A histomorphological atlas of resected mesothelioma discovered by self-supervised learning from 3446 whole-slide images

Farzaneh Seyedshahi[1,2], Kai Rakovic[1,2,8,+], Nicolas Poulain[2,+], Adalberto Claudio Quiros[1,7], Ian R. Powley[2], Cathy Richards[3], Hussein Uraiby[3], Sonja Klebe[4], Apostolos Nakas[3], Claire Wilson[5], Marco Sereno[5], Leah Officer-Jones[2], Catherine Ficken[2], Ana Teodosio[6], Fiona Ballantyne[2], Daniel Murphy[1,2], Ke Yuan[1,2,7,*], and John Le Quesne[1,2,8,*]

[1]School of Cancer Sciences, University of Glasgow, Glasgow, Scotland, UK
[2]Cancer Research UK Scotland Institute, Glasgow, Scotland, UK
[3]University Hospitals of Leicester, Leicester, UK
[4]Flinders Health and Medical Research Institute, Adelaide, Australia
[5]University of Leicester, Leicester, UK
[6]Birmingham Tissue Analytics, University of Birmingham, Birmingham, UK
[7]School of Computing Science, University of Glasgow, Glasgow, Scotland, UK
[8]Pathology Department, Queen Elizabeth University Hospital, NHS Greater Glasgow and Clyde, Glasgow, Scotland, UK
[+]These authors contributed equally to this work
[*]ke.yuan@glasgow.ac.uk, john.lequesne@glasgow.ac.uk

1 Abstract

Mesothelioma is a highly lethal and poorly biologically understood disease which presents diagnostic challenges due to its morphological complexity. This study uses self-supervised AI (Artificial Intelligence) to map the histomorphological landscape of the disease. The resulting atlas consists of recurrent patterns identified from 3446 Hematoxylin and Eosin (H&E) stained images scanned from resected tumour slides. These patterns generate highly interpretable predictions, achieving state-of-the-art performance with 0.65 concordance index (c-index) for outcomes and 85% AUC in subtyping. Their clinical relevance is endorsed by comprehensive human pathological assessment. Furthermore, we characterise the molecular underpinnings of these diverse, meaningful, predictive patterns. Our approach both improves diagnosis and deepens our understanding of mesothelioma biology, highlighting the power of this self-learning method in clinical applications and scientific discovery.

2 Main

Mesothelioma is a highly lethal cancer almost always caused by asbestos exposure [1, 2]. Early detection, critical for effective treatment, remains challenging [3, 4]. Mesothelioma's biological diversity complicates histopathological diagnosis, as early malignancy can be difficult to distinguish from reactive changes. Diagnosing mesothelioma from H&E (Hematoxylin and Eosin) images is a subjective and time-intensive process even for skilled subspecialty histopathologists. Definitive diagnosis often remains elusive, even with immunohistochemistry or FISH (fluorescence in situ hybridisation). These diagnostic challenges are at least in part due to the difficulties in devising robust, manually applicable systems of morphological characterisation, in addition to the well-known issues of inter-pathologist agreement. The emergence of AI methods provides an opportunity to comprehensively describe the morphological complexity of mesothelioma to generate a quantitative visual dictionary of the disease.

Recent AI methods in mesothelioma primarily focus on tile-based or cell-based approaches, using supervised or weakly supervised learning. MesoNet [5] used Whole Slide Images (WSI) tiles to predict patient survival through a risk score based on malignant morphologies but lacked insight into the diversity of the tumour microenvironment. Cell-based methods like MesoGraph [6] and SpindleMesoNET [7] quantified malignancy through tumour cell shapes, especially spindle cells, but required extensive annotations and were computationally intensive for slide-level applications. While these approaches provide valuable insights, their findings are

restricted by the nature and quality of their human annotations. Recent self-supervised models, such as Hierarchical Image Pyramid Transformer (HIPT)[8], CTransPath [9], HistoSSLscaling[10], UNI[11], and Histomorphological Phenotype Learning(HPL)[12], have been developed for H&E WSI histopathological analysis. Uniquely among these approaches, HPL focuses on identifying recurrent histomorphological patterns through clusters known as histomorphological phenotype clusters (HPCs). HPL leverages the Barlow Twins self-supervised framework [13], using ResNet for feature extraction from 224x224 WSI patches, followed by clustering of tile feature vectors via the Leiden algorithm [14]. Each HPC represents a unique morphological pattern that can be associated with specific molecular landscapes or used to predict patient outcomes and mesothelioma subtypes by quantifying HPC frequencies. (Further details in the Online Methods section.) Previously, HPL has been applied to lung cancer, revealing significant underlying patterns and giving impressive prognostic performance[12], but it has not been implemented in mesothelioma.

Self-supervised methods such as HPL depend upon accessing large volumes of training data, preferably from resected tumour material which offers large tissue areas with full morphological variance. This is especially challenging in mesothelioma, which is so often diagnosed from tiny biopsies and subsequently treated medically rather than surgically. So, in this study, we curated 3446 whole slide images of 485 resected mesothelioma cases to generate a uniquely powerful training resource called Leicester Archival Thoracic Tumour Investigation Cohort-Mesothelioma cohort (LATTICe-M), as shown with further details in Figure 1a. We then applied the HPL pipeline to our dataset to build a comprehensive atlas of mesothelioma H&E morphology. (Figure1b)

3 Results

3.1 Mapping the histomorphological phenotype landscape of mesothelioma

We have identified 47 recurrent histomorphological phenotype clusters (HPCs) (Figure2a) based on morphological features encoded by self-learned neural networks. These HPCs are identified from 3,239,939 tiles extracted from 3446 images at 5x equivalent resolution. 42 of 47 HPCs are shared in more than 20% of cases, and none of them are case-specific. The most frequent pattern in our dataset is HPC 11("talc, pleurodesis, epithelioid infiltration") (Figure 2b)

A team of subspecialty expert pathologists from 3 centres, who had no access to the WSI images or labels (blinded assessment), examined every HPC to achieve consensus morphological annotations for each one, derived from their defining features: epithelioid vs spindled morphology, inflammation, necrosis, cellularity, desmoplasia, atypia, and cluster purity and each HPC was given a summary title. They evaluated inflammation levels in each HPC, categorising them as None-Sparse, Mild-Moderate, or Marked. Most HPCs were None-Sparse, but some displayed notable patterns. (Figure 2c). These annotations reveal areas of the UMAP containing multiple HPCs with broad similarities, such as spindled/collagenous HPCs, epitheloid tumour growth patterns, and lymphocytic infiltration, as well as peripheral and projecting clouds of morphologically highly distinct lung tissue and chest wall muscle tiles (Figure 2d). This grouping is supported by the general co-occurrence of HPCs within each slide in the Supplementary figures. These HPCs enable the detailed automated spatial annotation of any mesothelioma whole slide image, as illustrated in Figure 2e, highlighting two cases with highly divergent outcomes and morphologies. The first case, a sarcomatoid malignancy which resulted in death at 66 days, is highly morphologically diverse and contains abundant tiles in spindle cell-associated clusters, while the second is predominantly made up of a single epithelioid cluster.

3.2 Histomorphological phenotype clusters predict mesothelioma subtypes

The crucial histopathological distinction in mesothelioma subtyping lies between epithelioid and non-epithelioid (i.e. sarcomatoid/biphasic) variants. To classify this, we generated a numerical vector representing the percentage or frequency of each HPC for every WSI. This vector was transformed using the centred log-ratio (clr) transformation to enhance stability and interpretability, then feed into a logistic regression model for classification as either non-epithelioid or epithelioid. Tumours labelled as biphasic and sarcomatoid were combined into a single group, creating a binary classification task. 8 HPCs are significantly associated with the epithelioid subtype (HPCs 14, 39, 24, 25, 27, 40, 8, and 18 with log-rank test$p - value < 0.05$), containing epithelioid malignancy, predominantly characterised by tubular patterns and solid sheets of epithelioid tumour cell growth. Of the 9 HPCs linked to the non-epithelioid subtype, 3 HPCs (15, 16, and 22— mostly containing disorderly spindle cells) are unanimously classified as non-epithelioid malignancy by our pathologists as well. The other 6 HPCs (6, 7, 35, 37, 45, and 28— mostly solid epithelioid growth pattern) contain more diverse appearances, including epithelioid HPCs, pleural plaque and muscle. This might be due to the inclusion of biphasic cases and lethal epithelioid patterns in this group and also suggests a possible link between sarcomatoid growth and invasion into the chest wall. (Figure3a) Our logistic regression classifier achieved 87% 5-fold cross-validated AUC (Area Under the Curve) Score on the LATTICe-M dataset and 76.5% on The Cancer Genome Atlas (TCGA) mesothelioma dataset and robust across varied clustering configurations (Figure 3b). We visualised HPC compositions at the case level using a PCA plot, colour-coding cases by subtype at diagnosis. The transition from epithelioid cases to sarcomatoid cases through biphasic cases is clearly visible.(Figure 3c)

3.3 Histomorphological phenotype clusters as predictors of patient survival

We aggregated HPC frequencies across all samples per patient, summarising each case into a readily interpretable composition of morphologies. The 5-fold cross-validated c-index values for patient prognosis outcomes were 0.67 and 0.65 for the training and test LATTICe-M primary datasets, respectively, and 0.65 for the fully unseen TCGA cohort as an external dataset. The addition of clinical information, including mesothelioma subtype, TNM stage, and age, only modestly improved the ability of the algorithm to predict outcomes, yielding an increment in C-index of 0.01. We further verified that our approach is robust across other clustering configurations. Compared to similar research on mesothelioma outcome prediction using WSIs, such as MesoNet [5], our model achieves at least a comparable c-index score for the same additional dataset (TCGA). While MesoNet reported a score of 0.656 for TCGA, we matched this performance, with c-index scores ranging from 0.64 to 0.7 across different folds, however prioritising model interpretability through our morphology-based HPCs, also using a fully self-supervised pipeline.

We identified HPCs 10 (Log Hazard Ratio=-0.089, p-value=0.001) and 27 (Log Hazard Ratio=-0.062, p-value=0.008) ("epithelioid nests in bland stroma" and "dense lymphocytes") as positive survival factors, while HPCs 15 (Log Hazard Ratio=0.052, p-value=0.026) and 22 (Log Hazard Ratio=0.042, p-value=0.016) ("disorderly spindle cells" and "transitional mesothelioma") emerge as strong predictors of poor outcome. (Figure4a) A comparison of tile-level UMAP plots colour-coded by hazard ratio reveals a high degree of similarity, further underscoring the strong links between sarcomatoid transformation and poor patient outcomes. The map highlights red areas (higher hazard ratio) like HPC 15 and 16 (Sarcomotiod HPCs) and blue areas (lower hazard ratios) like HPC 45, 27 and 5 (either non-tumourous tissue or lymphocyte HPCs) (Figure 4b).

Next, we categorised patients into high- and low-risk groups based on their calculated hazard ratios for 60 months. Kaplan-Meier plots were generated for LATTICe-M train ($p-value = 2.79e-15$) and test ($p-value = 9.96e-15$) datasets, as well as the TCGA-Meso additional cohort ($p-value = 1.38e-3$), as shown in Figure 4c. The model achieves impressive separation, predicting outcomes with surprising power in this extremely poor-prognosis population who face all the complex hazards of radical surgery and impaired respiratory physiology alongside the biology of their tumour burden. Figure 4d shows a comparison between classical histological grading of epithelioid pleural mesothelioma (as suggested in [15]) and our model, both applied to the TCGA mesothelioma just epithelioid sample cases. Our pipeline demonstrates superiority (p-value=0.025) against human grading for patient outcomes in this dataset. Figure 4e shows the SHAP (SHapley Additive exPlanations) [16] decision plots for our Cox model, comparing a high-risk sarcomatoid case (red) and a relatively low-risk epithelioid case (blue). The plot shows how the model assigns high or low-risk labels for these patients based on the abundance/scarcity of influential HPCs, such as the highly lethal HPC 15 ("disorderly spindle cells"), or the protective HPC 27 ("dense lymphocytes"), which both contribute to the calculated risk in these two cases.

We further demonstrate the ability of our model to predict patient outcomes within disease subtypes (epithelioid vs non-epithelioid) groups. HPC frequencies were calculated, and survival was predicted separately for each group, identifying HPCs which underscored subtype-specific traits (Figure 5a). For epithelioid cases, HPC 10 ("epithelioid nests in bland stroma") and HPC 22 ("transitional mesothelioma") emerged as significant predictors of good and bad outcomes, respectively. HPC 10 is likely to represent relatively indolent well-differentiated 'classically' epithelioid disease. Interestingly, HPC 22, which is very enriched for the appearances of transitional mesothelioma, is not uncommon in cases diagnostically subtyped as epithelioid and is strongly predictive of poor outcomes in this group. This supports the view that transitional appearances signal early stages of transition to sarcomatoid growth [17], and its presence in this group highlights the difficulty in human identification of this pattern[18]. For biphasic/sarcomatoid cases, HPC 23 ("bland spindle cells and collagen") predicts poor outcome, perhaps identifying areas of cytologically bland desmoplastic differentiation, while the good prognostic association of HPC 27 ("dense lymphocytes") suggests towards a particular role for the immune system in sarcomatoid disease. We also show example tiles for specific HPC groups of interest based on pathologist annotations, including inflamed clusters, classical desmoplastic appearances, and necrosis (Figure5c).

3.4 Molecular underpinnings of HPCs

We next investigated the biological underpinnings of HPC morphology to further explain our model's predictive capabilities in mesothelioma prognosis and subtyping. This was achieved by quantifying associations between gene expression signatures and HPC composition in the TCGA Mesothelioma RNASeq dataset [19].

3.4.1 Associations with Ki67 expression and tumour microenviroment signatures

We used the MCPcounter algorithm to estimate cell types, including fibroblasts, endothelial cells, T cells, B lineage cells, myeloid dendritic cells, NK cells, and CD8 T cells from RNASeq data. (Figure 6a) Expression of the proliferation marker Ki67 is also mapped, revealing especially high proliferation in spindle cell-enriched HPCs (HPCs 22, 16, 15, and 6). Critically, these HPCs are the most predictive of non-epithelioid subtype (3a). In contrast, HPCs defined by well-differentiated epithelioid disease and normal tissue (e.g., HPCs 17, 3, 10, and 19) show low proliferation.

Method	AUC-ROC	Avg. Precision	Sensitivity	Specificity
max-MIL	0.70 ± 0.01	0.54 ± 0.12	0.54 ± 0.07	73 ± 0.09
naive-MIL	0.84 ± 0.05	0.72 ± 0.11	0.72 ± 0.08	0.71 ± 0.1
PINS	0.85 ± 0.05	0.80 ± 0.07	0.82 ± 0.1	0.71 ± 0.13
CLAM	0.85 ± 0.07	0.74 ± 0.11	0.75 ± 0.11	0.77 ± 0.02
MesoGraph	0.90 ± 0.007	0.86 ± 0.02	0.88 ± 0.015	0.72 ± 0.01
HPL-MIL (Ours)	**0.93 ± 0.12**	**0.92 ± 0.11**	**0.94 ± 0.09**	**0.86 ± 0.19**

Table 1: AUC-ROC, Average Precision, Sensitivity, and Specificity scores for the HPL model performance using MIL approach to generate unique risk/sarcomatoid-ness scores. The St. George's Hospital TMA core dataset was used for benchmarking without any prior training, as reported in [6], through internal cross-validation, demonstrating the robustness of tile vector representations from our self-supervised network trained on mesothelioma WSIs. Additionally, we report HPL-MIL performance on our primary dataset (LATTICe-M) for comparison over different sets, using external (unseen) cross-validation for the subtyping task.

Fibroblast signatures are strongly pronounced in multiple clusters which either significantly determine sarcomatoid disease or contain fibroblastic/collagenous/stroma-rich morphology indicative of fibroblast-like mesenchymal dedifferentiation of mesothelioma cells. Fibroblast signatures are minimal in HPCs representing papillary or micropapillary epithelioid mesothelioma, solid pattern disease, lung tissue, and large-vessel-rich HPCs consistent with more specialised epithelioid/tissue-specific phenotypes.

Lymphocyte-rich HPC 27 shows strong correlations with T cells, B lineage cells, and myeloid dendritic cells. Similarly, inflamed HPCs (HPCs 29, 1, 24, 27), identified by Hover-Net, indicate active immune environments linked to better prognosis and likely improved immunotherapy response. HPCs 1 (inflamed fat) and 27 (dense lymphocytes) are high in both B- and T-cell signatures and show strong inter-correlation, suggesting dense inflammation and tertiary lymphoid structure formation in chest wall fatty tissues, supporting previous observations that tertiary lymphoid structures are related to good outcome.[20]

3.4.2 Pathways and Hallmarks

KEGG pathway correlations across HPCs again show clear separation between non-epithelioid and epithelioid subtypes (Figure 6b), with generally heightened mitogenic signalling pathway activity in a group of sarcomatoid and fibroblastic clusters associated with aggressive biology. In contrast, HPCs linked to epithelioid growth and normal tissues exhibit relative down-regulation.

An analysis of cancer hallmark pathways further identifies the most sarcomatoid HPCs as a group with strong positive links to multiple proliferation-associated pathways (Figure 6c), in addition to mitogenic signalling and multiple EMT-related pathways. Notably, the same group exhibits downregulation of oxidative phosphorylation components, indicating a metabolic shift towards hypoxia.

3.5 Validation on tiny tissue fragments

To assess the generalisability of our self-supervised model trained on the LATTICe-M dataset, we benchmarked its performance on the St. George's Hospital TMA dataset from the MesoGraph study [6]. This external evaluation is significant for two reasons. First, no additional training was applied to the new dataset, so the results represent the pre-trained model's performance on a fully unseen wholly exterior cohort. Second, although the model was trained on WSIs, it maintained strong performance on tissue microarray cores. These fragments are not only tiny (≈1 millimetre) but are selected to represent 'pure' tumour tissue. In contrast, our model was trained on large diagnostic WSIs including background tissues, and used unsupervised clustering to filter artefacts.

As the image size in TMA cores is insufficient to support the previous frequency-based method, we employed multiple instance learning (MIL) to predict mesothelioma subtypes by summarising information across tiles from a core or a biopsy. We called this method HPL-MIL and benchmarked HPL-MIL against state-of-the-art methods, max-MIL and naive-MIL (patch-based MILmethods), PINS [21], CLAM[22], MesoGraph, on the 235 cores from St. George's Hospital TMA cohort [6] (Table 1). HPL-MIL achieved higher AUC, Average Precision, Sensitivity and specificity scores across all the methods without pre-training on the cohort tissues.

4 Discussion

In this study, we applied our self-supervised HPL pipeline to the LATTICe-M cohort, which we believe to be the largest image collection in terms of area of mesothelioma tissue yet employed for AI training. We achieved state-of-the-art accuracy in two key clinically important tasks: a C-index of 0.65 in survival prediction across a 5-fold cross-validation and 85% accuracy in subtype classification (epithelioid vs sarcomatoid/biphasic). Furthermore, our method outperformed human grading in prognostication for epithelioid cases, and we identified survival-linked histomorphological patterns within each subtype, emphasising the interpretability

of self-supervised methods and identifying recurrent morphologies worthy of future study. Quantitative visual maps of HPCs (Figure 2e) and SHAP decision plots (Figure 4e) offer clinical utility for understanding AI diagnoses and future selection of therapies.

Our approach eliminates the need for retraining to retain performance on external mesothelioma datasets, thus addressing a key computational challenge in self-supervised models, as proven by its efficacy across three independent cohorts. It effectively extracts relevant morphological patterns from small TMA cores (e.g. the St. George's dataset) and WSIs of varied origins and quality (e.g. TCGA and LATTICe-M), enabling real-time clinical decision-making without extensive preprocessing. This suggests the possibility of a robust diagnostic tool for both resection material and mesothelioma biopsies, which remain a major diagnostic challenge.

Our model has essentially created a morphological atlas of mesothelioma, discovering *ab initio* the characteristic recurrent H&E morphologies which comprise the disease. The fact that these morphologies have clear biological and clinicopathological significance proves their meaning and value. For example, the discovered linkage of the tumour microenvironment to patient survival shows how crucial the morphology of immune system engagement is to tumour virulence and biology, and suggests biomarker potential in predicting responses to immunotherapy. Furthermore, RNASeq data annotation of histomorphological clusters further illustrates connections between tumour microenviroment signatures, molecular pathways, and survival, offering valuable molecular insights into the biology of the disease.

The molecular associations of HPCs help us to understand tumour virulence and suggest numerous hypotheses for mechanistic testing. For example, we see numerous mRNA cancer hallmark pathways linked to high-risk sarcomatoid HPCs, helping to explain links between morphology and outcome in molecular terms and highlighting possible areas of target discovery. Sarcomatoid clusters are directly linked to signatures of proliferation, hypoxia, and EMT in bulk sequence data, without any requirement for spatial methods or microdissection. This is in keeping with biological knowledge that sarcomatoid mesothelioma cells can proliferate rapidly under hypoxic conditions [23], and supportive of the idea that sarcomatoid dedifferentiation represents co-option of a physiological EMT pathway.

Additionally, subtle transitional morphologies in cases classified as being epithelioid overall appear to have significant prognostic value in our survival analysis. This highlights the continuous nature of epithelioid to sarcomatoid transition, and suggests the importance of accurate identification of transitional states, which is a challenging task by eye, and which is likely to benefit from our approach. Furthermore, HPL could also be used to target therapy by identifying such cases with subtle sarcomatoid changes, which are likely to be more responsive to immunotherapies [24]).

5 Online Methods

5.1 Histomorphological phenotype learning (HPL)

HPL is a tool developed to detect and categorise histomorphological patterns within large collections of whole-slide images. HPL employs an automated, self-supervised deep learning approach, eliminating the need for expert pathologists to prelabel or manually define histomorphological patterns. Once these patterns are identified, new whole-slide images can be introduced to the trained model and classified according to the pre-established patterns. This feature allows pathologists to quantify specific patterns in new patient samples precisely. The clustering of each whole-slide image into meaningful histomorphological patterns follows several sequential steps, which are described below (Figure 1b)

- **Whole-slide images pre-processing:** In this first step, whole-slide images are segmented into non-overlapping 224 × 224-pixel tiles at 5X magnification, which corresponds to a pixel size of approximately 1.8 micrometres. Tiles that do not contain at least 60% tissue coverage are filtered out to maintain relevance. Consistent pixel size and magnification are ensured during the tile processing phase to guarantee uniformity in the resulting tiles. The tiling code used for this process is accessible on DeepPATH GitHub[25] for further details.

- **Feature extraction:** HPL employs a self-supervised learning technique known as Barlow Twins [13], which matches or even exceeds the performance of other self-supervised methods. Barlow Twins delivers state-of-the-art results in standard pathology tasks compared to DINO[26], MoCo[27], and SwAV [28] mothods [29]. We previously compared Barlow Twins with DINO within HPL framework, and it showed improved performance with the cohort size similar to this study [12].

  One key feature of HPL is its ability to maintain consistent image representations, even with slight colour or zoom level variations. This capability ensures that differences in image scanning or processing across datasets do not affect the results. The aim is to capture diverse visual patterns in tissue samples and represent them as feature vectors, capturing distinct characteristics like texture. Each 224 × 224-pixel tile is converted into a vector representation, denoted as $\{z \in R^D; D = 128\}$. During training, the model is optimised to produce consistent outputs for twin inputs, ensuring robustness in vector representation.

- **Clustering:** After generating the vector representations, we employed the Leiden community detection algorithm [14] (from the Python ScanPy library [30]) to cluster tiles or vector representations with similar histomorphological features. Since

neighbouring vector representations in high-dimensional space exhibit similarity, this method effectively groups the tiles based on shared morphological patterns captured by their feature vector representations.

We began with a subsample of 750,000 tiles and constructed a nearest-neighbour graph between the tiles. From this initial set of detected clusters, we assigned the remaining vector representations to these clusters (or graph nodes) based on their distance. The number of clusters identified depends on the chosen Leiden algorithm resolution. For our analysis, multiple resolutions were applied to capture varying levels of granularity. The resulting histomorphological phenotype clusters (HPCs) enabled the quantification of patients or whole-slide images (WSIs) based on these clusters, streamlining further analysis and simplifying the understanding of complex tissue patterns.

- **Preparing compositional vectors:** At this stage, using the identified HPCs, we can characterise the entire tissue or patient by quantifying the frequency of each HPC. To achieve this, each whole-slide image (WSI) is transformed into a compositional vector **A**, where the dimensionality is equal to the total number of HPCs (c). Each element within the vector represents the percentage of the tissue area attributed to a specific HPC. This approach quantifies the contribution of each HPC to the overall tissue composition, allowing for a detailed analysis of the histomorphological landscape within a patient or a sample.

$$A = \{a_0, a_1, a_2, \cdots, a_{c-1}\} \text{ s.t. } \sum_{i=0}^{c-1} a_i = 1 \text{ and } a_i \in [0,1]$$

  For statistical compositional analysis and to prepare for the use of linear models, we apply the Centred Log-Ratio (clr) transformation [31] to our compositional vector **A** to minimise correlation between HPC frequencies. This transformation maps the vector composition from the c-part simplex into a c-dimensional Euclidean vector space. Additionally, to address zero elements in the dataset, we use multiplicative replacement [32].

- **Subtype classification:** For the diagnostic task, we employed the clr transformed compositional vectors derived from whole-slide images (WSI) and fed them into a logistic regression model (Scikit-learn [33] and Statsmodels Python library [34]). This approach is weakly supervised, utilising patient-level labels assigned by pathologists, where each patient label is applied to both the patient and their corresponding slides. We combined sarcomatoid and biphasic mesothelioma into a single class (non-epithelioid class) and compared it against the majority class, primarily consisting of epithelioid samples. Also, to address the class imbalance in the primary dataset (1:3 ratio for non-epithelioid to Epithelioid subtypes), we applied an undersampling strategy using the Edited Nearest Neighbour (ENN) technique [35] (Imbalanced-learn Python library [36]). This method reduced the majority class by removing redundant and noisy samples.

  Ultimately, the logistic regression model used the compositional vectors of WSIs to classify mesothelioma subtypes based on the contributions of HPCs. In this approach, individual HPCs serve as distinct features for our logistic regression classifier, enabling us to rank the importance of each HPC and its role in predicting specific tumour subtypes within each sample. The predicted probability of being two classes is given by:

$$Y = \frac{e^{b_0 + b_1 * clr(A)}}{1 + e^{b_0 + b_1 * clr(A)}}$$

  Where b_0 is the bias or intercept term and b_1 is the coefficient for compositional vector (A).

- **Survival analysis:** In the clinical outcome aspect of our study, we created a clr transformed compositional vector for each patient, reflecting the overall HPCs composition. We then used the Cox proportional hazards regression model [37] to analyse patient survival in relation to the HPC composition vector. Finally, Kaplan-Meier plots [38] were employed to visually distinguish between high-risk and low-risk patient groups within each dataset. For this step, we used Lifelines [39] and SciPy Python libraries [40]. For both subtype classification and survival prediction tasks, we employed five-fold cross-validation to ensure robust evaluation. The reported scores represent the average performance across all folds. Furthermore, we ensured no overlap of patients between the training and test sets, maintaining strict separation to prevent data leakage and guarantee unbiased assessments. However, for providing annotations and associations with the tumour microenvironment, we focused on a single fold for consistency and detailed analysis.

5.2 Expert pathologist annotation

Additionally, we engaged three expert pathologists to independently and blindly annotate HPCs without access to patient clinical data or additional HPC details. Each HPC was classified into one of three categories: epithelioid tumour, spindle cells/extracellular matrix, or non-tumour. The pathologists assessed each HPC's primary and secondary architectural features, HPC purity, inflammation, necrosis, nuclear atypia and biphasic components (in malignant groups). Also, they evaluated patterns such as desmoplastia and cellularity in spindle cell HPCs, as well as the tumour-stroma ratio and stromal cellularity in epithelioid HPCs.

5.3 HPC cell type enrichment analysis

We used the deep learning model HoVer-Net [41] to segment cells in each tile within the HPCs and calculate the abundance of only inflammatory cells in every HPC. While the tiles used in the HPL framework were at 5x magnification, the HoVer-Net model was trained on 20x tiles. To bridge this difference, we first applied HoVer-Net to 20x tiles, then combined 16 tiles (arranged in a 4x4 grid) to create 5x equivalents. This allowed us to map HoVer-Net's segmentation results, particularly the inflammation annotations, to specific tiles. Each tile was subsequently assigned to a corresponding HPC by calculating the average number of inflamed cells detected across the related tiles. Approximately 900 whole-slide images (WSIs) at 20x magnification were annotated using the HoVer-Net model for this analysis.

5.4 HPC tumour microenviroment signature associations

We correlated HPCs with tumour microenviroment features, hallmark pathways, and relevant biomarkers (Section 3.4). All WSIs from each patient were used to calculate the clr-transformed HPC compositional vectors. We then applied the single-sample Gene Set Enrichment Analysis (ssGSEA) to quantify pathway expression in both the Kyoto Encyclopedia of Genes and Genomes (KEGG) [42] and Molecular Signatures Database (MSigDB) [43] hallmark datasets. We also estimated immune cell subpopulation abundance using MCPcounter [44]. Ki67 RNA expression levels were also utilised across the entire sample set to assess cellular proliferation in each case, providing further insight into tumour growth activity within different morphological HPCs. Correlations between the clr-transformed HPC compositions and pathway expression levels were calculated, with only correlations having a p-value below 0.01 retained, ensuring statistical significance.

5.5 Tissue microarray benchmarking

Finally, we also benchmarked and compared the HPL model (trained on WSIs) against other state-of-the-art AI methods using an independent small dataset of tissue microarray (TMA) cores to demonstrate its robustness. We used the St. George's Hospital dataset, publicly released with the MesoGraph study [6], comprising 235 cores with associated mesothelioma subtype labels. The study benchmarked methods such as max-MIL, naive-MIL (patch-based MIL approaches), PINS [21], CLAM [22], and MesoGraph on this dataset. We evaluated our self-supervised model, pretrained on the LATTICeM WSI dataset, against the reported metrics (AUC-ROC, Precision, Sensitivity, and Specificity). Features were extracted from TMA core tiles (four tiles per core) using our pretrained model and subsequently integrated at the core level through Multiple Instance Learning (MIL). Employing gated attention in MIL [45], we predicted the probability of each core belonging to a specific mesothelioma subtype, naming this approach HPL-MIL. While MIL lacks interpretability, it allowed us to benchmark mesothelioma subtype classification against existing MIL-based methods, showcasing the generalisability and scalability of the HPL pipeline. Despite this success, the remainder of the study prioritises the more interpretable histomorphological clusters and their morphological insights to address the complexity of mesothelioma.

5.6 Datasets

The primary dataset for this study, the Leicester Archival Thoracic Tumour Investigation Cohort-Mesothelioma (LATTICe-M) [46], includes 512 patients. These images are from patients diagnosed with plural mesothelioma, resected after surgery, and categorised into epithelioid (372 cases), sarcomatoid (107 cases), and biphasic (33 cases). The dataset predominantly features male patients (436 cases) over female patients (76 cases). Figure 1a presents additional clinical details. The WSIs were sectioned and stained with Hematoxylin and Eosin (H&E) at Leicester University Hospital, scanned at 10X, 20X, or 40X magnifications. After tiling and background removal, WSIs with fewer than 100 tiles were excluded, leaving 485 patients and 3446 WSIs for the pipeline and downstream analysis. To identify significant clinical factors in this cohort, we employed a Cox proportional hazards model and found that age, mesothelioma subtype, and TNM stage significantly contributed to survival prediction. (Figure 1a)

To validate our results, we used the publicly available Cancer Genome Atlas (TCGA)-mesothelioma cohort [19], an entirely differently-scanned dataset, still comprising WSIs but obtained from multiple centres. It includes 86 samples from 74 patients with both WSIs and RNAseq data available. This cohort was primarily used to discover links between HPCs and tumour microenviroment features, pathways, and hallmarks. All HPL pipeline steps were performed on the primary dataset (LATTICe-M), and evaluation scores were reported on the fully unseen additional TCGA dataset, without any further training.

Finally, we utilised the St. George's Hospital dataset, consisting of H&E-stained TMAs from tumour biopsies collected at St. George's Hospital, London. This dataset includes four TMA slides scanned at 20× magnification using a Hamamatsu Nanozoomer S360 scanner, comprising 235 cores labelled as epithelioid, biphasic, or sarcomatoid, as the only available clinical information. The dataset, introduced in the Mesograph study [6], was used for training and testing. We employed it to demonstrate the robustness and generalisability of our trained WSI model by benchmarking and comparing its performance on TMA cores against different methods reported in the study. (Section 3.5)

6 Code and Data availability

The code for our study is on this GitHub. The LATTICe cohort was initially constructed under REC 14/EM/1159 (East Midlands REC), with ongoing resource management by the Greater Glasgow and Clyde Biorepository approved via an amendment granted by the Leicester South REC. Current use of the collection is managed under REC 16/WS/0207. The dataset will be made available upon request for individual academic research purposes. TCGA mesothelioma RNAseq data has been retrieved from UCSC Xena and images from Genomic Data Commons (GDC) portal. St. George Hospital TMA Dataset is available on MesoGraph GitHub.

7 Acknowledgements

The authors would like to express their gratitude to the CRUK Scotland Institute, the NHS Greater Glasgow and Clyde Biorepository, the University of Leicester, the Mazumdar-Shaw Chair Endowment, and the University of Glasgow for their invaluable support and contributions to this study. The manuscript was critically reviewed by Catherine Winchester (CRUK Scotland Institute). CRUK Early Detection Programme, IAMMED-Meso, EDDPGM-Nov21\100001 supported the research. KY acknowledges support from Cancer Research UK (EDDPGM-Nov21\100001 and DRCMDP-Nov23\100010), BBSRC BB\V016067\1, Prostate Cancer UK MA-TIA22-001 and EU Horizon 2020 grant ID: 101016851.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] J Christopher Wagner, Christopher A Sleggs, and Paul Marchand. "Diffuse pleural mesothelioma and asbestos exposure in the North Western Cape Province". In: *British journal of industrial medicine* 17.4 (1960), p. 260.

[2] Linda Molinari. *Mesothelioma survival rates & factors that affect patients*. Nov. 2023. URL: `https://www.mesothelioma.com/mesothelioma/prognosis/survival-rate/`.

[3] Sara Ricciardi et al. "Surgery for malignant pleural mesothelioma: an international guidelines review". In: *Journal of thoracic disease* 10.Suppl 2 (2018), S285.

[4] Arnaud Scherpereel et al. "ERS/ESTS/EACTS/ESTRO guidelines for the management of malignant pleural mesothelioma". In: *European respiratory journal* 55.6 (2020).

[5] Pierre Courtiol et al. "Deep learning-based classification of mesothelioma improves prediction of patient outcome". In: *Nature medicine* 25.10 (2019), pp. 1519–1525.

[6] Mark Eastwood et al. "MesoGraph: Automatic Profiling of Malignant Mesothelioma Subtypes from Histological Images". In: *arXiv preprint arXiv:2302.12653* (2023).

[7] Julia R Naso et al. "Deep-learning based classification distinguishes sarcomatoid malignant mesotheliomas from benign spindle cell mesothelial proliferations". In: *Modern Pathology* 34.11 (2021), pp. 2028–2035.

[8] Richard J Chen et al. "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16144–16155.

[9] Xiyue Wang et al. "Transformer-based unsupervised contrastive learning for histopathological image classification". In: *Medical image analysis* 81 (2022), p. 102559.

[10] Alexandre Filiot et al. "Scaling self-supervised learning for histopathology with masked image modeling". In: *medRxiv* (2023), pp. 2023–07.

[11] Richard J Chen et al. "Towards a General-Purpose Foundation Model for Computational Pathology". In: *Nature Medicine* (2024).

[12] Adalberto Claudio Quiros et al. "Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unannotated pathology slides". In: *Nature Communications* 15.1 (2024), p. 4596.

[13] Jure Zbontar et al. "Barlow twins: Self-supervised learning via redundancy reduction". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12310–12320.

[14] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific reports* 9.1 (2019), p. 5233.

[15] Andrew G Nicholson et al. "EURACAN/IASLC proposals for updating the histologic classification of pleural mesothelioma: towards a more multidisciplinary approach". In: *Journal of Thoracic Oncology* 15.1 (2020), pp. 29–49.

[16] Scott Lundberg. "A unified approach to interpreting model predictions". In: *arXiv preprint arXiv:1705.07874* (2017).

[17] Francoise Galateau Salle et al. "Comprehensive molecular and pathologic evaluation of transitional mesothelioma assisted by deep learning approach: a multi-institutional study of the International Mesothelioma Panel from the MESOPATH Reference Center". In: *Journal of thoracic oncology* 15.6 (2020), pp. 1037–1053.

[18] Sanja Dacic et al. "Interobserver variation in the assessment of the sarcomatoid and transitional components in biphasic mesotheliomas". In: *Modern Pathology* 33.2 (2020), pp. 255–262.

[19] Julija Hmeljak et al. "Integrative molecular characterization of malignant pleural mesothelioma". In: *Cancer discovery* 8.12 (2018), pp. 1548–1565.

[20] Laura Mannarino et al. "Epithelioid pleural mesothelioma is characterized by tertiary lymphoid structures in long survivors: results from the MATCH study". In: *International Journal of Molecular Sciences* 23.10 (2022), p. 5786.

[21] Mark Eastwood et al. "Malignant mesothelioma subtyping of tissue images via sampling driven multiple instance prediction". In: *International Conference on Artificial Intelligence in Medicine*. Springer. 2022, pp. 263–272.

[22] Ming Y Lu et al. "Data-efficient and weakly supervised computational pathology on whole-slide images". In: *Nature biomedical engineering* 5.6 (2021), pp. 555–570.

[23] Darya Karatkevich et al. "Chemotherapy increases CDA expression and sensitizes malignant pleural mesothelioma cells to capecitabine treatment". In: *Scientific Reports* 14.1 (2024), p. 18206.

[24] Kentaro Hashimoto et al. "Sarcomatoid malignant pleural mesothelioma treated with nivolumab: A case series". In: *Oncology Letters* 24.5 (2022), pp. 1–7.

[25] Nicolas Coudray et al. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning". In: *Nature medicine* 24.10 (2018), pp. 1559–1567.

[26] Mathilde Caron et al. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.

[27] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[28] Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: *Advances in neural information processing systems* 33 (2020), pp. 9912–9924.

[29] Mingu Kang et al. "Benchmarking self-supervised learning on diverse pathology datasets". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 3344–3354.

[30] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome biology* 19 (2018), pp. 1–5.

[31] Juan José Egozcue et al. "Isometric logratio transformations for compositional data analysis". In: *Mathematical geology* 35.3 (2003), pp. 279–300.

[32] Josep A Martín-Fernández, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn. "Dealing with zeros and missing values in compositional data sets using nonparametric imputation". In: *Mathematical Geology* 35 (2003), pp. 253–278.

[33] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[34] Skipper Seabold and Josef Perktold. "Statsmodels: econometric and statistical modeling with python." In: *SciPy* 7.1 (2010).

[35] Dennis L Wilson. "Asymptotic properties of nearest neighbor rules using edited data". In: *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1972), pp. 408–421.

[36] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5. URL: `http://jmlr.org/papers/v18/16-365`.

[37] Danyu Y Lin and Lee-Jen Wei. "The robust inference for the Cox proportional hazards model". In: *Journal of the American statistical Association* 84.408 (1989), pp. 1074–1078.

[38] Bradley Efron. "Logistic regression, survival analysis, and the Kaplan-Meier curve". In: *Journal of the American statistical Association* 83.402 (1988), pp. 414–425.

[39] Cameron Davidson-Pilon. "lifelines: survival analysis in Python". In: *Journal of Open Source Software* 4.40 (2019), p. 1317.

[40] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature methods* 17.3 (2020), pp. 261–272.

[41] Simon Graham et al. "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images". In: *Medical Image Analysis* (2019), p. 101563.

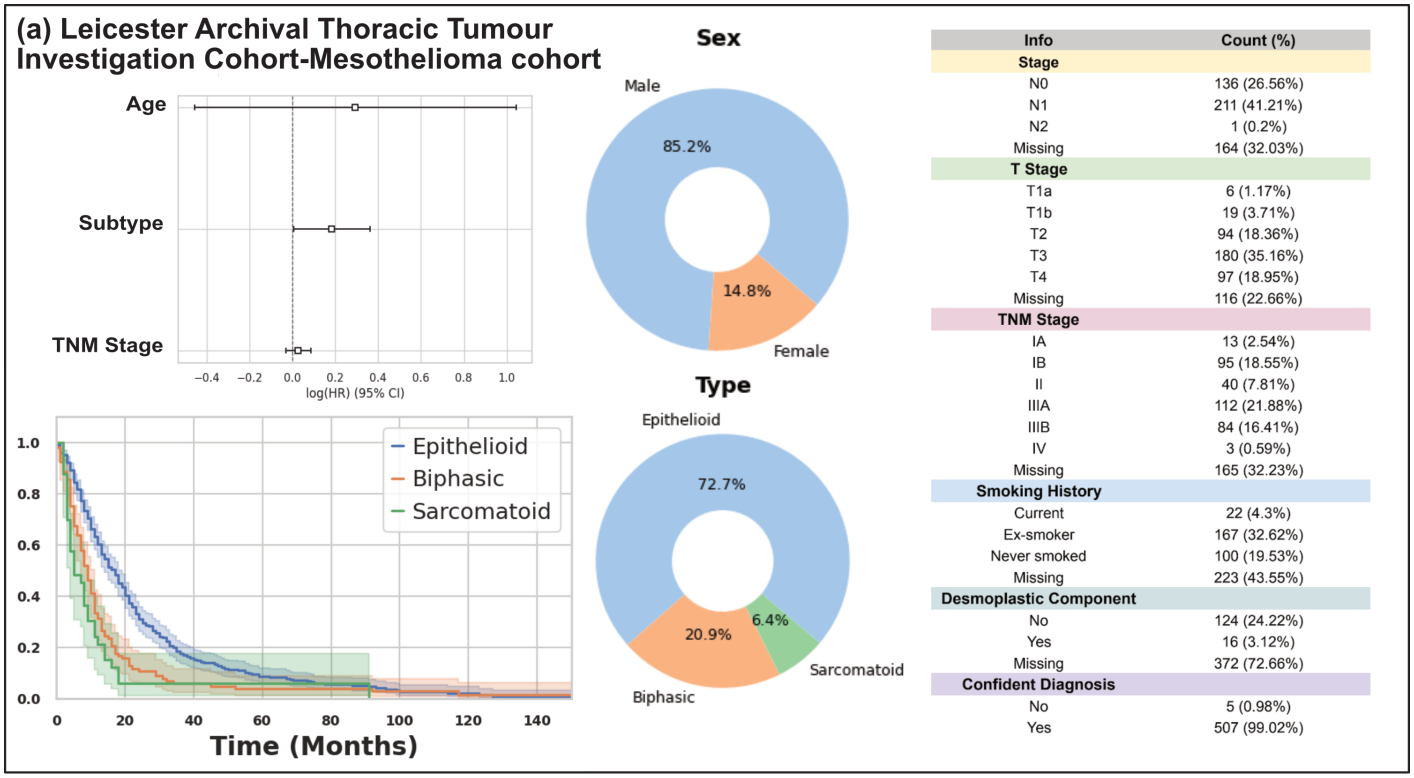
[42] Minoru Kanehisa and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[43] Arthur Liberzon et al. "Molecular signatures database (MSigDB) 3.0". In: *Bioinformatics* 27.12 (2011), pp. 1739–1740.

[44] Etienne Becht et al. "Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression". In: *Genome biology* 17 (2016), pp. 1–20.

[45] Maximilian Ilse, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning". In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136.

[46] David Allan Moore et al. "In situ growth in early lung adenocarcinoma may represent precursor growth or invasive clone outgrowth—a clinically relevant distinction". In: *Modern Pathology* 32.8 (2019), pp. 1095–1105.



Info	Count (%)
Stage	
N0	136 (26.56%)
N1	211 (41.21%)
N2	1 (0.2%)
Missing	164 (32.03%)
T Stage	
T1a	6 (1.17%)
T1b	19 (3.71%)
T2	94 (18.36%)
T3	180 (35.16%)
T4	97 (18.95%)
Missing	116 (22.66%)
TNM Stage	
IA	13 (2.54%)
IB	95 (18.55%)
II	40 (7.81%)
IIIA	112 (21.88%)
IIIB	84 (16.41%)
IV	3 (0.59%)
Missing	165 (32.23%)
Smoking History	
Current	22 (4.3%)
Ex-smoker	167 (32.62%)
Never smoked	100 (19.53%)
Missing	223 (43.55%)
Desmoplastic Component	
No	124 (24.22%)
Yes	16 (3.12%)
Missing	372 (72.66%)
Confident Diagnosis	
No	5 (0.98%)
Yes	507 (99.02%)

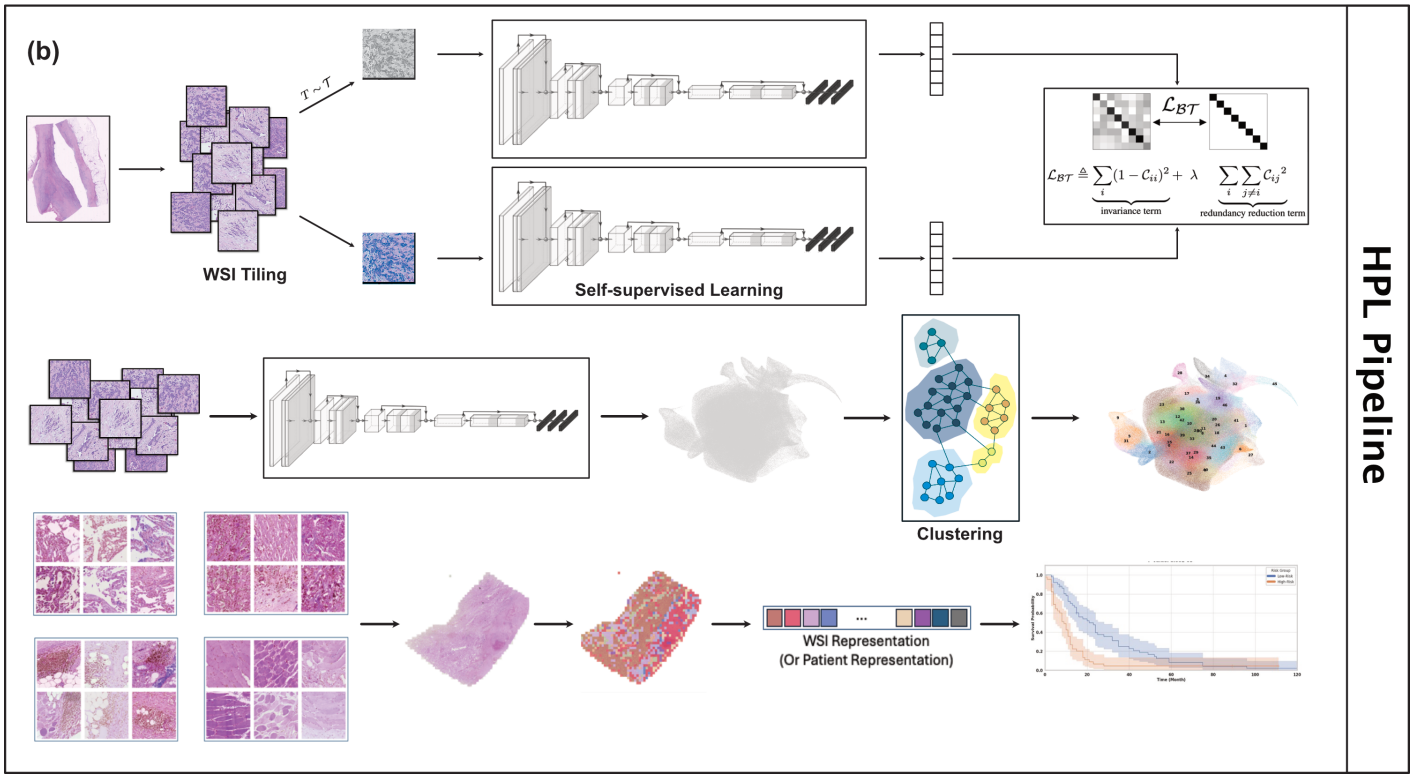


Figure 1: **(a)** Overview of the LATTICe-M dataset clinical information, highlighting key fields. Also, a forest plot for clinical data−Age, Subtype, TNM Stage, as well as Kaplan-Meier for the mesothelioma subtype variable, significantly affect outcome prediction for this cohort. **(b)** HPL Pipeline: Each WSI was divided into 224x224 px tiles. After applying various distortions, these tiles were used as input for the Barlow Twins self-supervised model. In the second step, once the model was trained, the ResNet backbone generated 128D vectors per tile, representing prominent features. These vectors were then clustered using the Leiden community clustering algorithm. In the final step, at the patient level, the clusters (HPCs) were analysed to quantify the proportion of each morphologically significant HPC within the WSI. This information was used to predict mesothelioma subtypes and outcomes using logistic regression and Cox hazard ratio models.

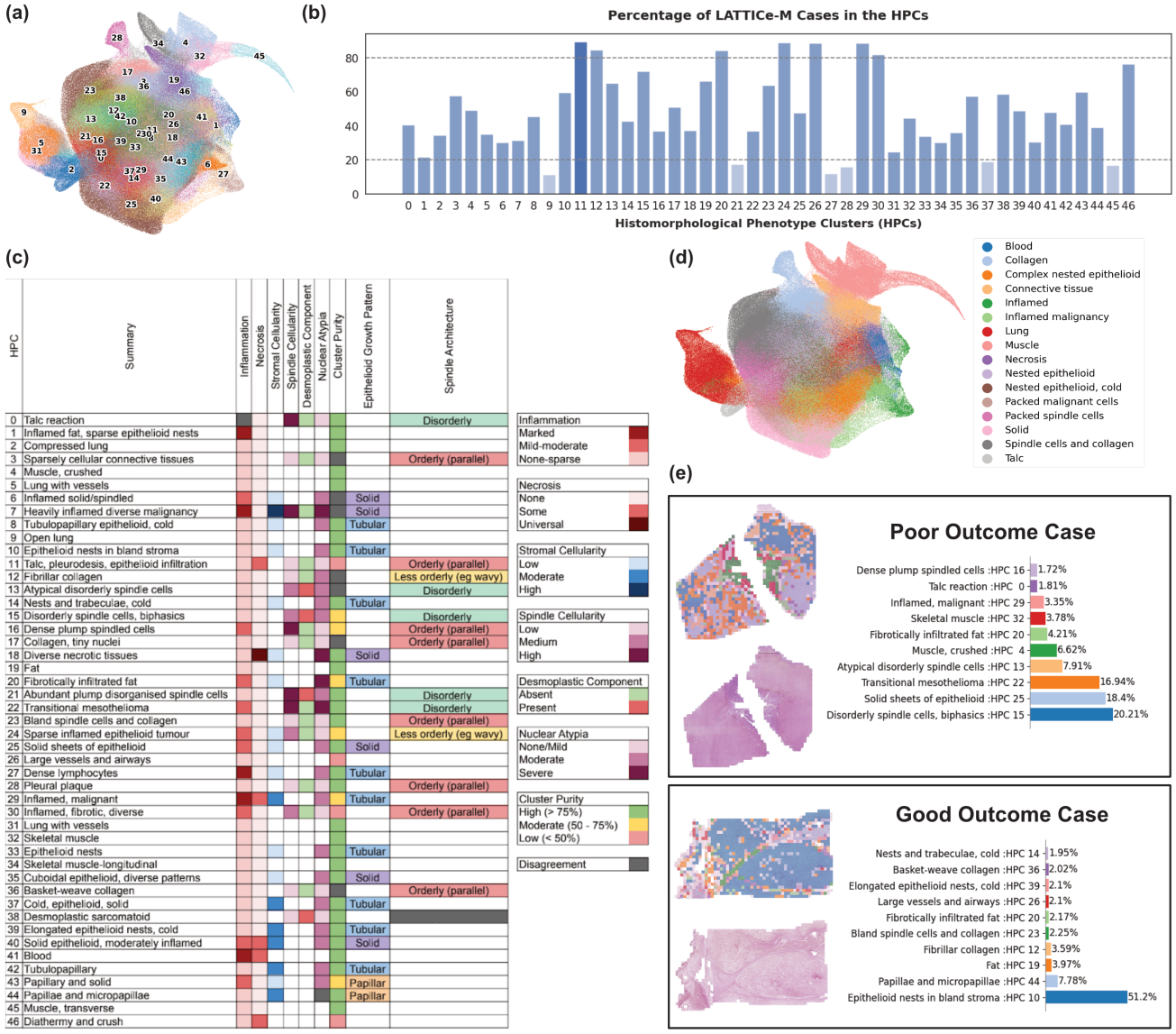


Figure 2: **(a)** Colour-coded UMAP plot for 47 HPCs. **(b)**Percentage of patients from the LATTICe-M cohorts associated with each HPC. **(c)** Pathologists' panel annotations for HPCs; three pathologists independently annotated the clusters. The table shows the majority opinion for histopathological features identified in each cluster. **(d)** Main patterns – supercluster – colour-coded UMAP; annotations were given by a pathologist following the clustering step to group HPCs with similar morphologies.**(e)**A WSI from a poor outcome and good outcome cases overlaid and quantified by HPCs, representing the differences in HPC composition and the percentage of the top 10 HPCs for each sample, alongside morphology annotations.

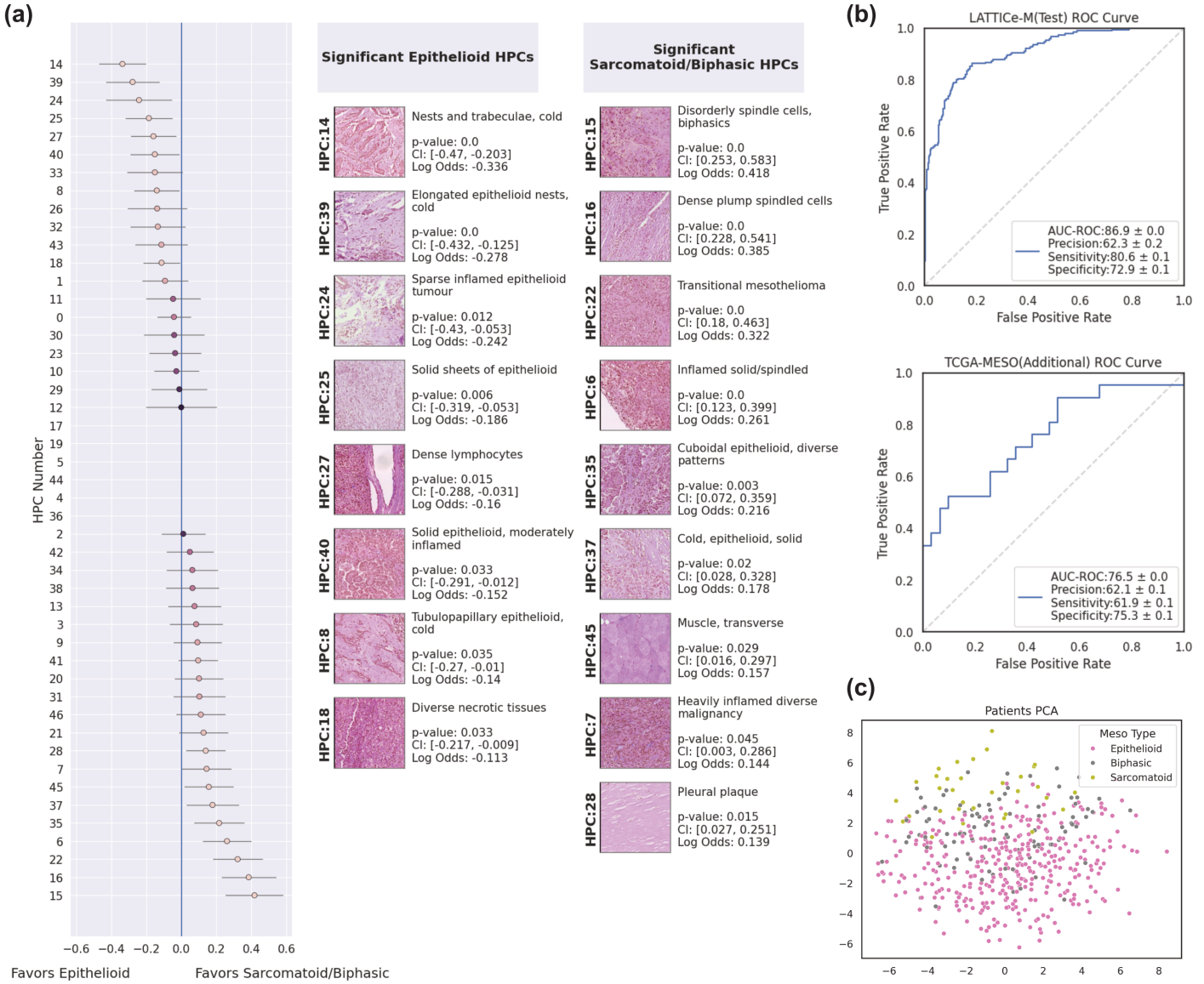


Figure 3: **(a)** Forest plot for the logistic regression model used in subtype classification, showing each HPC's contribution (regression coefficients), along with significant HPCs for each class (epithelioid vs. non-epithelioid), including their p-values, confidence intervals, log odds ratios, and pathologist annotations on HPC morphology. **(b)**The ROC (Receiver operating characteristic) curve for LATTICe-M and TCGA-MESO Datasets showing their subtype classification performance, AUC-ROC, Precision, Sesitivity and Specifity Scores **(c)** The PCA plot shows patient vector representations, colour-coded by mesothelioma subtypes.

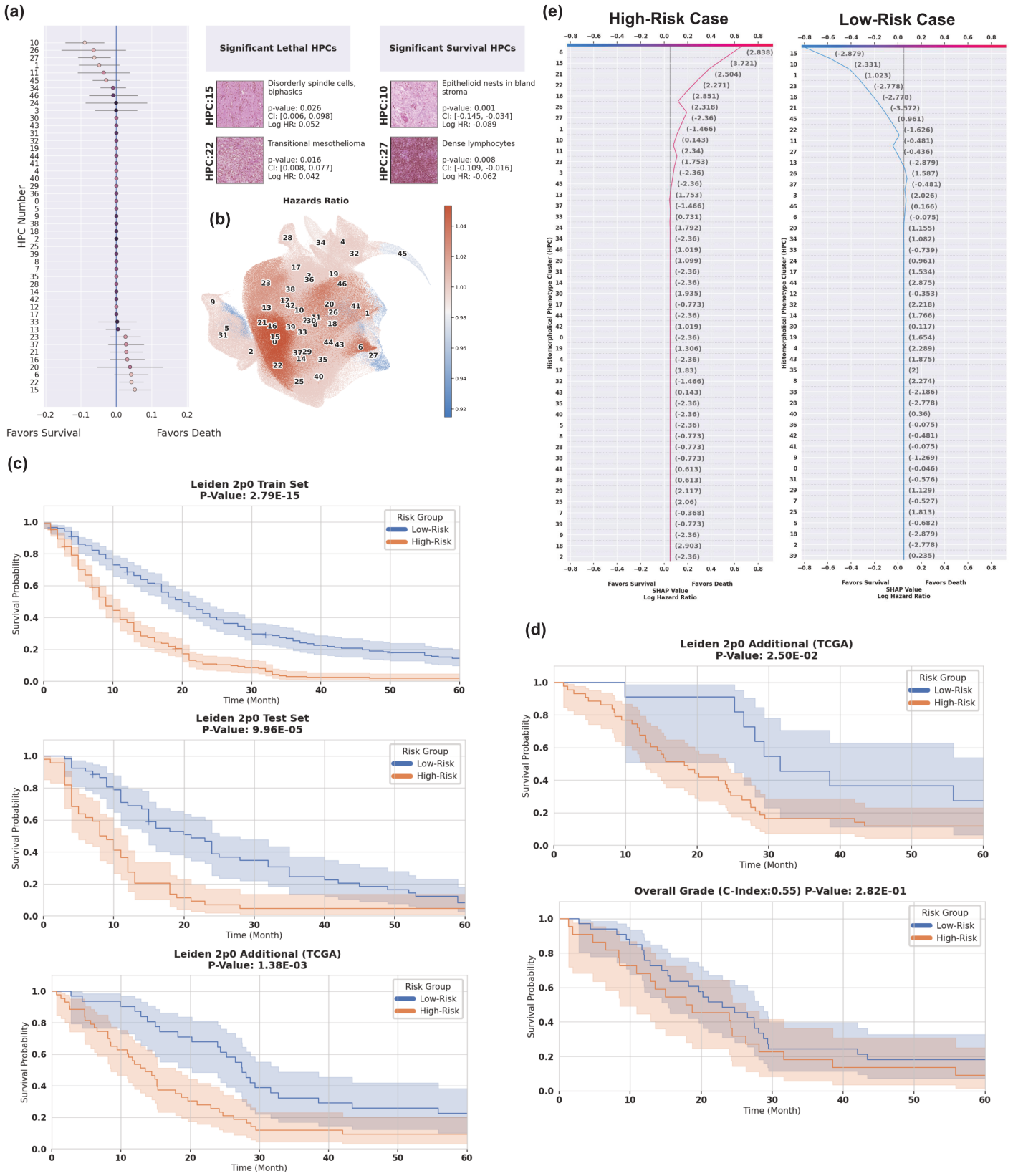


Figure 4: **(a)** Forest plot for the Cox proportional hazards model used in subtype classification, showing each HPC's contribution (regression coefficients), along with significant HPCs for each class (epithelioid vs. non-epithelioid), including their p-values, confidence intervals, log odds ratios, and pathologist annotations on HPC morphology.**(b)** UMAP plot of tile vector representations, colour-coded by hazard ratio assigned to each HPC. **(c)**Kaplan-Meier plots showing low-risk and high-risk patient groups for LATTICe-M and TCGA-Meso datasets, with their reported p-value performance scores.**(d)** Kaplan-Meier plots comparing the performance of our risk and survival predictor, pre-trained on the LATTICe-M dataset, tested on TCGA-MESO (only epithelioid cases), versus predictions based on the traditional systemic overall grade of the epithelioid cases.**(e)**An example of a SHAP decision plot for a high-risk (red) and low-risk (blue) patient, displaying the percentage of HPCs within their samples that contribute to their outcomes.

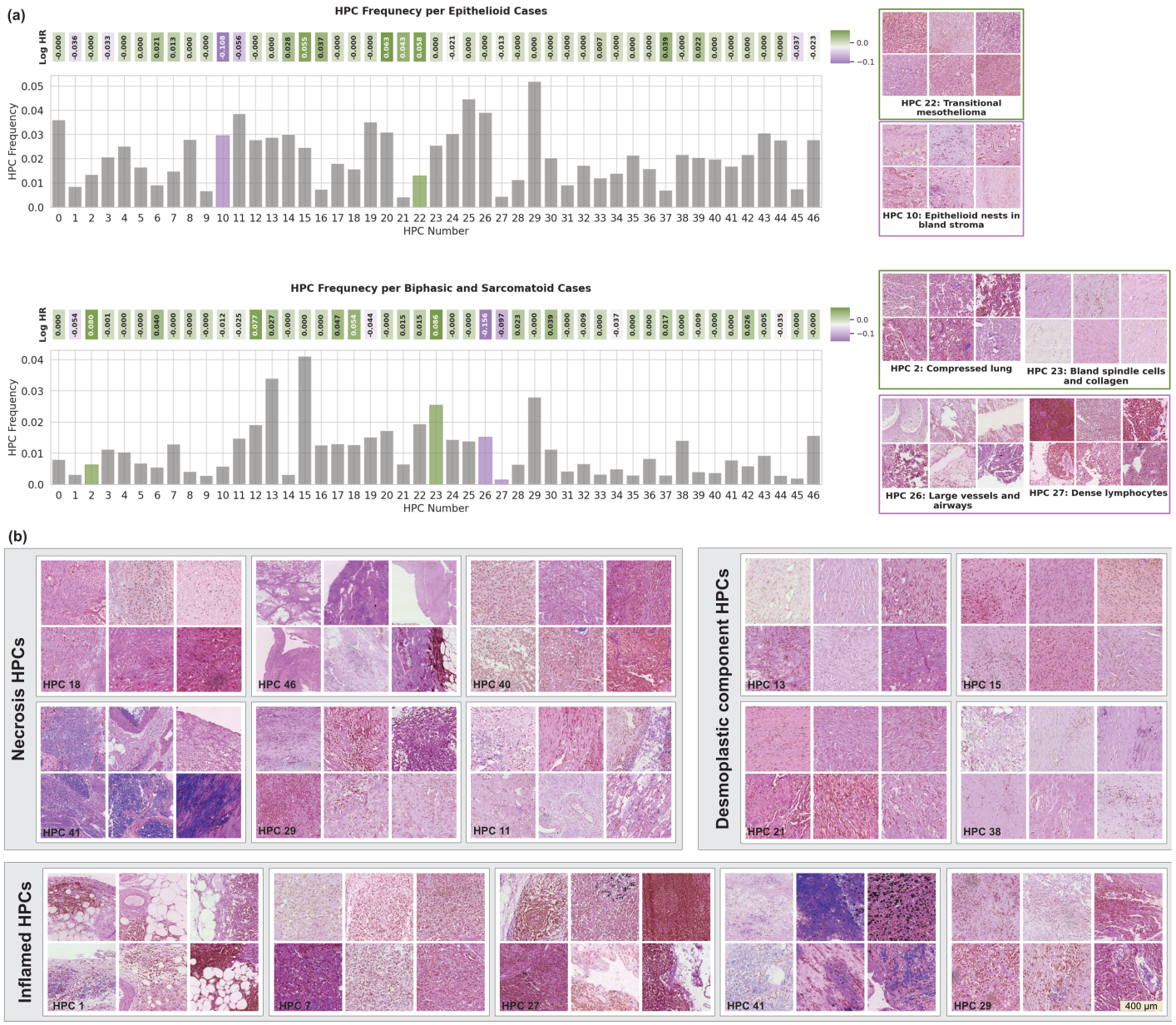


Figure 5: **(a)** HPC frequency plot for biphasic/sarcomatoid and epithelioid cases, illustrating the co-occurrence of HPCs for each subtype in comparison to the other. Based on a Cox model trained on subtype-filtered cases, significant HPCs in the model ($p - value < 0.05$) are highlighted in bold colours with respect to their log hazard ratio, with sample tiles shown in the legend. **(b)**Grouped HPCs based on pathologist annotations, highlighting key histopathological features such as necrosis, desmoplastic components, and inflammation in the tumour microenvironment. These HPCs provide insights into the different morphologies and their potential roles in disease progression and patient prognosis.

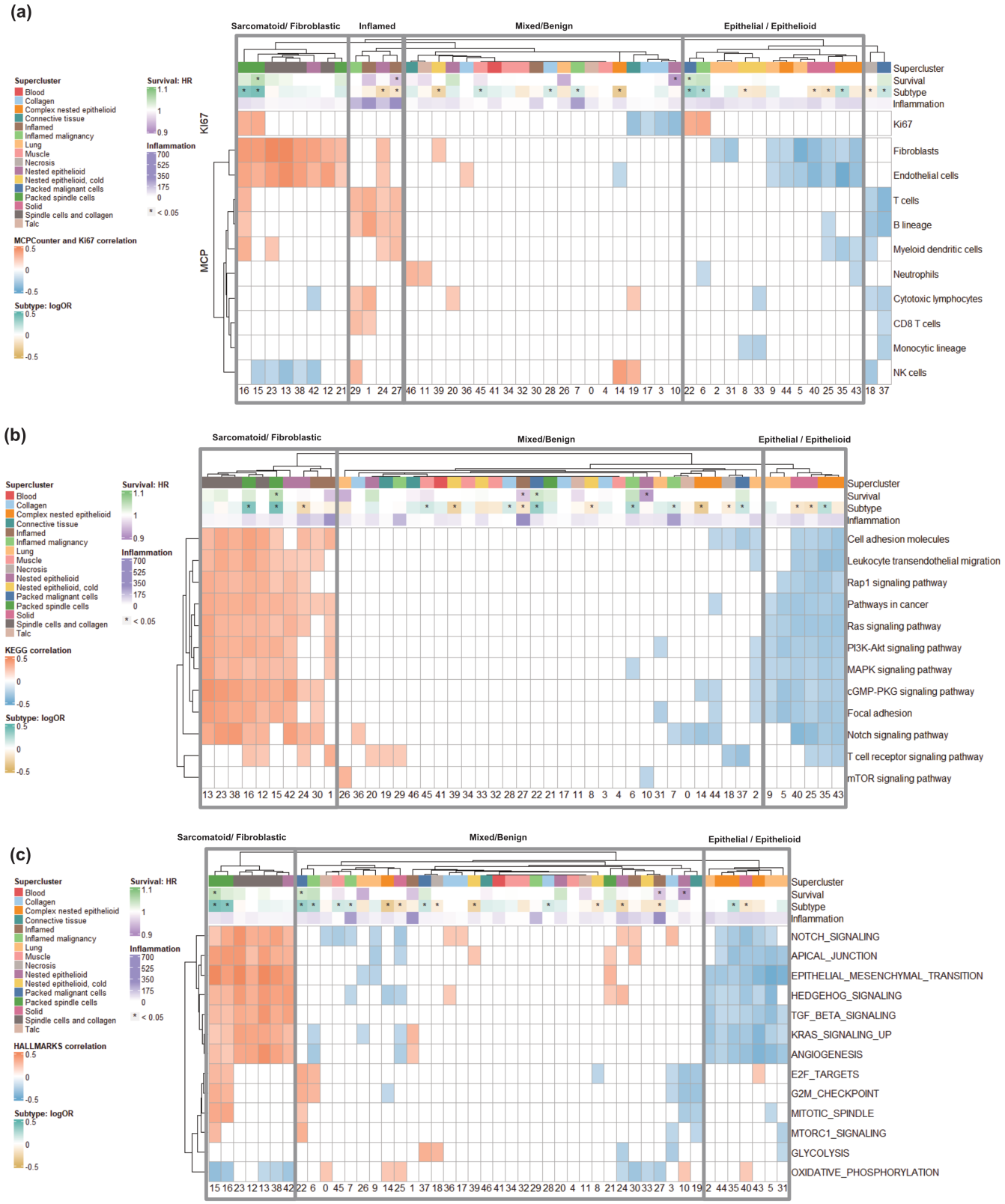


Figure 6: Correlation between WSI level HPC compositions and transcriptomic signatures in bulk RNA-seq from the TCGA MESO cohort. Only correlation coefficients with associated p-value lower than 0.05 are represented; supercluster annotations were provided by an expert pathologist. The subtype row shows the log odds ratio from the logistic regression model, while survival row is based on the partial hazards ratio predicted by the Cox model. The inflammation row was generated using inference from the HoVer-Net model. **(a)** Ki67 expression and MCPcounter tumour microenviroment signatures. **(b)** Kyoto Encyclopedia of Genes and Genomes (KEGG). **(c)** Molecular Signatures Database (MSigDB) hallmarks.