Patterns, Volume *3*

Supplemental information

Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study

Mustafa Umit Oner, Jianbin Chen, Egor Revkov, Anne James, Seow Ye Heng, Arife Neslihan Kaya, Jacob Josiah Santiago Alvarez, Angela Takano, Xin Min Cheng, Tony Kiat Hon Lim, Daniel Shao Weng Tan, Weiwei Zhai, Anders Jacobsen Skanderup, Wing-Kin Sung, and Hwee Kuan Lee

SUPPLEMENTAL ITEMS

Table S1: The number of samples, slides, and patches in each TCGA cohort. Each patient has only one tumor sample and one normal sample if available. Note that "tumor slide" and "normal slides" refer to the slides of tumor samples and normal samples, respectively. Similarly, "tumor patches" and "normal patches" refer to patches cropped over "tumor slides" and "normal slides", respectively. Related to Table 1.

	# samples			# slides			# patches		
	normal	tumor	total	normal	tumor	total	normal	tumor	total
BRCA	133	929	1,062	312	1,280	1,592	84,196	710,446	794,642
GBM	0	474	474	0	917	917	0	618,649	618,649
KIRC	364	435	799	454	841	1295	466,883	655,625	1,122,508
LGG	0	454	454	0	625	625	0	347,065	347,065
LUAD	171	446	617	200	694	894	108,876	490,401	599,277
LUSC	220	453	673	333	714	1,047	166,181	544,778	710,959
OV	84	516	600	142	1,031	1,173	72,385	1,122,620	1,195,005
PRAD	111	428	539	111	535	646	75,798	338,120	413,918
THCA	83	428	511	83	443	526	30,234	199,275	229,509
UCEC	32	449	481	34	589	623	17,359	314,624	331,983

Table S2: The number of samples in different genomic tumor purity and percent tumor nuclei groups (<10% , 10-25%, 25-50%, and $\geq$50%). Related to Table 1.

	genomic tumor purity				percent tumor nuclei			
	<10%	10-25%	25-50%	$\geq$50%	<10%	10-25%	25-50%	$\geq$50%
BRCA	1	44	247	637	0	0	10	919
GBM	0	6	43	425	1	0	4	469
KIRC	0	7	158	270	0	0	0	435
LGG	0	10	54	390	0	0	3	451
LUAD	1	40	225	180	0	0	5	441
LUSC	0	40	188	225	0	0	5	448
OV	0	1	28	487	0	0	0	516
PRAD	0	24	117	287	0	0	7	421
THCA	0	3	50	375	0	0	0	428
UCEC	0	6	43	400	0	0	6	443

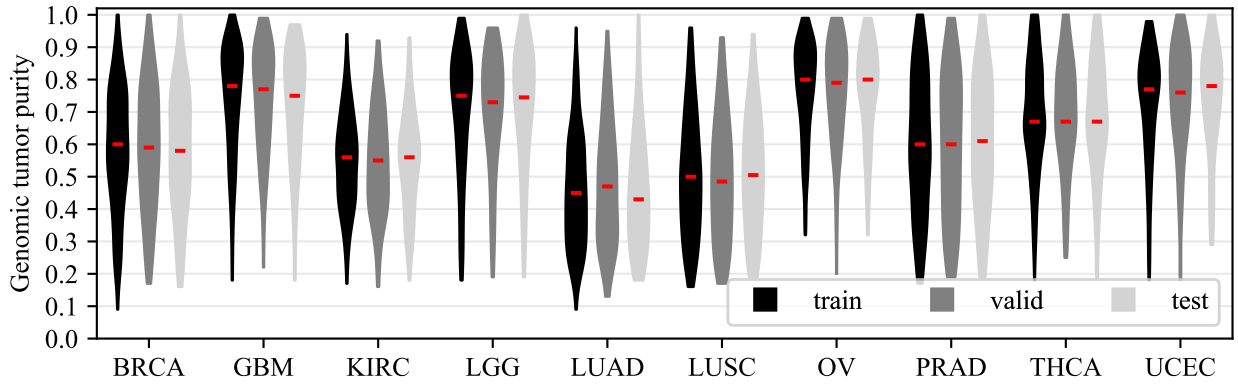


Figure S1: Violin plots of genomic tumor purity values (obtained using ABSOLUTE[1]) in the training, validation, and test sets of each TCGA cohort. Related to Table 1.

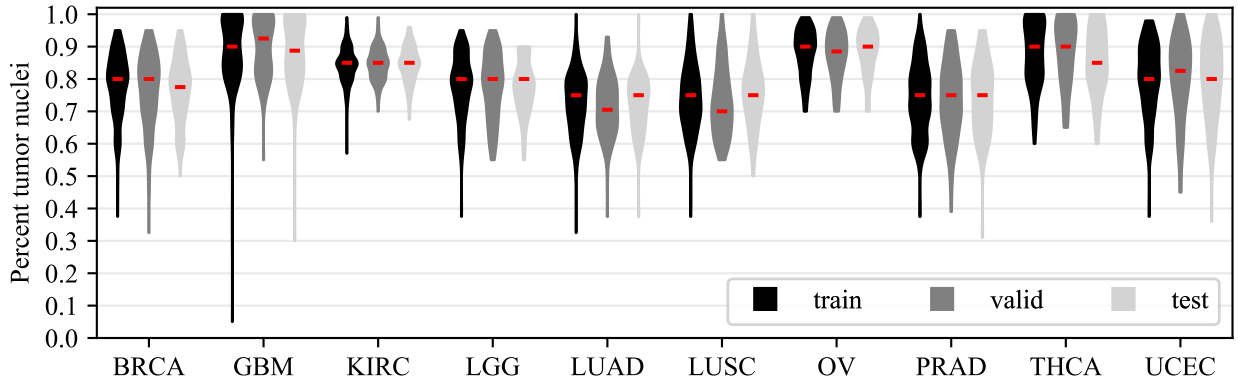


Figure S2: Violin plots of percent tumor nuclei values (collected from TCGA data portal) in each TCGA cohort's training, validation, and test sets. Related to Table 1.

Table S3: Comparison of methods based on Spearman's correlation coefficients in the test sets of different cohorts. Spearman's correlation coefficients between genomic tumor purity values and MIL predictions (ρ_{mil}) and genomic tumor purity values and pathologists' percent tumor nuclei estimates (ρ_{path}) in the test sets of different cohorts are calculated for only the tumor samples. Then, they are compared using the method in Meng et al.[2]. The number of tumor samples (n), Spearman's correlation coefficients together with calculated p-values ($P_{\rho_{mil}}$ and $P_{\rho_{path}}$) and 95% confidence intervals ($CI_{\rho_{mil}}$ and $CI_{\rho_{path}}$), and calculated p-values in statistical tests (P_{comp}) are presented. Note that if the calculated correlation in any method is not significant (i.e., $P_{\rho_{mil}} > 5.0e-02$ or $P_{\rho_{path}} > 5.0e-02$), the statistical test is not conducted. It is indicated by 'x'. The best methods are highlighted in bold. Related to Figure 2 and Figure 3A.

		MIL prediction			Pathologist's estimate			Comparison
	n	ρ_{mil}	$P_{\rho_{mil}}$	$CI_{\rho_{mil}}$	ρ_{path}	$P_{\rho_{path}}$	$CI_{\rho_{path}}$	P_{comp}
BRCA	185	**0.655**	**4.6e-24**	**0.547 - 0.743**	0.299	3.6e-05	0.162 - 0.429	**1.4e-07**
GBM	94	0.572	1.7e-09	0.389 - 0.721	0.104	3.2e-01	-0.102 - 0.309	x
LGG	90	0.418	4.1e-05	0.226 - 0.574	0.201	5.7e-02	-0.029 - 0.392	x
LUAD	90	**0.515**	**2.1e-07**	**0.320 - 0.660**	0.255	1.5e-02	0.036 - 0.448	**1.2e-02**
LUSC	90	0.467	3.5e-06	0.280 - 0.627	0.324	1.8e-03	0.118 - 0.503	1.7e-01
OV	103	**0.581**	**1.3e-10**	**0.429 - 0.711**	0.328	7.1e-04	0.132 - 0.518	**9.4e-03**
PRAD	85	0.424	5.3e-05	0.224 - 0.597	0.293	6.5e-03	0.074 - 0.504	2.0e-01
UCEC	89	**0.579**	**2.7e-09**	**0.408 - 0.720**	0.344	9.8e-04	0.139 - 0.531	**2.6e-02**

Table S4: Spearman's correlation coefficients. Spearman's correlation coefficients between (i) genomic tumor purity values from ABSOLUTE[1] (ABS) and MIL predictions (MIL), (ii) genomic tumor purity values from ESTIMATE[3] (EST) and MIL predictions, and (iii) genomic tumor purity values from ABSOLUTE and genomic tumor purity values from ESTIMATE are calculated for the tumor samples having corresponding values in the test sets. The number of tumor samples (n), correlation coefficients (ρ) together with calculated p-values (P) and 95% confidence intervals (CI) are presented.

		ABS vs. MIL			EST vs. MIL			EST vs ABS		
	n	ρ	P	CI	ρ	P	CI	ρ	P	CI
BRCA	186	0.655	4.6e-24	0.547 - 0.743	0.519	4.0e-14	0.401 - 0.615	0.611	2.4e-20	0.496 - 0.709
GBM	22	0.610	3.3e-03	0.162 - 0.882	0.528	1.4e-02	0.112 - 0.821	0.732	1.6e-04	0.439 - 0.898
LGG	91	0.418	4.1e-05	0.226 - 0.574	0.139	1.9e-01	-0.076 - 0.333	0.352	6.6e-04	0.142 - 0.531
LUAD	91	0.515	2.1e-07	0.320 - 0.660	0.546	2.5e-08	0.391 - 0.674	0.645	6.7e-12	0.468 - 0.779
LUSC	88	0.447	1.4e-05	0.264 - 0.611	0.350	8.9e-04	0.150 - 0.524	0.628	7.5e-11	0.466 - 0.752
OV	52	0.596	3.9e-06	0.360 - 0.768	0.579	8.5e-06	0.323 - 0.763	0.708	6.2e-09	0.532 - 0.824
PRAD	86	0.424	5.3e-05	0.224 - 0.597	0.319	3.0e-03	0.109 - 0.496	0.447	1.8e-05	0.241 - 0.634
UCEC	40	0.574	1.3e-04	0.284 - 0.788	0.400	1.2e-02	0.057 - 0.695	0.580	1.1e-04	0.291 - 0.789

Table S5: Comparison of methods based on absolute errors in the test sets of different cohorts. Absolute errors between genomic tumor purity values and MIL predictions (e_{mil}) and genomic tumor purity values and pathologists' percent tumor nuclei estimates (e_{path}) in the test sets of different cohorts are calculated for only the tumor samples. Then, they are compared using the Wilcoxon signed-rank test[4]. The number of tumor samples (n), mean absolute errors ($\mu_{e_{mil}}$ and $\mu_{e_{path}}$) together with standard deviations ($\sigma_{e_{mil}}$ and $\sigma_{e_{path}}$), median absolute errors ($m_{e_{mil}}$ and $m_{e_{path}}$) together with interquartile ranges ($IQR_{e_{mil}}$ and $IQR_{e_{path}}$), and calculated p-values in the statistical tests (P_{comp}) are presented. The best methods are highlighted in bold. Related to Figure 2 and Figure 3A.

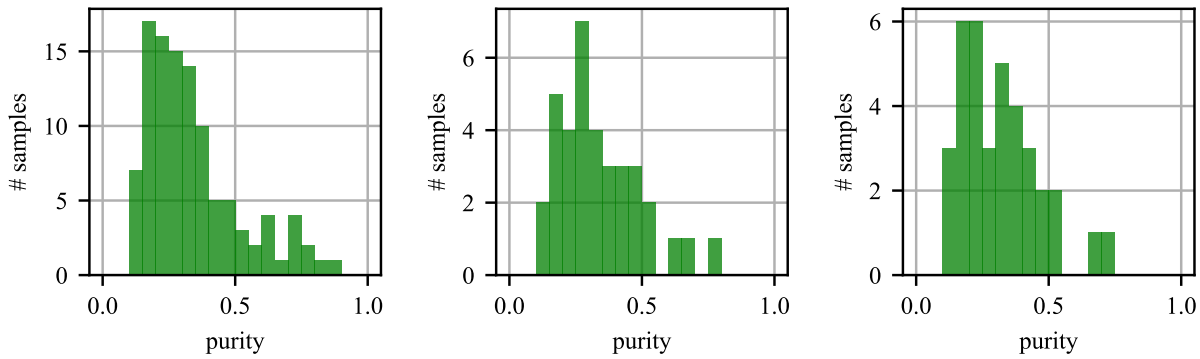
		MIL prediction				Pathologist's estimate				Comp.
	n	$\mu_{e_{mil}}$	$\sigma_{e_{mil}}$	$m_{e_{mil}}$	$IQR_{e_{mil}}$	$\mu_{e_{path}}$	$\sigma_{e_{path}}$	$m_{e_{path}}$	$IQR_{e_{path}}$	P_{comp}
BRCA	185	**0.116**	**0.097**	**0.104**	**0.043 - 0.159**	0.220	0.147	0.200	0.105 - 0.310	**2.5e-13**
GBM	94	**0.113**	**0.106**	**0.074**	**0.046 - 0.142**	0.195	0.158	0.145	0.080 - 0.260	**2.1e-07**
LGG	90	0.136	0.119	0.105	0.052 - 0.188	0.152	0.122	0.130	0.060 - 0.200	5.4e-02
LUAD	90	**0.132**	**0.109**	**0.112**	**0.060 - 0.175**	0.280	0.151	0.275	0.170 - 0.395	**3.9e-09**
LUSC	90	**0.148**	**0.122**	**0.125**	**0.054 - 0.196**	0.266	0.150	0.250	0.140 - 0.375	**5.8e-06**
OV	103	**0.105**	**0.091**	**0.086**	**0.043 - 0.127**	0.136	0.126	0.110	0.030 - 0.190	**1.6e-02**
PRAD	85	**0.173**	**0.154**	**0.130**	**0.068 - 0.240**	0.204	0.141	0.180	0.090 - 0.285	**1.4e-02**
UCEC	89	**0.109**	**0.120**	**0.072**	**0.027 - 0.142**	0.132	0.124	0.100	0.040 - 0.170	**1.4e-02**

Note S1: Singapore Cohort

Singapore cohort consists of 179 lung adenocarcinoma patients having East Asian ancestry. Each patient has one tumor sample, and one slide is prepared from each tumor sample (except one sample in the training set). The slides are prepared from formalin-fixed paraffin-embedded sections (FFPE). On the contrary to FFPE sections in the Singapore cohort, slides in the TCGA cohorts are prepared from fresh-frozen sections. These two tissue preservation methods are quite different from each other. While the FFPE method preserves morphology better and is the routine in histopathology, the fresh-frozen method preserves nucleic acids better and is preferred for molecular analysis[5]. The number of samples, slides and patches in the training, validation and test sets of the Singapore cohort are presented below.

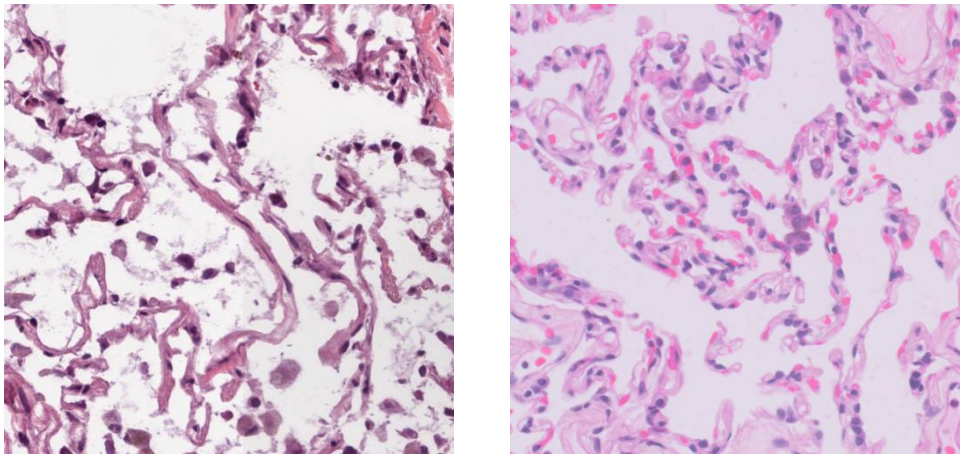
Singapore cohort: the number of samples, slides, and patches. Note that each patient has only one tumor sample. Related to Table 1.

	# samples			# slides			# patches		
dataset	normal	tumor	total	normal	tumor	total	normal	tumor	total
training	0	107	107	0	108	108	0	525,961	525,961
validation	0	36	36	0	36	36	0	190,971	190,971
test	0	36	36	0	36	36	0	182,383	182,383
all	0	179	179	0	180	180	0	899,315	899,315



(a) Training **(b)** Validation **(c)** Test

Figure S3: Singapore cohort: genomic tumor purity histograms for (a) training, (b) validation, and (c) test sets. Related to Table 1.



(a) TCGA LUAD: Fresh-frozen - Normal

(b) Singapore LUAD: FFPE - Normal

(c) TCGA LUAD: Fresh-frozen - Cancerous

(d) Singapore LUAD: FFPE - Cancerous

Figure S4: Example patches cropped from slides of fresh-frozen and formalin-fixed paraffin-embedded (FFPE) sections. (**a, c**) A normal patch and a cancerous patch cropped from slides of fresh-frozen sections in the TCGA LUAD cohort. (**b, d**) A normal patch and a cancerous patch cropped from slides of FFPE sections in the Singapore LUAD cohort. Related to Figure 4.

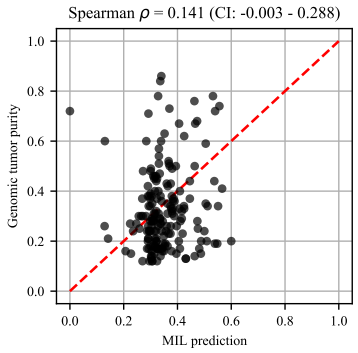


Figure S5: External validation on Singapore cohort. We checked the performance of the TCGA LUAD model directly on the Singapore LUAD cohort (with n=179 tumor samples) used as an external validation set. Scatter plot of genomic tumor purity vs. MIL model prediction. Diagonal red dotted line shows the y=x line.

Table S6: Statistics of the absolute difference between the predictions of a tumor sample's top and bottom slides. In the test set of each cohort, for a tumor sample with two slides, the absolute difference (d_{abs}) between the tumor purity predictions of the slides is calculated. Then, the number of tumor samples with two slides (n), the mean absolute difference ($\mu_{d_{abs}}$), the standard deviation of the absolute difference ($\sigma_{d_{abs}}$), the median absolute difference ($m_{d_{abs}}$), and the interquartile range ($IQR_{d_{abs}}$) are presented. Related to Figure 3C.

	n	$\mu_{d_{abs}}$	$\sigma_{d_{abs}}$	$m_{d_{abs}}$	$IQR_{d_{abs}}$
BRCA	73	0.101	0.106	0.063	0.031 - 0.115
GBM	90	0.090	0.083	0.068	0.016 - 0.141
LGG	31	0.086	0.089	0.054	0.023 - 0.139
LUAD	44	0.100	0.110	0.059	0.023 - 0.125
LUSC	52	0.106	0.123	0.062	0.030 - 0.144
OV	102	0.125	0.156	0.080	0.032 - 0.150
PRAD	21	0.144	0.189	0.086	0.027 - 0.134
UCEC	23	0.063	0.056	0.042	0.021 - 0.089

Table S7: Comparing the absolute errors of sample-level predictions and the expected value of the absolute errors of slide-level predictions in the test sets of different cohorts. In the test set of each cohort, for a tumor sample with two slides, the absolute errors between genomic tumor purity values and sample-level MIL predictions (e_{smpl}) and the expected value of absolute errors between genomic tumor purity values and slide-level MIL predictions (e_{sld}) are calculated. Then, the number of samples with two slides (n), the mean absolute errors ($\mu_{e_{smpl}}$ and $\mu_{e_{sld}}$) together with standard deviations ($\sigma_{e_{smpl}}$ and $\sigma_{e_{sld}}$), the median absolute errors ($m_{e_{smpl}}$ and $m_{e_{sld}}$) together with interquartile ranges ($IQR_{e_{smpl}}$ and $IQR_{e_{sld}}$), and the calculated p-values in the statistical tests (P_{comp}) are presented. Note that the PRAD (n=21) and UCEC (n=23) cohorts were excluded from this study due to few samples with two slides. The best methods are highlighted in bold. Related to Figure 3D.

		Sample level				Slide level				
	n	$\mu_{e_{smpl}}$	$\sigma_{e_{smpl}}$	$m_{e_{smpl}}$	$IQR_{e_{smpl}}$	$\mu_{e_{sld}}$	$\sigma_{e_{sld}}$	$m_{e_{sld}}$	$IQR_{e_{sld}}$	P_{comp}
BRCA	73	**0.114**	**0.082**	**0.092**	**0.043 - 0.166**	0.126	0.073	0.129	0.060 - 0.171	**2.8e-03**
GBM	90	0.115	0.107	0.076	0.046 - 0.145	0.118	0.096	0.089	0.062 - 0.161	7.1e-01
LGG	31	0.178	0.149	0.146	0.100 - 0.218	0.168	0.152	0.106	0.067 - 0.198	5.6e-01
LUAD	44	**0.118**	**0.102**	**0.084**	**0.050 - 0.168**	0.138	0.102	0.121	0.067 - 0.181	**3.7e-04**
LUSC	52	**0.124**	**0.092**	**0.109**	**0.039 - 0.168**	0.150	0.096	0.143	0.085 - 0.201	**1.7e-03**
OV	102	**0.106**	**0.091**	**0.086**	**0.043 - 0.128**	0.135	0.100	0.105	0.073 - 0.176	**5.0e-03**

Table S8: Spearman's correlation coefficients between absolute errors in MIL predictions and percent necrosis values (ρ) are calculated in the test set of each cohort. The number of samples (n), correlation coefficients together with calculated p-values (P) and 95% confidence intervals (95% CI) are presented for tumor samples only. There is no significant correlation (P>0.05) in any cohorts except LUSC, in which the correlation is 0.253 (P=1.6e-02 < 0.05). The LGG cohort is excluded from analysis since all samples have percent necrosis of 0.

	n	ρ	95% CI	P
BRCA	185	0.089	-0.054, 0.236	2.3e-01
GBM	94	-0.040	-0.232, 0.150	7.0e-01
LUAD	90	0.034	-0.187, 0.267	7.5e-01
LUSC	90	0.253	0.062, 0.432	1.6e-02
OV	103	0.044	-0.157, 0.236	6.6e-01
PRAD	85	-0.050	-0.262, 0.170	6.5e-01
UCEC	89	-0.023	-0.230, 0.187	8.3e-01

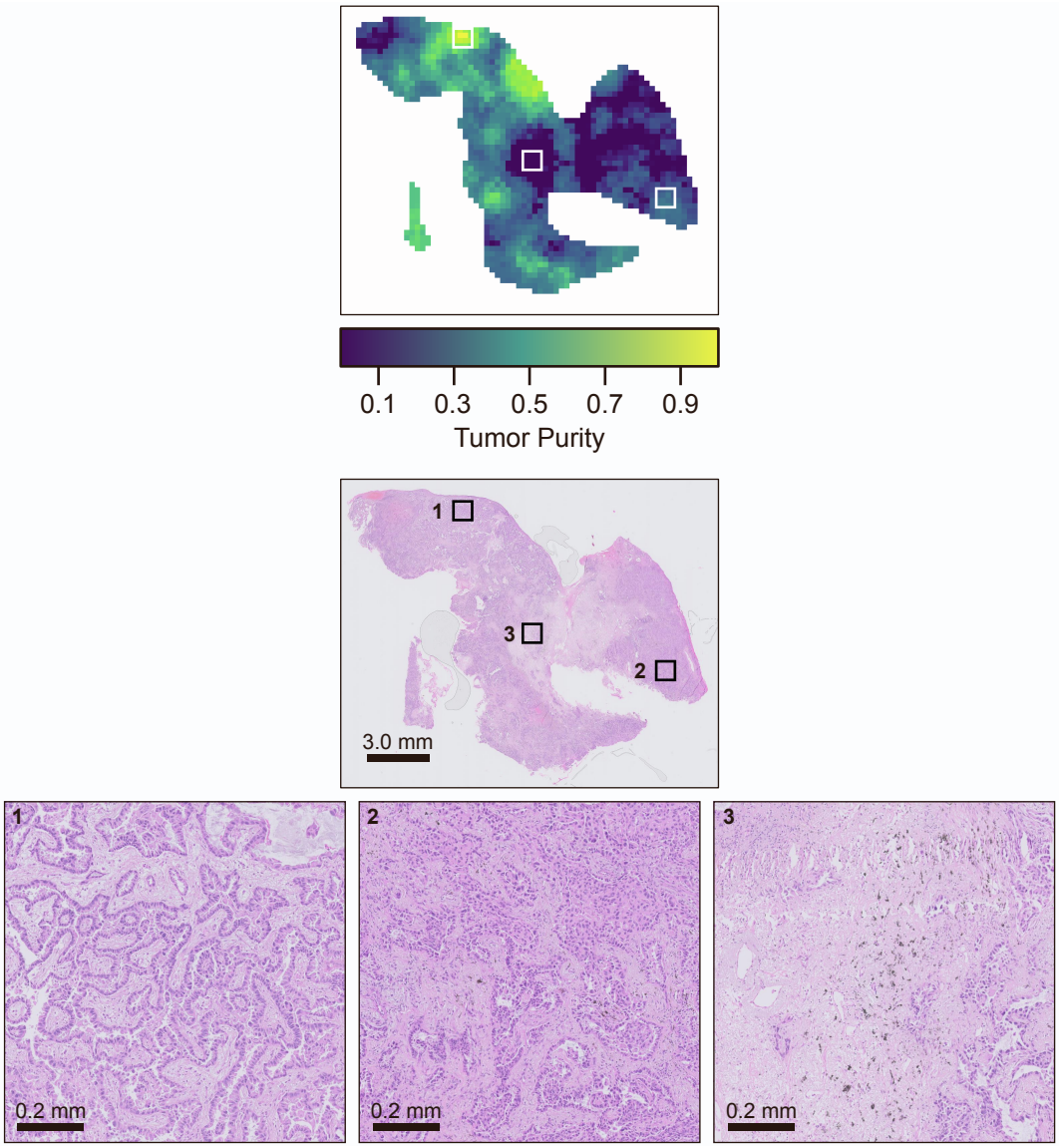


Figure S6: Tumor purity map for A186 in the Singapore Cohort. Genomic tumor purity was 0.340 and our MIL model predicted tumor purity as 0.339, so the absolute error was 0.001. Related to Figure 4.

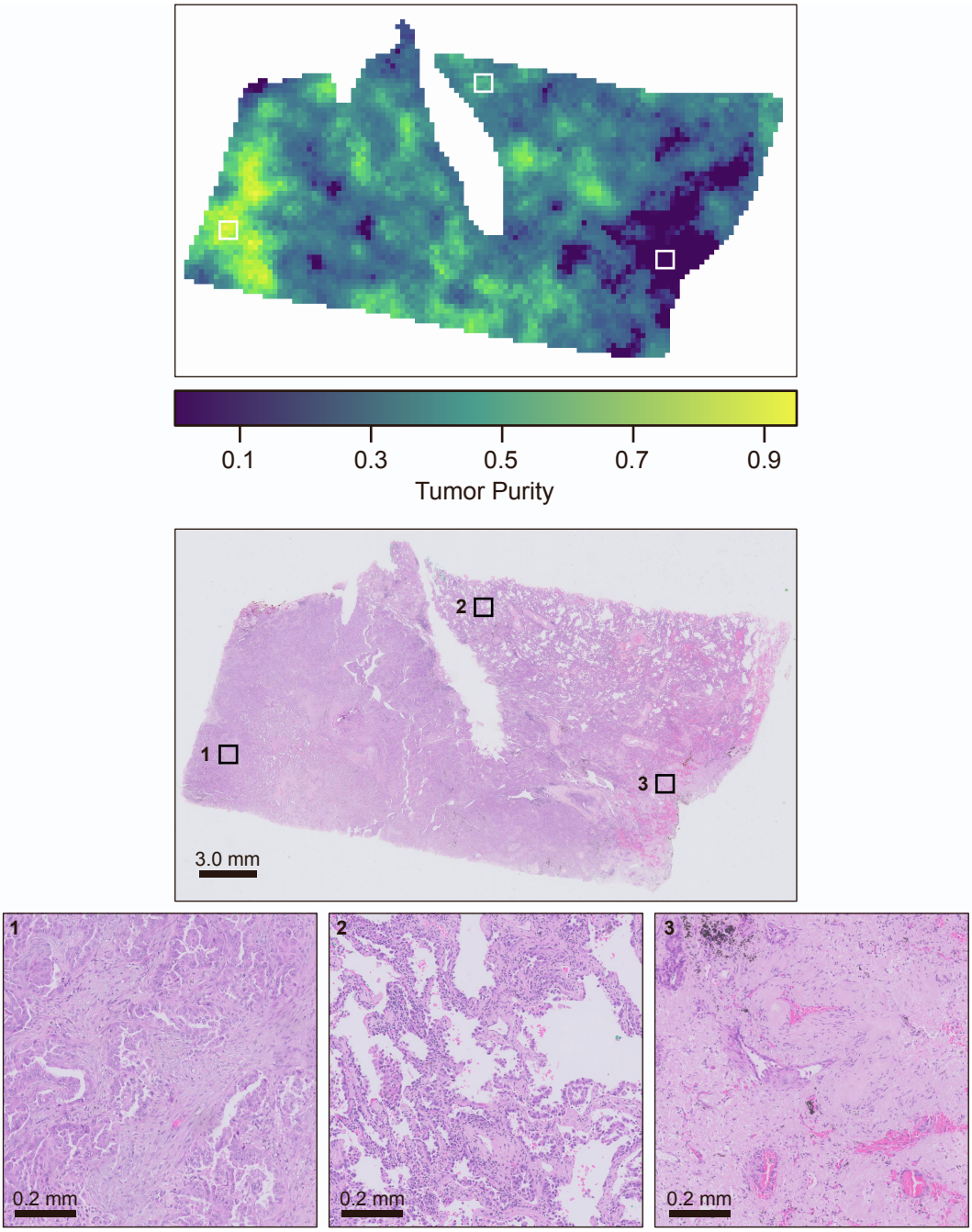


Figure S7: Tumor purity map for A537 in the Singapore Cohort. Genomic tumor purity was 0.420 and our MIL model predicted tumor purity as 0.380, so the absolute error was 0.04. Related to Figure 4.

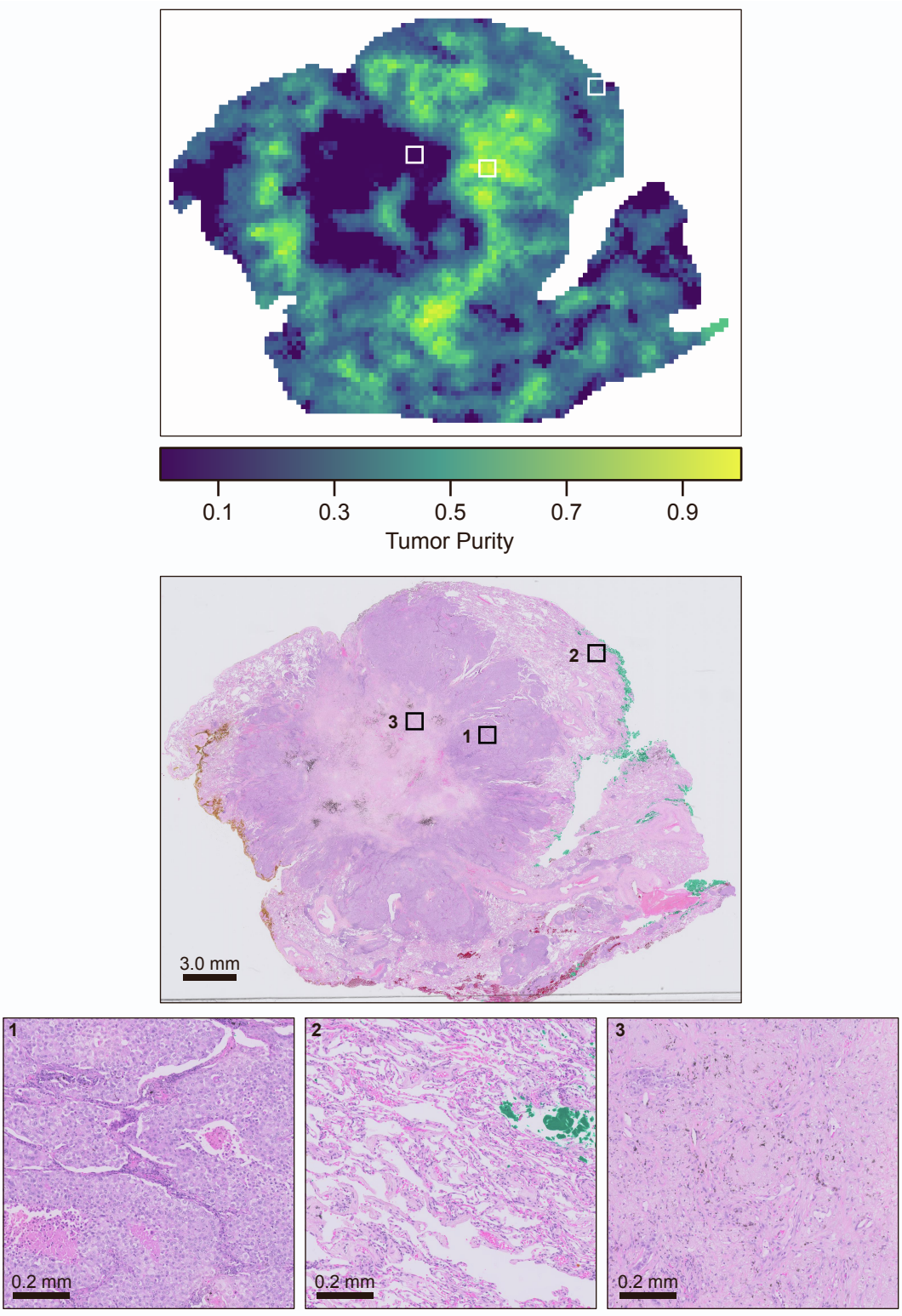


Figure S8: Tumor purity map for A143 in the Singapore Cohort. Genomic tumor purity was 0.240 and our MIL model predicted tumor purity as 0.339, so the absolute error was 0.099. Related to Figure 4.

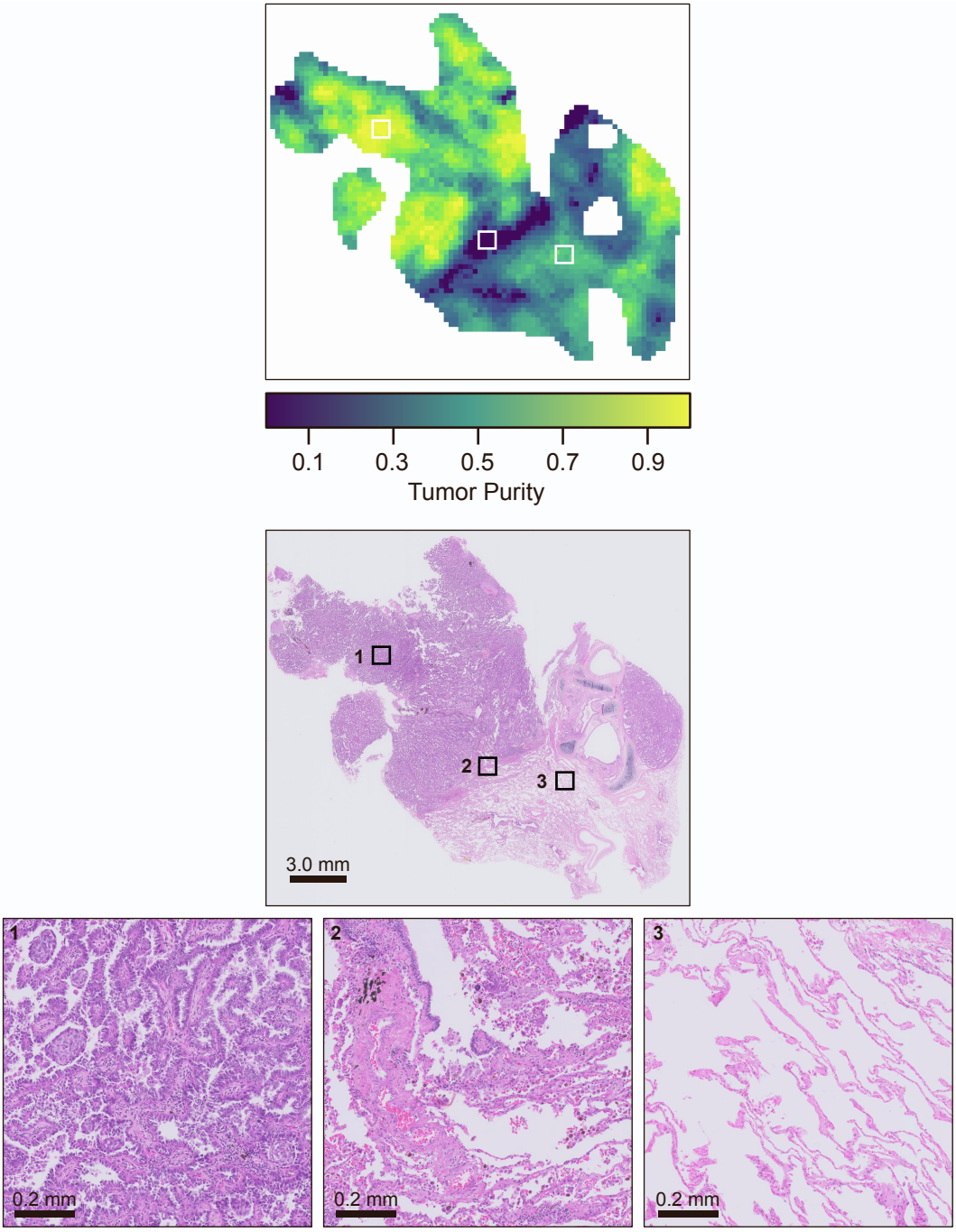


Figure S9: Tumor purity map for A219 in the Singapore Cohort. Genomic tumor purity was 0.410 and our MIL model predicted tumor purity as 0.584, so the absolute error was 0.174. Related to Figure 4.

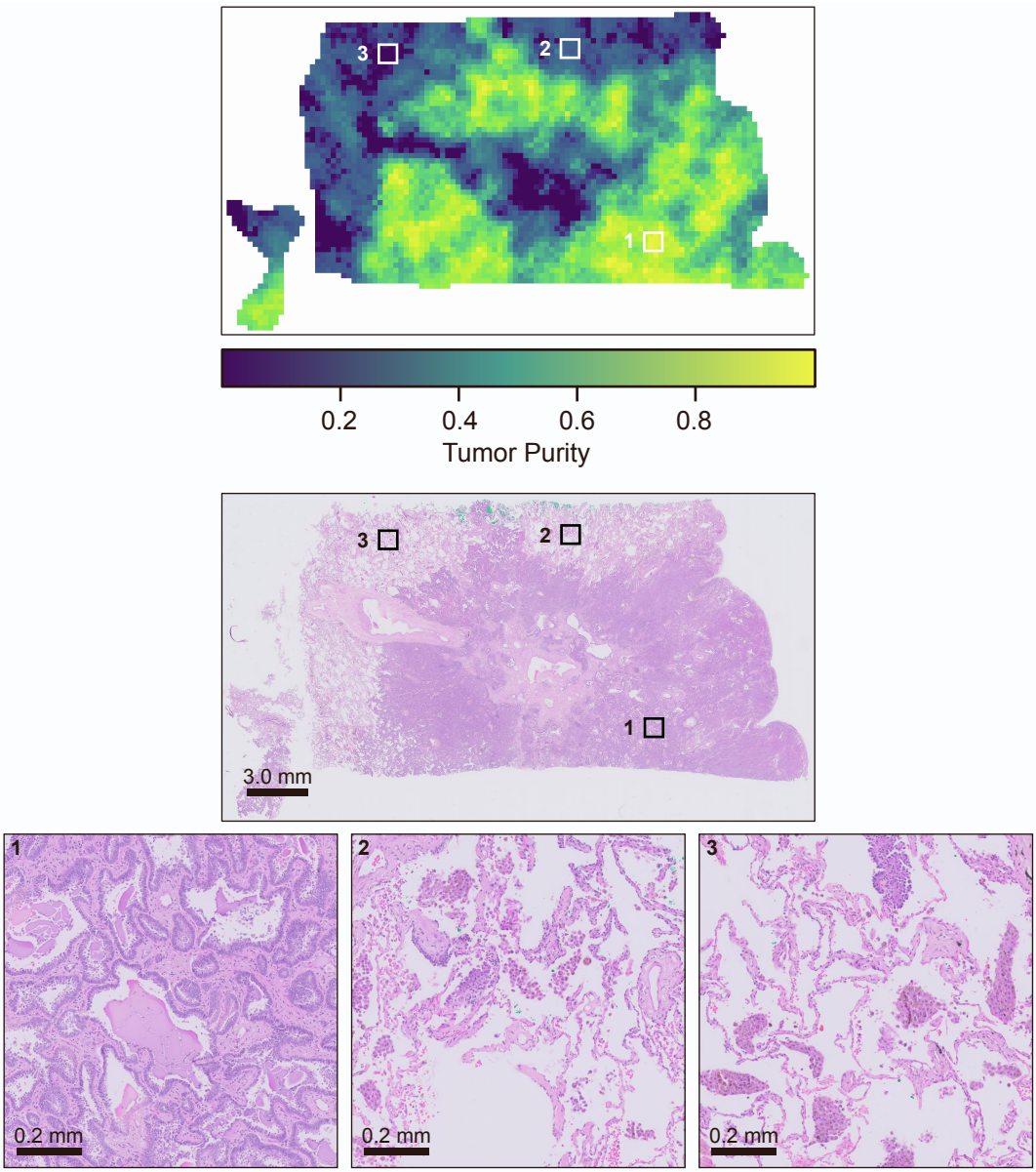


Figure S10: Tumor purity map for A126 in the Singapore Cohort. Genomic tumor purity was 0.160 and our MIL model predicted tumor purity as 0.527, so the absolute error was 0.367. Related to Figure 4.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

MIL Framework

Problem formulation and notation

Let $\mathcal{D}$ be a MIL dataset such that for each $(X, Y) \in \mathcal{D}$, $X = \{x_1, x_2, \cdots, x_N\} \subseteq \mathcal{I}$ and $Y \in \mathcal{Y}$, where $\mathcal{I}$ is the instance space, and $\mathcal{Y}$ is the bag label space. Note that we fix the number of instances in a bag to N for clarity of notation, yet our formulation is also valid for bags with the variable number of instances.

Given any pair $(X, Y) \in \mathcal{D}$, our objective is to predict bag label Y for a given bag of instances X. Here, a bag label Y is the genomic tumor purity of a sample, and a bag X is a collection of cropped patches over the sample's slides. Let $\hat{Y}$ be the predicted bag label of X. To obtain $\hat{Y}$, we designed a novel MIL framework consisting of three stages.

The first stage is a *feature extractor* module $\theta_{\text{feature}} : \mathcal{I} \rightarrow \mathcal{F}$, where $\mathcal{F}$ is the feature space. For each $x_i \in X$, the *feature extractor* module takes x_i as input, extracts J features and outputs a feature vector: $\boldsymbol{f}_{x_i} = \theta_{\text{feature}}(x_i) = [f^1_{x_i}, f^2_{x_i}, \cdots, f^J_{x_i}] \in \mathcal{F}$, where $f^j_{x_i} \in \mathbb{R}$ is the j^{th} feature value and $\mathcal{F} = \mathbb{R}^J$. Let $\boldsymbol{F}_X = [\boldsymbol{f}_{x_1}, \boldsymbol{f}_{x_2}, \cdots, \boldsymbol{f}_{x_N}] \in \mathbb{R}^{JN}$ be feature matrix constructed from extracted feature vectors such that i^{th} column corresponds to $\boldsymbol{f}_{x_i}$.

The second stage is a *MIL pooling filter* module $\theta_{\text{filter}} : \mathbb{R}^{JN} \rightarrow \mathcal{H}$, where $\mathcal{H}$ is the bag-level representation space. The *MIL pooling filter* module takes the feature matrix $\boldsymbol{F}_X$ as input and aggregates the extracted feature vectors to obtain a bag-level representation: $\boldsymbol{h}_X = \theta_{\text{filter}}(\boldsymbol{F}_X) \in \mathcal{H}$.

The last stage is a *bag-level representation transformation* module $\theta_{\text{transform}} : \mathcal{H} \rightarrow \mathcal{Y}$. It transforms the bag-level representation into the predicted bag label: $\hat{Y} = \theta_{\text{transform}}(\boldsymbol{h}_X)$.

We use neural networks to implement θ_{feature} and $\theta_{\text{transform}}$ so that we can fully parameterize the learning process. For θ_{filter}, we use our novel 'distribution' pooling filter. This system of neural networks is end-to-end trainable.

Distribution Pooling Filter

Our previous study[6] defined the family of distribution-based pooling filters as: Given a feature matrix $\boldsymbol{F}_X = [f^j_{x_i} | f^j_{x_i} \in \mathbb{R},\ i = 1, 2, \cdots, N \text{ and } j = 1, 2, \cdots, J]$ obtained from a bag $X = \{x_1, x_2, \cdots, x_N\}$, its bag level representation is obtained by estimating a marginal distribution over each extracted feature. Let $\tilde{p}^j_X : \mathbb{R} \rightarrow \mathbb{R}^+ \cup \{0\}$ be the estimated marginal distribution obtained over j^{th} extracted feature and $\tilde{p}^j_X \in \mathbb{P}$ where $\mathbb{P}$ is the set of all possible marginal distributions. $\tilde{p}^j_X$ is calculated by using kernel density estimation[7], which employs a Gaussian kernel with standard deviation σ, as shown in the Eq. 1. Each instance has two attention based weights, feature weight α_i and kernel weight β_i, obtained from neural network modules. Hence, the bag level representation $\boldsymbol{h}_X = [\tilde{p}^j_X\ | \tilde{p}^j_X \in \mathbb{P}, j = 1, 2, \cdots, J] \in \mathcal{H}$ where $\mathcal{H} = \mathbb{P}^J$. Note that the estimated marginal distributions are uniformly binned during training neural network models for computational purposes.

$$\tilde{p}^j_X(v) = \sum_{i=1}^{N} \beta_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(v - \alpha_i f^j_{x_i}\right)^2} \ \forall_{j=1,2,\cdots,J} \tag{1}$$

Our previous study formally proved that the distribution-based pooling filters are more expressive than the point estimate-based counterparts (like max and mean pooling) regarding the amount of information captured while obtaining bag-level representations[6]. Then, we empirically showed that models with distribution-based pooling filters perform equal or better than that with point estimate-based pooling filters on distinct real-world MIL tasks.

In this study, we used standard deviation of $\sigma = 0.05$ and the estimated marginal distributions were uniformly binned into 21 bins. Note that attention weights in 'distribution' pooling were fixed to $\alpha_i = 1\ \forall_i$ and $\beta_i = \frac{1}{N}\ \forall_i$ where N is the number of instances per bag.

Neural network architectures and hyper-parameters

We used a ResNet18[8] model as the *feature extractor* module and a three-layer multi-layer-perceptron as the *bag-level representation transformation* module.

During the training of the models, we prepared bags on the go. A bag was created by randomly sampling 200 patches (instances) from all available patches previously cropped over a sample's slides. The patch size was 512×512. Data augmentation (random cropping with a size of 299×299 and random horizontal/vertical flipping) was also applied on the patches. We extracted 128 features for each instance inside the bag.

The architecture and list of hyper-parameters used in MIL models are given below.

Neural network architecture and list of hyper-parameters used in the MIL models.

Architecture	input - $299 \times 299 \times 3$ ResNet18 (128 nodes in the last fc layer) 'distribution' pooling Dropout(0.5) fc-384 + ReLU Dropout(0.5) fc-192 + ReLU Dropout(0.5) fc-1 *(regression)*
patch size	512×512
random crop size	299×299
# instances per bag (N)	200
# features (J)	128
# bins in 'distribution' filters	21
σ in Gaussian kernel	0.05
Optimizer	ADAM
Learning rate	$1e-4$
$L2$ regularization weight decay	0.0005
batch size	1

Segmentation of Histopathology Slides in The TCGA LUAD Cohort

In the TCGA LUAD cohort, for each patient with a matching normal sample, we used the trained feature extractor module of our MIL model to extract features of patches cropped over the slides of the tumor and normal samples of the patient. Then, we clustered the patches by using hierarchical clustering over the extracted feature vectors. We determined the distance threshold in hierarchical clustering such that there were 4 clusters among the patches from slides of the normal sample. This made our clustering approach robust against patient-to-patient variations. Indeed, this was why we decided to use both tumor and normal samples of the patient. In other words, instead of determining a global distance threshold for all patients, we calculated patient-specific distance threshold values to capture inter-patient variations.

Each cluster can be assigned one of two labels: cancerous or normal. Ideally, a cluster with a cancerous label can contain patches only from slides of the tumor sample. On the other hand, a cluster with a normal label can contain patches from slides of both the tumor and the normal samples since the tumor sample may also contain normal tissue components. As a post-processing step, we analyzed normal clusters. If the number of patches from slides of the normal sample in a normal cluster was less than 10%, we split this cluster into two such that patches from slides of the tumor sample were assigned to a new cancerous cluster. Finally, we created segmentation masks for slides of the tumor sample by using cluster labels assigned to the patches.

SUPPLEMENTAL REFERENCES

[1] Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., et al. (2012). Absolute quantification of somatic dna alterations in human cancer. Nature biotechnology, 30(5):413–421.

[2] Meng, X.-L., Rosenthal, R., and Rubin, D. B. (1992). Comparing correlated correlation coefficients. Psychological bulletin, 111(1):172.

[3] Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P. W., Levine, D. A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications, 4(1):1–11.

[4] Wilcoxon, F. (1992). Individual comparisons by ranking methods. In Breakthroughs in statistics, pages 196–202. Springer.

[5] Spencer, D. H., Sehn, J. K., Abel, H. J., Watson, M. A., Pfeifer, J. D., and Duncavage, E. J. (2013). Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. The Journal of molecular diagnostics, 15(5):623–633.

[6] Oner, M. U., Kye-Jet, J. M. S., Lee, H. K., and Sung, W.-K. (2020). Studying the effect of mil pooling filters on mil tasks. arXiv preprint arXiv:2006.01561.

[7] Parzen, E. (1962). On estimation of a probability density function and mode. The annals of mathematical statistics, 33(3):1065–1076.

[8] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.