cumulative information from a cell graph of its vicinity rooted at the center of the nucleus. It includes information about cell morphology and the interaction with the surrounding cells (Figure 4c). Being able to visually represent the outputs of the model at various resolutions *in situ* allows in-depth analysis of the computational biomarker. One can observe that non-responders have a higher proportion of cluster 1 (p=0.026), while cluster 2 shows a trend towards enrichment in responders (p=0.085) (Figure 4d).

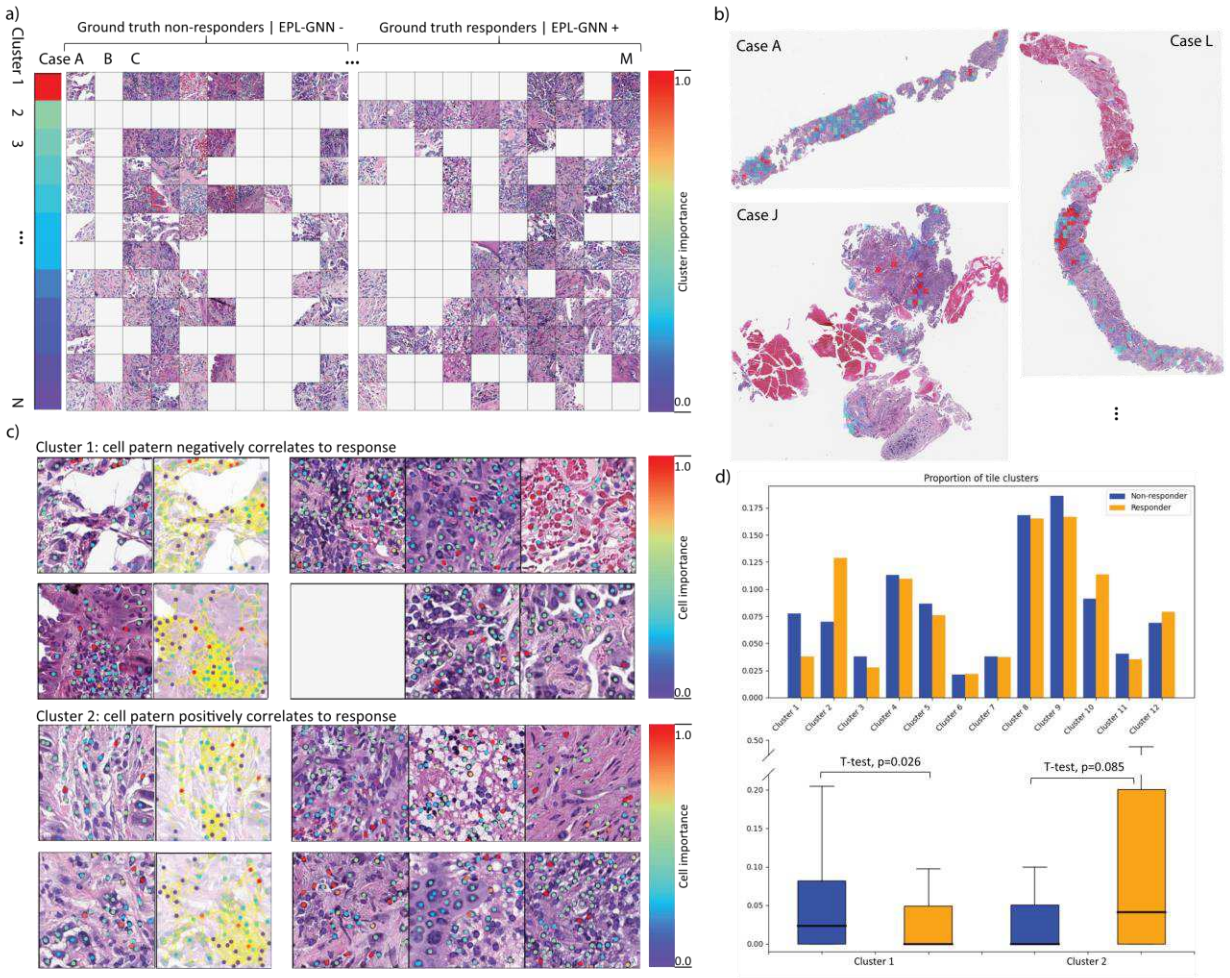


Figure 4. Visualization of the learned information of EPL-GNN at cell and tissue levels. a) The representative tiles learned and analyzed by EPL-GNN for ICI response prediction on 10 responder and 10 non-responder slides with high model confidence. Each column represents one slide (case), and each row is a tissue subtype (tile clusters). A light grey padding means that a specific tissue subtype is missing in the individual slide. Clusters are ranked vertically with their importance to response prediction. The top 2 important clusters show strong correlation to either response or non-response. b) Tile importance mapped back onto the original tissue visually showing the in situ pattern. c) Visualization of the cell graphs and cell importance on tiles of the two most important clusters. For each cluster, 2 tiles are shown side by side with the cell graphs. Importance of each node (cell) for ICI response prediction are colored. d) Top: The proportion of the 12 tile clusters averaged over all test slides. Bottom: statistical

comparison of cluster 1 and 2 proportion in responders and non-responders. Non-responders have higher proportion of cluster 1 (p=0.026), while cluster 2 shows a trend toward enrichment in responders (p=0.085).

Discussion

Clinical utility of computational biomarkers based on H&E stain

The hematoxylin and eosin (H&E) stained tissue is the cornerstone of surgical pathology. Every tissue reviewed by a surgical pathologist receives an H&E stain and thus H&E stained slides are the most ubiquitous and standardized images generated in histopathology [32]. The ability for a computational biomarker to utilize H&E stained slides allows for broad adoption at point of care as no additional technical procedures are required. In contrast, models trained to predict ICI response from IHC slides require cutting additional sections and applying less standardizable protocols with high pre-analytical variability.

The challenge of ICI response prediction

To overcome the challenge of ICI response prediction from H&E slides, we approached the problem from multiple different angles. Our initial efforts included simple hypothesis-driven cell-based spatial statistics, traditional convolutional neural network (CNN) based tile classification, and explorative end-to-end mapping. After modest results with these techniques, we eventually converged on a hybrid of hypothesis-driven and explorative hierarchical spatial modelling (EPL-GNN).

The driving hypothesis for modeling the relationship between tumor immune microenvironment and ICI response is that patients with a high level of tumor infiltrating lymphocytes (TILs) are more likely to respond to ICI treatment in multiple cancer types [19, 20, 21]. In our experiments, the most predictive metrics are two scores describing the infiltration of lymphocytes over the whole slides: infiltration degree and infiltration density (c.f. methods Section, Table 2). Next, we considered traditional convolutional neural network (CNN) based tile classification. The simplest baseline is to train a CNN model that maps each image tile of a slide to a response probability score, and then average their scores as the response probability for the whole slide. Alternatively, one can average the feature vectors of all tiles to represent the global slide feature vector which can then be mapped to the response score. Both of these traditional CNN approaches didn't show strong signal for ICI response (Table 2). We also tested multiple instance learning (MIL), which has proven to be tremendously successful for cancer detection on datasets with tens of thousands of patients [33]. However, we could not achieve a performance comparable to other methods for ICI response prediction with MIL (Table 2). We thus concluded that the local morphological features carried in single tiles alone do not contain adequate information to predict patient outcome of ICI treatment.

To overcome these deficiencies, we explored the EPL framework [25] for end-to-end modeling on diverse features over whole slides. The original EPL model uses a CNN as encoder to learn subtypes of morphological features. While this approach improved beyond slide-wide infiltration features, it is known that CNNs are suboptimal for learning Cartesian, spatial features [34]. Motivated by the hypothesis that the relevant features for ICI response are likely carried by lymphocyte-tumor cell interactions, we replaced the EPL encoder with a graph neural network (GNN) which learns local spatial features on cell graphs. Finally, the two scores describing global infiltration were concatenated to the slide embedding followed by the EPL aggregator. This hybrid approach of hypothesis-driven and

explorative hierarchical spatial modelling displayed significant improvement over all competing methods listed in Table 2.

Table 2. Performance of baseline approaches for ICI response prediction measured in AUC on cross-validation. The hypothesis-driven cell-based spatial statistics (top panel) are predictive for the ICI response. The CNN baselines (mid panel), including multiple instance learning (MIL) [33], do not identify enough signal for the prediction. EPL-GNN (bottom panel) further improves on vanilla EPL and shows comparable performance to manual estimation of PD-L1 expression.

Comparison between EPL-GNN and baseline models

Approach	Cross Validation ROC AUC	
	All cases (N=417)	PD-L1 estimated cases (N=286)
Infiltration degree	0.616	0.624
Infiltration density	0.603	0.605
CNN average score	0.540	0.543
CNN average embedding	0.580	0.532
MIL	0.549	0.564
EPL-CNN	0.657	0.606
EPL-GNN	**0.701**	**0.719**

From diagnosis to response prediction

Research in computational pathology often focuses on diagnostic problems, including the detection and classification of cancer [33], cancer grading [35], or metastasis analysis [36]. These approaches use human assessment as gold standard and produce a computational replacement or adjunct tool. Alternatively, predictive computational biomarkers are trained against clinical outcome as gold standard, for example response to treatment in case of ICIs. However, for tasks such as ICI response prediction, the location and nature of predictive features are unknown to pathologists. An approach that is able to model on a wide array of information over the whole slide is a prerequisite.

The prediction function for cancer identification in literature can be summarized as a weighted average of tumor features in tiles: $Y_{diagnosis} = \sum_i \alpha Z_i$, where α is the learnable linear weights, and Z_i is the feature in tile i. EPL-GNN describes a slide in terms of K subtypes of features, and maps from the concatenation of these features to the response: $Y_{response} = f(MLP(< Z_1, Z_2, \dots, Z_K >))$, where MLP represents a multi-layer perceptron aggregator, and f is the activation function for the output. This approach enables end-to-end explorative studies for tasks in which the predictive features are unknown *a priori*.

The largest reported cohort for lung ICI response

The patient cohort presented here is the largest reported dataset for lung ICI response prediction based on RECIST criteria to-date. There are multiple challenges of obtaining a large and relevant dataset for this study. These include the identification of patients who meet inclusion criteria (c.f. Data Section) and have response to ICI documented in a standard protocol such as RECIST. Likewise, these patients must have a pre-treatment biopsy and long-term follow-up. While we consider the size of the cohort a strength of the study, the patients were all treated at the same academic medical center. Multiple attempts were made to identify external cohorts, but no satisfactory dataset of any significant size could

be obtained. With growing adoption of ICIs in clinical practice, we hope to overcome this challenge in future work.

Conclusion

We present the first computational biomarker to predict ICI response directly from H&E images. The system outperforms the FDA-approved biomarkers for predicting response to ICI therapy in terms of cost, speed, reproducibility and increased sensitivity. This was achieved by curating the largest cohort of stage 4 lung cancer patients with ICI treatment and RECIST response quantification. The novel deep learning architecture for spatial pathology was developed to analyze more than 2 billion lymphocytes and cancer cells. Given that the system relies only on standard H&E images, this computational biomarker could be easily implemented in clinical practice to overcome many of the limitations of the current FDA-approved biomarkers.

Method

Data

Patients were selected for metastatic NSCLC treated with PD-L1 blockade-based immunotherapy between 2013-2019 for the discovery and validation cohort. In both cohorts, those treated with combination therapy of PD-(L)1 and chemotherapy were excluded. Inclusion criteria were digitized H&E images as well as available outcome data from their response to PD-(L)1 therapy. Objective overall response was determined by RECIST, performed by a blinded thoracic radiologist. Patients who did not progress were censored at the time of their last available imaging assessment. Clinical data was locked as of 09/2019. All handling of slide image data was through a digital slide viewer developed at MSK [37].

Lymphocyte and tumor cell detection with VOCA

We used a state-of-the-art cell nuclei detection method called vector-oriented confidence accumulation (VOCA) [22] for the detection of lymphocytes and tumor cells. Over 6000 lymphocytes and tumor cells were manually annotated on a software called “MSK Slide Viewer” [38] and used as the ground truth to train the VOCA model. The final model was used to for detection of more than 2 billion cells on the whole cohort. The threshold of filtering out false positive detections was also learned automatically by the model. The coordinates of each lymphocyte and tumor cell were saved as the input for further feature extraction.

Cell feature extraction with Nuc2Vec

For each detected nucleus, we extracted the surrounding image patch of size 64x64 pixels, rescale it to 128x128, and did a forward pass through a resnet-34 model pretrained with the Nuc2Vec method on slides from the MSKCC-IMPACT patient cohort, to obtain a 128-dimenstional embedding vector. Briefly, Nuc2Vec utilizes a contrastive learning method to learn vector embeddings of local image patches centered around each nucleus, such that the Euclidean distance in the embedding space is approximating a metric of the morphological similarity between nuclei. Thus, the embedding vector represents the morphology of each nucleus. For a detailed description of the method, training dataset, and quality of the learned vector embeddings, we refer to [24].

Infiltration analysis

Using the detection results of VOCA, we applied DBSCAN [23] to identify cell clusters, represented as polygons, of both tumor cells and lymphocytes. For lymphocytes, we consider cells belonging to the same cluster if they are located less than 40 pixels (20 microns) from each other, while 50-pixel was chosen for tumor cells. The polygons generated by DBSCAN represent cell clusters, and the intersection between the lymphocyte clusters and tumor cell clusters are considered potential infiltration regions. In addition to the local spatial configuration of cells, we calculated two scores capturing infiltration information at the whole slide level. One is the infiltration degree:

$$infiltration\ degree = \frac{intersection\ area}{tumor\ area}$$

This score describes the degree of infiltration over the whole slide defined as the ratio between the area of lymphocyte-tumor cell cluster intersection and the area of tumor cell clusters. The second score is:

$$infiltration\ density = \frac{\#\ TILs}{tumor\ area}$$

The infiltration density counts the number of lymphocytes within the lymphocyte-tumor cell cluster intersection and divides it by the area of tumor cell clusters. Experiments of using these two scores alone for ICI response prediction showed a predictive signal for ICI (c.f. Table 2).

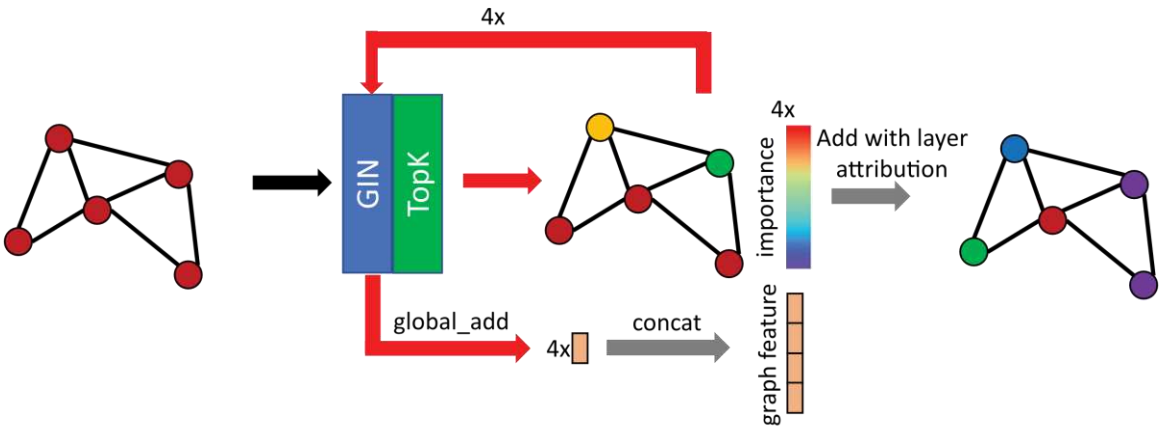
Graph Neural Networks (GNN) of cells

From all infiltration hotspots we sampled image tiles of size 112x112 microns without overlap. The cell graphs on these image tiles are used as input to the GNN encoder. To construct the cell graphs, the nodes represent the detected cells and are described using the feature representation from Nuc2Vec. Finally, cell graphs are constructed as radius graphs [39] where all co-occurrences within 25 microns are modeled as an edge in the graph.

GNN encoder architecture

We constructed a GNN encoder based on graph isomorphism networks (GIN) [40]. The model contains four GIN layers, each is followed by a top-K pooling layer [41]. The output feature of each cell is the concatenation of the node features after each convolution, and the graph readout is the sum of all concatenated node features. The importance scores s for each node at each layer are learned by the top-K pooling function. Also, the layer attribution α can be estimated by the gradient as in the original EPL paper [25].Therefore, the node importance score is calculated as:

$$Node\ importance = \sum_{l=1}^{4} \alpha_l s_l$$



The GNN encoder gradually pools important cell patterns for the response prediction and encodes the cellular information of each infiltration tile, then multiple groups of these encoded cellular patterns are represented by their centroid in feature space and aggregated. Along with the quantitative slide level infiltration scores, these graphical deep features can produce a slide response score. Details of the overall framework architectures and training strategy are described in [25].

EPL framework for end-to-end WSI outcome mapping

End-to-end part learning [25] models a WSI as consisting of K groups of image tiles of similar features. Each group of tiles form a cluster in feature space, and the one tile nearest to the centroid is used as the representative tile for that cluster. K feature vectors, each of length l are concatenated and fed forward to the aggregation module, which can be multiple fully connected layers, to generate the output score. Meanwhile, the tile features of each cluster are also pushed towards the centroid to ensure that representative tile can approximate the common features of the cluster. In general, this framework enables the end-to-end WSI-outcome mapping, and theoretically can be applicable for any slide-wide labels (c.f. [25] for details).

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal and F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians,* vol. 71, no. 3, 2021.

[2] K. R. Voong, J. Feliciano, D. Becker and B. Levy, "Beyond PD-L1 testing-emerging biomarkers for immunotherapy in non-small cell lung cancer," *Annals of Translational Medicine,* vol. 5, no. 18, 2017.

[3] R. Cristescu, R. Mogg, M. Ayers, A. Albright, E. Murphy, J. Yearley, X. Sher, X. Q. Liu, H. Lu, M. Nebozhyn, C. Zhang, J. K. Lunceford, A. Joe, J. Cheng, A. L. Webber, N. Ibrahim, E. R. Plimack, P. A. Ott, T. Y. Seiwert, A. Ribas, T. K. McClanahan, J. E. Tomassini, A. Loboda and D. Kaufman, "Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy," *Science,* vol. 362, no. 6411, 2018.

[4] D. Lau, A. M. Bobe and A. A. Khan, "RNA Sequencing of the Tumor Microenvironment in Precision Cancer Immunotherapy," *Trends in Cancer,* vol. 5, no. 3, 2019.

[5] F. Sanger, S. Nicklen and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors.," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 74, no. 12, 1977.

[6] W. A. Cooper, P. A. Russell, M. Cherian, E. E. Duhig, D. Godbolt, P. J. Jessup, C. Khoo, C. Leslie, A. Mahar, D. F. Moffat, V. Sivasubramaniam, C. Faure, A. Reznichenko, A. Grattan and S. B. Fox, "Intra- and interobserver reproducibility assessment of PD-L1 biomarker in non–small cell lung cancer," *Clinical Cancer Research,* vol. 23, no. 16, 2017.

[7] J. A. Rehman, G. Han, D. E. Carvajal-Hausdorf, B. E. Wasserman, V. Pelekanou, N. L. Mani, J. McLaughlin, K. A. Schalper and D. L. Rimm, "Quantitative and pathologist-read comparison of the heterogeneity of programmed death-ligand 1 (PD-L1) expression in non-small cell lung cancer," *Modern Pathology,* vol. 30, no. 3, 2017.

[8] H. Brunnström, A. Johansson, S. Westbom-Fremer, M. Backman, D. Djureinovic, A. Patthey, M. Isaksson-Mettävainio, M. Gulyas and P. Micke, "PD-L1 immunohistochemistry in clinical diagnostics of lung cancer: Inter-pathologist variability is higher than assay variability," *Modern Pathology,* vol. 30, no. 10, 2017.

[9] A. Stenzinger, J. D. Allen, J. Maas, M. D. Stewart, D. M. Merino, M. M. Wempe and M. Dietel, "Tumor mutational burden standardization initiatives: Recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions," *Genes Chromosomes and Cancer,* vol. 58, no. 8, 2019.

[10] M. E. Gutierrez, K. Choi, R. B. Lanman, E. J. Licitra, S. M. Skrzypczak, R. Pe Benito, T. Wu, S. Arunajadai, S. Kaur, H. Harper, A. L. Pecora, E. V. Schultz and S. L. Goldberg, "Genomic Profiling of Advanced Non–Small Cell Lung Cancer in Community Settings: Gaps and Opportunities," *Clinical Lung Cancer,* vol. 18, no. 6, 2017.

[11] FDA, "FDA Authorizes Software that Can Help Identify Prostate Cancer," 21 09 2021. [Online]. Available: https://www.fda.gov/news-events/press-announcements/fda-authorizes-software-can-help-identify-prostate-cancer. [Accessed 26 10 2021].

[12] M. G. Hanna, V. E. Reuter, O. Ardon, D. Kim, S. J. Sirintrapun, P. J. Schüffler, K. J. Busam, J. L. Sauter, E. Brogi, L. K. Tan, B. Xu, T. Bale, N. P. Agaram, L. H. Tang, L. H. Ellenson, J. Philip, L. Corsale, E. Stamelos, M. A. Friedlander, P. Ntiamoah, M. Labasin, C. England, D. S. Klimstra and M. Hameed, "Validation of a digital pathology system including remote review during the COVID-19 pandemic," *Modern Pathology,* vol. 33, no. 11, 2020.

[13] J. A. Diao, J. K. Wang, W. F. Chui, V. Mountain, S. C. Gullapally, R. Srinivasan, R. N. Mitchell, B. Glass, S. Hoffman, S. K. Rao, C. Maheshwari, A. Lahiri, A. Prakash, R. McLoughlin, J. K. Kerner, M. B. Resnick, M. C. Montalto, A. Khosla, I. N. Wapinski, A. H. Beck, H. L. Elliott and A. Taylor-Weiner, "Human-interpretable image features derived from densely mapped cancer pathology slides predict

diverse molecular phenotypes," *Nature Communications,* vol. 12, no. 1, 2021.

[14] L. Sha, B. Osinski, I. Ho, T. Tan, C. Willis, H. Weiss, N. Beaubier, B. Mahon, T. Taxter and S. Yip, "Multi-field-of-view deep learning model predicts nonsmall cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images," *Journal of Pathology Informatics,* vol. 10, no. 1, 2019.

[15] C. Fang, D. Xu, J. Su, J. R. Dry and B. Linghu, "DeePaN: deep patient graph convolutional network integrating clinico-genomic evidence to stratify lung cancers for immunotherapy," *npj Digital Medicine,* vol. 4, no. 1, 2021.

[16] X. Bao, R. Shi, T. Zhao and Y. Wang, "Immune landscape and a novel immunotherapy-related gene signature associated with clinical outcome in early-stage lung adenocarcinoma," *Journal of Molecular Medicine,* vol. 98, no. 6, 2020.

[17] Y. Yang, J. Yang, L. Shen, J. Chen, L. Xia, B. Ni, L. Ge, Y. Wang and S. Lu, "A multi-omics-based serial deep learning approach to predict clinical outcomes of single-agent anti-PD-1/PD-L1 immunotherapy in advanced stage non-small-cell lung cancer," *American Journal of Translational Research,* vol. 13, no. 2, 2021.

[18] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe and J. Verweij, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *European Journal of Cancer,* vol. 45, no. 2, 2009.

[19] M. Binnewies, E. W. Roberts, K. Kersten, V. Chan, D. F. Fearon, M. Merad, L. M. Coussens, D. I. Gabrilovich, S. Ostrand-Rosenberg, C. C. Hedrick, R. H. Vonderheide, M. J. Pittet, R. K. Jain, W. Zou, T. K. Howcroft, E. C. Woodhouse, R. A. Weinberg and M. F. Krummel, "Understanding the tumor immune microenvironment (TIME) for effective therapy," *Nature Medicine,* vol. 24, no. 5, 2018.

[20] C. U. Blank, J. B. Haanen, A. Ribas and T. N. Schumacher, "The "cancer immunogram"," *Science,* vol. 352, no. 6286, 2016.

[21] D. S. Chen and I. Mellman, "Elements of cancer immunity and the cancer-immune set point," *Nature,* vol. 541, no. 7637, 2017.

[22] C. Xie, C. M. Vanderbilt, A. Grabenstetter and T. J. Fuchs, "VOCA: Cell Nuclei Detection In Histopathology Images By Vector Oriented Confidence Accumulation," *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning,* vol. 102, 2019.

[23] D. Arribas-Bel, M. Garcia-López and E. Viladecans-Marsal, "Building(s and) cities: Delineating urban areas with a machine learning algorithm," *Journal of Urban Economics,* 2019.

[24] C. Feng, F. Org, C. Vanderbilt and T. J. Fuchs, "Nuc2Vec: Learning Representations of Nuclei in Histopathology Images with Contrastive Loss," in *Medical Imaging with Deep Learning*, 2021.

[25] C. Xie, H. Muhammad, C. M. Vanderbilt, R. Caso, D. Vijay Kumar Yarlagadda, G. Campanella and T. J. Fuchs, "Beyond Classification: Whole Slide Tissue Histopathology Analysis By End-To-End Part Learning," in *Medical Image with Deep Learning*, 2020.

[26] H. Muhammad, C. Xie, C. S. Sigel, M. Doukas, L. Alpert, A. L. Simpson and T. J. Fuchs, "EPIC-Survival: End-to-end Part Inferred Clustering for Survival Analysis, with Prognostic Stratification Boosting," in *Medical Imaging with Deep Learning*, Lübeck, 2021.

[27] M. Reck, D. Rodríguez-Abreu, A. G. Robinson, R. Hui, T. Csoszi, A. Fülöp, M. Gottfried, N. Peled, A. Tafreshi, S. Cuffe, M. O'Brien, S. Rao, K. Hotta, K. Vandormael, A. Riccio, J. Yang, M. Catherine Pietanza and J. R. Brahmer, "Updated analysis of KEYNOTE-024: Pembrolizumab versus platinum-based chemotherapy for advanced non–small-cell lung cancer with PD-L1 tumor proportion score of 50% or greater," *Journal of Clinical Oncology,* vol. 37, no. 7, 2019.

[28] M. Reck, D. Rodríguez-Abreu, A. G. Robinson, R. Hui, T. Csőszi, A. Fülöp, M. Gottfried, N. Peled, A. Tafreshi, S. Cuffe, M. O'Brien, S. Rao, K. Hotta, M. A. Leiby, G. M. Lubiniecki, Y. Shentu, R. Rangwala and J. R. Brahmer, "Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer," *New England Journal of Medicine,* vol. 375, no. 19, 2016.

[29] R. S. Herbst, G. Giaccone, F. de Marinis, N. Reinmuth, A. Vergnenegre, C. H. Barrios, M. Morise, E. Felip, Z. Andric, S. Geater, M. Özgüroğlu, W. Zou, A. Sandler, I. Enquist, K. Komatsubara, Y. Deng, H. Kuriki, X. Wen, M. McCleland, S. Mocci, J. Jassem and D. R. Spigel, "Atezolizumab for First-Line Treatment of PD-L1–Selected Patients with NSCLC," *New England Journal of Medicine,* vol. 383, no. 14, 2020.

[30] K. AbdulJabbar, S. E. Raza, R. Rosenthal, M. Jamal-Hanjani, S. Veeriah, A. Akarca, T. Lund, D. A. Moore, R. Salgado, M. Al-Bakir, L. Zapata, C. T. Hiley, L. Officer, M. Sereno, C. R. Smith, S. Loi, A. Hackshaw, T. Marafioti, S. A. Quezada, N. McGranahan, J. Le Quesne, C. Swanton, M. Jamal-Hanjani, J. Le Quesne, A. Hackshaw, S. A. Quezada, N. McGranahan, R. Rosenthal, C. T. Hiley, S. Veeriah, D. A. Moore, T. Marafioti, R. Salgado, Y. Ngai, A. Sharp, C. Rodrigues, O. Pressey, S. Smith, N. Gower, H. Dhanda, J. Riley, L. Primrose, L. Martinson, N. Carey, J. A. Shaw, D. Fennell, G. A. Wilson, N. J. Birkbak, T. B. Watkins, M. Escudero, A. Stewart, A. Rowan, J. Goldman, P. Van Loo, R. K. Stone, T. Denner, E. Nye, S. Ward, E. L. Lim, S. Boeing, M. Greco, K. Litchfield, J. Nicod, C. Puttick, K. Enfield, E. Colliver, B. Campbell, C. Abbosh, Y. Wu, M. Skrzypski, R. E. Hynds, A. Georgiou, M. W. Sunderland, J. L. Reading, K. S. Peggs, J. A. Hartley, P. Gorman, H. L. Lowe, L. Ensell, V. Spanswick, A. Karamani, D. Biswas, M. Razaq, S. Beck, A. Huebner, M. Dietzen, C. Naceur-Lombardelli, M. A. Akther, H. Zhai, N. Kannu, E. Manzano, S. K. Bola, E. Ghorani, M. R. de Massy, E. Hoxha, E. Hatipoglu, S. Ogwuru, B. Chain, G. Price, S. Dubois-Marshall, K. Kerr, S. Palmer, H. Cheyne, J. Miller, K. Buchan, M. Chetty, M. Khalil, V. Ezhil, V. Prakash, G. Anand, S. Khan, K. Lau, M. Sheaff, P. Schmid, L. Lim, J. Conibear, R. Schwarz, J. Tugwood, J. Pierce, C. Dive, G. Brady, D. G. Rothwell, F. Chemi, E. Kilgour, F. Blackhall, L. Priest, M. G. Krebs, P. Crosbie, A. Nakas, S. Rathinam, L. Nelson, K. Ryanna, M. Tuffail, A. Bajaj, J. Brozik, F. Morgan, M. Kornaszewska, R. Attanoos, H. Adams, H. Davies, M. Carter, C. R. Lindsay, F. Gomes, Z. Szallasi, I. Csabai, M. Diossy, H. Aerts, A. Kirk, M. Asif, J. Butler, R. Bilanca, N. Kostoulas, M. MacKenzie, M. Wilcox, S. Busacca, A. Dawson, M. R. Lovett, M. Shackcloth, S. Feeney, J. Asante-Siaw, J. Gosney, A. Leek, N. Totten, J. D. Hodgkinson, R. Waddington, J. Rogan, K. Moore, W.

Monteiro, H. Marshall, K. G. Blyth, C. Dick, A. Kidd, E. Lim, P. De Sousa, S. Jordan, A. Rice, H. Raubenheimer, H. Bhayani, M. Hamilton, L. Ambrose, A. Devaraj, H. Chavan, S. Begum, A. Mani, D. Kaniu, M. Malima, S. Booth, A. G. Nicholson, N. Fernandes, J. E. Wallen, P. Shah, S. Danson, J. Bury, J. Edwards, J. Hill, S. Matthews, Y. Kitsanta, J. Rao, S. Tenconi, L. Socci, K. Suvarna, F. Kibutu, P. Fisher, R. Young, J. Barker, F. Taylor, K. Lloyd, T. Light, T. Horey, D. Papadatos-Pastos, P. Russell, S. Lock, K. Gilbert, D. Lawrence, M. Hayward, N. Panagiotopoulos, R. George, D. Patrini, M. Falzon, E. Borg, R. Khiroya, A. Ahmed, M. Taylor, J. Choudhary, P. Shaw, S. M. Janes, M. Forster, T. Ahmad, S. M. Lee, J. Herrero, D. Carnell, R. Mendes, J. George, N. Navani, M. Scarci, E. Bertoja, R. C. Stephens, E. M. Hoogenboom, J. W. Holding, S. Bandula, B. Naidu, G. Langman, A. Robinson, H. Bancroft, A. Kerr, S. Kadiri, C. Ferris, G. Middleton, M. Djearaman, A. Patel, C. Ottensmeier, S. Chee, B. Johnson, A. Alzetani, E. Shaw, J. Lester, Y. Summers, R. Califano, P. Taylor, R. Shah, P. Krysiak, K. Rammohan, E. Fontaine, R. Booton, M. Evison, S. Moss, J. Novasio, L. Joseph, P. Bishop, A. Chaturvedi, H. Doran, F. Granato, V. Joshi, E. Smith, A. Montero, C. Swanton and Y. Yuan, "Geospatial immune variability illuminates differential evolution of lung adenocarcinoma," *Nature Medicine,* vol. 26, no. 7, 2020.

[31] I. Vitale, E. Shema, S. Loi and L. Galluzzi, "Intratumoral heterogeneity in cancer progression and response to immunotherapy," *Nature Medicine,* vol. 27, no. 2, 2021.

[32] A. T. Feldman and D. Wolfe, "Tissue processing and hematoxylin and eosin staining," *Methods in Molecular Biology,* vol. 1180, 2014.

[33] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine,* vol. 25, no. 8, 2019.

[34] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev and J. Yosinski, "An intriguing failing of convolutional neural networks and the CoordConv solution," in *Advances in Neural Information Processing Systems*, 2018.

[35] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D. M. Berney, D. G. Bostwick, A. J. Evans, D. J. Grignon, P. A. Humphrey, K. A. Iczkowski, J. G. Kench, G. Kristiansen, T. H. van der Kwast, K. R. Leite, J. K. McKenney, J. Oxley, C. C. Pan, H. Samaratunga, J. R. Srigley, H. Takahashi, T. Tsuzuki, M. Varma, M. Zhou, J. Lindberg, C. Lindskog, P. Ruusuvuori, C. Wählby, H. Grönberg, M. Rantalainen, L. Egevad and M. Eklund, "Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study," *The Lancet Oncology,* vol. 21, no. 2, 2020.

[36] M. Y. Lu, T. Y. Chen, D. F. Williamson, M. Zhao, M. Shady, J. Lipkova and F. Mahmood, "AI-based pathology predicts origins for cancers of unknown primary," *Nature,* vol. 594, no. 7861, 2021.

[37] P. J. Schüffler, L. Geneslaw, D. V. K. Yarlagadda, M. G. Hanna, J. Samboy, E. Stamelos, C. Vanderbilt, J. Philip, M. H. Jean, L. Corsale, A. Manzo, N. H. Paramasivam, J. S. Ziegler, J. Gao, J. C. Perin, Y. S. Kim, U. K. Bhanot, M. H. Roehrl, O. Ardon, S. Chiang, D. D. Giri, C. S. Sigel, L. K. Tan, M. Murray, C. Virgo, C. England, Y. Yagi, S. J. Sirintrapun, D. Klimstra, M. Hameed, V. E. Reuter and T. J. Fuchs, "Integrated digital pathology at scale: A solution for clinical diagnostics and cancer research at a large academic medical center," *Journal of the American Medical Informatics Association,* vol. 28,

no. 9, 2021.

[38] M. G. Hanna, V. E. Reuter, M. R. Hameed, L. K. Tan, S. Chiang, C. Sigel, T. Hollmann, D. Giri, J. Samboy, C. Moradel, A. Rosado, J. R. Otilano, C. England, L. Corsale, E. Stamelos, Y. Yagi, P. J. Schüffler, T. Fuchs, D. S. Klimstra and S. J. Sirintrapun, "Whole slide imaging equivalency and efficiency study: experience at a large academic center," *Modern Pathology,* vol. 32, no. 7, 2019.

[39] J. E. L. Matthias Fey, "Fast Graph Representation Learning with PyTorch Geometric," *arXiv.org,* 2019.

[40] K. Xu, S. Jegelka, W. Hu and J. Leskovec, "How powerful are graph neural networks?," *7th International Conference on Learning Representations, ICLR 2019,* 2019.

[41] B. Knyazev, G. W. Taylor and M. R. Amer, "Understanding attention and generalization in graph neural networks," *Advances in Neural Information Processing Systems,* vol. 32, 2019.

Acknowledgement

Funding/Support. This work was supported by the Warren Alpert Foundation Center for Digital and Computational Pathology at Memorial Sloan Kettering Cancer Center, the NIH/NCI Cancer Center Support Grant P30 CA008748, Weill Cornell Graduate School of Medical Sciences and the Tri-I Computational Biology and Medicine Program. Matthew D. Hellmann is a Damon Runyon Clinical Investigator (including Continuation Award) supported in part by the Damon Runyon Cancer Research Foundation grant no. CI-98-18.

Access to data and code. This paper is dependent on the code of two previous works [22, 25]. Code for minimal implementation of these methods will be made public for non-commercial purpose.