AI-driven computational pathology for recurrence risk assessment in early-stage breast cancer

Chaoyang Yan[1,♯], Linwei Li[2,3,4,♯], Xiaolong Qian[3,4,5,♯], Yang Ou[2,3,4], Zhidong Huang[2,3,4], Zhihan Ruan[1], Wenting Xiang[1], Hong Liu[2,3,4,*], Jian Liu[6,*]

Affiliations

1. Centre for Bioinformatics and Intelligent Medicine, Nankai University, Tianjin, China.
2. The Second Surgical Department of Breast Cancer, Tianjin Medical University Cancer Institute & Hospital, National Clinical Research Center for Cancer, Tianjin, China.
3. Tianjin's Clinical Research Center for Cancer, Tianjin, China.
4. Key Laboratory of Breast Cancer Prevention and Therapy, Tianjin Medical University, Ministry of Education, Tianjin, China.
5. Department of Breast Pathology and Lab, Tianjin Medical University Cancer Institute & Hospital, National Clinical Research Center for Cancer, Tianjin, China.
6. State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin, China

[♯] These authors contributed equally to this work.

[*] **Corresponding Authors:**

Jian Liu (jianliu@nankai.edu.cn)

Hong Liu (liuhong@tjmuch.com)

Abstract

Recurrence associated with poor prognosis is a primary death cause among breast cancer (BC) patients. The genomic MammaPrint (MP) test is designed to stratify risk and evaluate chemotherapy benefits for early-stage HR+/HER2- BC patients. However, it fails to reveal spatial tumor morphology and is limited by high costs. In this study, we establish a Chinese BC MP cohort and develop a weakly supervised agent-attention transformer model, CPMP, to predict MP recurrence risk from annotation-free BC histopathological slides. Our method achieves an average AUROC of 0.824 [range: 0.790-0.852] in predicting MP risk groups. CPMP is further leveraged for spatial and morphological analysis to explore histological patterns related to MP risk groups. It reveals the spatial localization of tumors at the whole-slide level and highlights differences in intercellular interaction patterns associated with MP. CPMP characterizes the diversity in tumor morphology and uncovers MP High-specific, Low-specific, and colocalized morphological phenotypes that differ in quantitative cellular compositions. Prognostic analysis in the external cohort exhibits a significant distant metastasis risk stratification (HR: 3.14, 95% CI: 1.83-5.37, *p*-value = 0.0014), underscoring the prognostic power of our model. These findings demonstrate the capability of our method in MP risk prediction, offering a flexible supplement to recurrence risk assessment in early-stage BC patients.

Introduction

Breast cancer (BC) is the most common cancer and the leading cause of cancer deaths among women worldwide[1]. Recurrence and distant metastasis are significant factors that contribute to poor prognostic outcomes and primary causes of deaths among BC patients[2,3]. Recurrence risk stratification is essential for informing therapeutic regimens in the clinical management of BC. Currently, recurrence and metastasis prediction remains a challenging task, typically assessed through clinical characteristics, the Nottingham Grading system or the Tumor-Node-Metastasis (TNM) staging system[4]. However, it can be difficult to determine the optimal treatment for individuals using these schemes, which may lead to overtreatment in some cases.

The 70-gene MammaPrint® (MP) test, which has drawn attention among clinicians, is specifically designed to evaluate recurrence risk in early-stage Hormone Receptor positive (HR+), Human Epidermal Growth Factor Receptor 2 negative (HER2-) invasive BC patients using formalin-fixed paraffin-embedded (FFPE) or fresh frozen tissue[5–7]. This test estimates the 10-year risk of distant metastasis and classifies patients into Low or High-risk groups for recurrence. The MP test effectively identifies High-risk patients who may benefit from adjuvant chemotherapy and Low-risk patients who, despite having clinical high-risk features, may safely be exempt from chemotherapy[7,8]. Although the genomic MP test offers significant clinical value, it falls short in revealing the spatial characteristics of tumors for recurrence risk diagnosis. Additionally, the MP test is time-consuming, labor-intensive, and requires substantial resources. Its high cost presents a financial barrier to many patients[9]. This situation ultimately hinders patients from receiving precise treatments based on a risk assessment of their condition. Therefore, there is a need to develop a precise and cost-efficient method for patient recurrence risk stratification.

Histopathology provides extensive insights into the spatial distribution of tissue, cellular interaction, and tumor morphology[10], establishing itself as the “gold standard” for cancer diagnosis. The increasing accessibility of digital histopathological whole slide images (WSIs), coupled with advancements in artificial intelligence (AI) technology, has propelled the research of computational pathology (CPATH) forward[11–13] and made remarkable strides in molecular biomarker discovery[14–20]. Recent studies[21–25] focused on the application of deep learning in BC histology to discover prognostic biomarkers and predict cancer recurrence. Wahab et al. designed an AI-based BRACE

marker, derived from the stromal and tumor heterogeneity, for survival prediction from routine WSIs of early-stage luminal BC[26]. Similarly, a computer-derived image feature-based risk score system was developed to evaluate the association between pathological features and recurrence prognosis for early-stage BC patients[27], enhancing risk stratification within risk categories defined by the multi-gene Oncotype DX® (ODX) assay. Lazard et al. presented a deep learning architecture for genomic homologous recombination status prediction using pathological WSIs from surgical resection specimens of BC patients[21] and Howard et al. developed a model that integrates clinicopathological variables with image-based features based on ODX assay scores to predict recurrence risk score of BC[28]. These studies underscore the potential of identifying specific biomarkers from pathological WSIs that hold prognostic significance for early-stage BC patients. However, paired genomic MP and histopathological data are limited, and previous work failed in utilizing MP-based morphological features for BC recurrence prediction. There remains a gap in research exploring histological patterns relevant to genomic MP for distant metastasis risk stratification and prognostic assessment in BC patients.

To bridge this gap, we construct the first Chinese clinical MP cohort of HR+/HER2- invasive BC patients who underwent standardized MP genomic testing at Tianjin Medical University Cancer Institute and Hospital (TJMUCH), National Clinical Research Center for Cancer. We present an AI-driven computational pathology framework (CPMP) for MP-informed recurrence risk assessment using histopathological WSIs from early-stage BC patients. We develop a spatially-aware, weakly supervised learning model with a regression-based design and an agent-attention transformer paradigm to predict continuous recurrence risk scores from annotation-free digital WSIs. We further generate spatial attention heatmap visualization at the whole-slide level to reveal intratumoral spatial localization, as well as the cellular interaction patterns associated with genomic MP within the tumor ecosystem. We characterize the phenotypic diversity in tumors and uncover High-specific, Low-specific, and colocalized morphological phenotypes among MP risk groups, each differing in their quantitative cellular compositions. Moreover, we expand our model to conduct the generalization analysis in the external TCGA-BRCA cohort and demonstrate its predictive capability for risk stratification across multiple prognostic indicators. Our CPMP framework is capable of predicting MP recurrence risk, facilitating the discovery of prognostic knowledge, and has the potential to serve as a flexible supplement to the routine clinical diagnostic workflow for early-stage BC patients.

Results

1. Overview of the CPMP model and study design

We establish a Chinese early-stage HR+/HER2- invasive BC cohort, TJMUCH-MP, comprising 477 female patients (Fig. 1a and Supplementary Fig. S1a). The patient cohort is comprised of clinicopathologic characteristics, digital WSIs, and genomic MP diagnostic results (Fig. 1b-d). The clinicopathological characteristics of patients in the TJMUCH-MP cohort can be found in Supplementary Tab. S1. MP diagnostic tests have been performed on tumor specimens of patients who met the inclusion criteria with invasive BC, HR+/HER2-, staged T1-2, lymph node-negative or 1-3 metastases from the StGallen recommendations[29]. Patients are categorized into High-risk and Low-risk based on the risk values of the genomic MP test (Fig. 1c and Supplementary Fig. S1b).

We present CPMP, a weakly supervised multiple instance learning (MIL) model for MP-informed recurrence risk assessment of BC patients based on annotation-free histopathological WSIs (Fig. 1e). Our proposed CPMP framework first preprocesses WSIs for the tissue region identification and tessellates local tiles with the size of 256×256 within the foreground regions at 20× objective magnification (Supplementary Fig. S1c). The median number of tiles across WSIs in the TJMUCH-MP cohort is 8,229, and the distribution of the number of tiles in WSIs has been reported (Supplementary Fig. S1d). Then, the CPMP framework captures diverse patterns and generates low-dimensional feature representation of local tile instances using self-supervised foundation models (Supplementary Fig. S1e). Given that the spatial relationships among tissue regions suggest cellular interactions that potentially reveal biological behaviors and prognostic significance, the Positional coordinate information Embedding (PE) of each tile is encoded and integrated with low-dimensional tissue embeddings to enhance spatial correlation among local regions. CPMP introduces the transformer architecture (Fig. 1f) for feature aggregation and employs the agent-attention mechanism[30] (Fig. 1g) to achieve a balance between global representation capability and computational efficiency. Besides, CPMP adopts the regression-based multi-layer perceptron (MLP) predictor to yield a continuous numerical value that represent genomic MP risk scores.

We developed and trained the CPMP model to predict MP-informed patient-level recurrence risk (Fig. 1h) using the clinical TJMUCH-MP cohort. In addition, we leverage CPMP to generate whole-slide-level spatial attention heatmaps using histopathological slides (Fig. 1i), which provide key

morphological patterns and spatial localization associated with MP recurrence risk. We investigate tumor morphological patterns by clustering the learned tile embeddings from CPMP to characterize tumor diversity and identify specific, colocalized morphological phenotypes across MP risk groups (Fig. 1j). Moreover, we generalize our trained CPMP model to the external TCGA-BRCA cohort for recurrence risk stratification and prognostic analysis, aiming to validate the potential capabilities of CPMP in assessing distant metastasis risk (Fig. 1k). More details are provided in the Methods section.

2. CPMP outperforms other methods for MP recurrence risk prediction

The model's ability to predict the recurrence risk from histopathological WSIs was measured on both continuous risk scores and discrete risk groups using the TJMUCH-MP cohort. To mitigate data bias and underlying scanning effects (Supplementary Fig. S1f), we employed a patient-level 5-Time 5-Fold cross-validation strategy (Fig. 2a) for model development and evaluation. Our model achieved an average Spearman correlation coefficient of 0.648 [$p < 0.0001$, range: 0.606-0.684], indicating a favorable consistency between genomic MP risk scores and recurrence risk prediction probabilities (Fig. 2b and Supplementary Fig. S2a-e). In classifying the MP recurrence risk groups, CPMP attained an average area under the receiver operating characteristic curve (AUROC) of 0.824 [range: 0.790-0.852] (Fig. 2c), with an accuracy of 0.762 [range: 0.750-0.773]. CPMP provides Low-as-positive and High-as-positive modes, which means that we aim to focus on Low-risk patients to avoid overtreatment or, conversely, to observe the High-risk patients who may be at risk from recurrence. CPMP reached average areas under the precision-recall curves (AUPRCs) of 0.873 [range: 0.835-0.900] on the Low-as-positive mode (Fig. 2d) and 0.774 [range: 0.751-0.795] on the High-as-positive mode (Fig. 2e).

The performance of our spatially-aware, agent-attention transformer model was further compared with other weakly supervised MIL methods, including ABMIL[31], CLAM[32], Wagner et al.[33], and TransMIL[34]. Details of all comparative models are provided in the Methods section. CPMP outperformed all of the comparative approaches in MP recurrence risk prediction in terms of systematic metrics of AUROC, balanced accuracy, weighted average recall, precision, F1 scores, ROC curves, and precision-recall curves (Fig. 2f, Supplementary Fig. S2g-i, and Supplementary Tab. S2). The average AUROC value of our model is 0.824, while those of other methods, namely CLAM, TransMIL, ABMIL, and Wagner et al., are 0.798 [range: 0.679-0.876], 0.755 [range: 0.608-0.841], 0.810 [range: 0.756-0.873], and 0.806 [range: 0.749-0.847], respectively. It is noteworthy that the

Transformer-based methods (Wagner et al. and TransMIL) obtained acceptable performance but fell short in comparison with our transformer-based CPMP. We visualized the predicted risk probabilities from comparative methods alongside genomic MP risk scores of patients (Fig. 2h). The results indicate that these classification-based methods tend to predict risk probabilities that are skewed toward both ends. Meanwhile, CPMP achieved improvements of 0.12, 0.06, 0.04, and 0.05 in the Spearman correlation coefficient over TransMIL, CLAM, ABMIL, and Wagner et al., respectively. This is attributed to the regression-based design in our method. The computational operations and runtime costs for inference were conducted on a workstation with RTX 4090 GPUs for comparison. We can see that CPMP reduces inference operations (G/MACs) by 6.50× and increases inference time by 1.82× times, while improving AUC performance by 2% ($p < 0.05$, statistically significant difference), compared with the Wagner et al. transformer (Fig. 2g). Also, CPMP requires fewer parameters than other methods (Supplementary Fig. S2f). CPMP attains the most superior efficiency-AUC performance during inference, striking a balance between the slide representation capability and computation complexity.

3. CPMP highlights its effectiveness in leveraging annotation-free WSIs

Blurry areas, pen markings, and non-tissue regions are inevitably present in annotation-free histopathological WSIs from real-world clinical data, introducing noise into the training data. Thus, we investigated CPMP's sensitivity to noise by removing noisy tiles. Specifically, we performed zero-shot classification to determine tissue types (tumor, adipose, stroma, immune infiltrates, gland, necrosis or hemorrhage, background or black, and non) across all tiles through the visual-language foundation model PLIP[35] (Supplementary Fig. S3a, b). Local tile images along with their corresponding zero-shot classification types were visualized (Supplementary Fig. S3c). Tiles classified as “background or black” and “non” types are typically blurry, pen-marked, or lacking morphological detail. In contrast, tiles identified as the six specific tissue types (tumor, adipose, stroma, immune infiltrates, gland, and necrosis or hemorrhage) contain rich morphological information and are thus utilized for training the other model. Evaluation results for both models indicate that the model with noisy tile filtering achieved comparable performance to the CPMP model without such filtering (Fig. 3a and Supplementary Tab. S3), showing no statistically significant difference between the two. This reflects the robustness of our weakly supervised CPMP

approach in handling local noisy information within annotation-free slides, making it more suitable for clinical scenarios.

We next evaluated the scale flexibility of CPMP in the number of local tiles in WSIs. We conducted random sampling from the set of local tiles in each WSI, employing sample sizes of 1,000, 5,000, and 10,000 tiles as subsets for model training. The trained models on various subsets of tiles exhibit consistent performance trends on the various number of tiles, with differences in AUROC remaining within 0.01 compared to the model using all tiles (p-value > 0.05, indicating no statistically significant differences, Fig. 3b and Supplementary Tab. S4). Notably, these models demonstrated robust performance with AUROC values exceeding 0.810 when making predictions based on as few as 100 tiles up to the full set of tiles. This suggests that CPMP is both flexible and efficient in recurrence risk inference using histopathological WSIs.

We further compared CPMP with a clinical-based logistic regression model that was trained on clinicopathological features for MP recurrence risk prediction for patients. The detailed settings for this clinical-based model were listed in the Methods section and Supplementary Tab. S5. Clinical-based model obtained an acceptable AUC value of 0.800 (Fig. 3c and Supplementary Tab. S6) but it was relatively inferior to CPMP, especially given the fact that the clinical-based logistic regression model requires molecular information such as hormone receptor status and Ki-67 percentages, whereas CPMP relies solely on routine Hematoxylin and Eosin (H&E) stained histopathological slides.

We conducted ablation studies on CPMP with respect to attention mechanisms, spatial positional encoding schemes, predictor head strategies, and tile-level feature representation models. The design outlines and the experimental configurations utilized for the ablation models were illustrated in Supplementary Fig. S4. Details about these modules are provided in the Methods section. The quantitative results reported that our CPMP method with the agent-attention mechanism brought a 1.8% improvement in AUROC over the self-attention transformer model and it was insensitive to the number of agent tokens (Fig. 3d and Supplementary Tab. S7). Our PE-based model outperformed the PPEG-based[34] and No-PE models in AUROC, weighted average recall, and F1 score (Fig. 3e and Supplementary Tab. S8). Additionally, a similar improvement can be observed in our regression-based mean square error supervision scheme when compared to the cross-entropy approach (Fig. 3f and Supplementary Tab. S9). In the ablation study of tile-level feature representation, the Prov-

GigaPath[36] and Virchow[37] models excelled UNI[38] in terms of AUROC (0.827 vs. 0.824), while H-optimus-0[39] achieved the highest values across other metrics (Fig. 3g and Supplementary Tab. S10). Additionally, the performance of CTransPath[40] and Phikon[41] was inferior to that of other models in representing tile embeddings for MP recurrence risk prediction in the TJMUCH-MP cohort. Given that the embedding dimensions of tile representations of Prov-GigaPath, Virchow, and H-optimus-0 are larger than those of UNI. We reported the performance results of the UNI-based model in the subsequent analysis.

4. CPMP reveals the spatial localization of tumors at the whole-slide level

Our proposed weakly supervised CPMP model not only predicts patient recurrence risk using spatial histopathological WSIs but also infers tile-level attention scores. These scores can offer insights into spatial factors that influence patient outcomes from the perspective of whole-slide level. We performed whole-slide-level attention heatmap inference on test cases. Spatial heatmap results were generated by mapping the attention scores of tiles to their corresponding locations within the slide. This allows us to visualize the contributions of different regions and display key morphological patterns. Details are provided in the Methods section. The representative histopathological WSIs selected from genomic MP Low-risk and High-risk groups were presented, along with their spatial visualization results at both the whole-slide and region-of-interest (ROI) levels (Fig. 4a-c, e-g and Supplementary Fig. S5a-c, e-g). Genomic MP diagnostic results and recurrence risk values predicted by the CPMP model were also marked in the box (Fig. 4a, e and Supplementary Fig. S5a, e). One can observe that CPMP exhibits varying levels of attention confidence in tissue regions, with high confidence indicated in red and low confidence in blue-purple, at the whole-slide level for both Low-risk and High-risk cases (Fig. 4b, f and Supplementary Fig. S5b, f). The pathologists confirmed that the areas identified as high confidence closely correspond to regions where tumor cells proliferate, based on their spatial locations. It can also be observed from the magnified ROIs that the high-attention heatmap focuses on the tumor regions and effectively outlines the boundary between tumor and normal stroma tissue (Fig. 4c, g and Supplementary Fig. S5c, g). Hot regions rendered by the spatial heatmap visualization are highly consistent with the tumor regions pre-marked by experts (Supplementary Fig. S5e, f). CPMP identifies the top 10 local tiles with the highest attention scores from each slide (Fig. 4d, h and Supplementary Fig. S5d, h). These selected tiles display clear patterns of tumor morphology applicable to both Low-risk and High-risk cases. The spatial heatmap results

indicate that our model can comprehensively pinpoint tumor growth locations and capture the spatial localization of tumor ROIs at the whole-slide level. This capability of tumor spatial localization offers a visually explainable approach to recurrence risk prediction that complements the genomic MP test with valuable spatial insights.

5. CPMP deciphers cellular interaction patterns within the tumor ecosystem

To gain deeper insights into the intercellular spatial relationships within the tumor ecosystem across MP recurrence risk groups, high-attention ROIs localized by CPMP were extracted for simultaneous nuclear instance segmentation and classification (Supplementary Fig. S6a). HoverNet[42] was utilized for the identification of five types of cell nuclei: neoplastic (tumor) cells, inflammatory cells, necrotic (dead) cells, connective (stroma) cells, and non-neoplastic epithelial cells. The contours of the identified nuclei were delineated on the fine-grained histopathological ROIs. From the nuclei segmentation visualization results for both Low-risk and High-risk ROIs (Fig. 5a, b), we can observe that the identified cell nuclei are predominantly tumor cells (outlined in red), while stromal cells (indicated in blue) are sparsely distributed around the tumor tissue. The results of nuclei identification and visualization of the top 10 local tiles further support this finding (Fig. 5c, e and Supplementary Fig. S6c-h). It is noteworthy that High-risk cases exhibit more dispersed spatial distribution among tumor cells compared to Low-risk cases. We thus quantified the cellular composition of the five cell types in each local tile and defined the density of tumor cells (TcD) as the number of identified neoplastic cells per square millimeter (Fig. 5d, f). We can see that there is a clear difference in tumor cell density between the two MP risk groups. The number of neoplastic cells in Low-risk tiles is generally larger than that in High-risk tiles (median: 103 vs. 53, Low-risk vs. High-risk). The measured TcD values for Low-risk tiles are consistently greater than 4,000 tumor cells per mm^2 (median: 5,698, [range: 4,149-7,081]), while those for High-risk tiles are all below 4,000 neoplastic cells per mm^2 (median: 2,932, [range: 1,604-3,983]).

Further, neoplastic cells, connective cells, and inflammatory cells (the main cellular components in the tumor ecosystem[43]) were employed to construct topological graphs among these cell nuclei to facilitate our understanding of intercellular interactions within the tumor microenvironment (Fig. 5g, h). The cellular topological graphs were visually characterized by interactions between tumor cells and stromal cells, with minimal evidence of lymphocyte presence and interaction with other cell types within both Low-risk and High-risk ROIs. To quantitatively characterize the spatial

intercellular interaction patterns within the tumor ecosystem, the sc-MTOP[44] framework was utilized to extract the topological graph-based features at the single-cell level. We only focus on profiling the topological features of tumor cells that interacted with other cells, specifically examining the connectivity graphs of tumor-tumor, tumor-stroma, and tumor-inflammatory interactions (Supplementary Fig. S6b). The distribution of the MeanEdgeLength features (the average edge lengths of a cell interacting with other cells) related to graph-edge lengths of tumor cells was analyzed and depicted in Fig. 5i-k. Among Low-risk patients, the graph-edge length distribution of tumor-tumor interaction graphs exhibit a higher peak, suggesting a denser and more abundant distribution of tumor cells. In contrast, High-risk patients display a sparser distribution of tumor cells (Fig. 5i, $p < 0.0001$, statistically significant difference). The edge length distribution for tumor-stroma and tumor-inflammatory interaction graphs remains generally consistent across both patient groups (Fig. 5j, $p = 0.2467$, and Fig. 5k, $p = 0.2729$, no statistically significant difference). These results provide evidence that CPMP harnesses whole-slide-level heatmap visualization for deciphering the patterns of intercellular interactions within the tumor ecosystem, which are associated with genomic MP recurrence risk groups.

6. CPMP characterizes morphological phenotypes associated with MP risk groups

Breast cancer ecosystems often exhibit both intra-tumoral and inter-tumoral heterogeneity, which influences tumor progression, metastasis formation, and patient prognosis[45]. To this end, we aimed to characterize the tumor morphological patterns throughout the entire test set and explore the specific and common phenotypes associated with MP risk groups using CPMP. Specifically, the top 100 tiles with the highest attention scores in each slide were identified to create a collection of high-confidence local regions. Each tile in this collection was also fed into CPMP for inference. It can be observed from the uniform manifold approximation and projection (UMAP) plot that tile instances display a clear distinction among different levels of the genomic MP risk (Fig. 6a). The recurrence risk prediction at both the WSI-level and the single-tile-level exhibit similar distribution characteristics in the UMAP (Fig. 6b, c), in keeping with the observations from Fig. 6a. The model's capability of predicting recurrence risk at single-tile-level is more pronounced, as it demonstrates a clear trajectory direction of recurrence risk development (indicated by the arrow in Fig. 6c). These results suggest that the high-confidence local tiles identified by the CPMP model provide a sufficient representation of WSIs.

We further performed Leiden clustering on tile embeddings to characterize phenotypic diversity of tumor morphology and attached the Low-risk or High-risk group information for each subcluster based on the genomic MP risk group of the WSI from which the local tile originated. High-attention tumor tiles were classified into 27 subclusters (named SC0-SC26) with different morphological embeddings (Fig. 6d). Also, some subclusters only present in a specific risk group and others appear in both risk groups (Fig. 6e-g and Supplementary Fig. S7a, b). Our model identifies 4 subclusters (SC14, SC21, SC23, and SC24) that are exclusively associated with MP High-risk group (Fig. 6e) and 10 subclusters (SC1, SC5, SC6, SC7, SC9, SC10, SC20, SC22, SC25, and SC26) specifically related to MP Low-risk group (Fig. 6f). The remaining 13 subclusters colocalize in both MP risk groups (Fig. 6g), reflecting the consistent tumor morphological characteristics of early-stage BC patients to whom the genomic MP test is applicable.

Representative local tiles from each subcluster along with the corresponding nuclear identification results were depicted to highlight the morphological and phenotypic characteristics linked to genomic MP risk groups (Fig.6h-j, Supplementary Fig. S7c, d, and Supplementary Fig. S8a). The MP High-specific subcluster phenotypes (Fig. 6h and Supplementary Fig. S7c) exhibit loosely distributed tumor cells interconnected with stromal cells, while the MP Low-specific subcluster phenotypes (Fig. 6i and Supplementary Fig. S7d) present a high percentage of aggregated tumor cells. These visually morphological patterns are also reflected in the colocalized subcluster phenotypes (Fig. 6j and Supplementary Fig. S8a). These results reinforce that the morphological differences among subcluster phenotypes mainly depend on the patterns of intercellular interactions and the aggregation abundance of tumor cells.

We next quantified the cellular composition of the five cell types for each subcluster and performed hierarchical clustering on the 27 subclusters (Fig. 6k and Supplementary Fig. S8b). Four MP High-specific subclusters are grouped together, while the MP Low-specific subclusters are clustered at the opposite end, forming four subgroups (SC26, SC25, and SC5; SC22, SC9, and SC20; SC6; and SC10, SC1, and SC7). The distinct cellular compositions between the High-specific and Low-specific subcluster phenotypes underline the capacity of CPMP to disclose the morphological patterns related to molecular risk groups.

7. CPMP serves as a potential tool for prognostic risk assessment

The predictive and generalization capabilities of CPMP need to be validated in external cohorts. However, existing public datasets in digital pathology lack MP recurrence risk information although they offer a wealth of histopathological WSIs. Considering that the activity levels of specific genes in MP influence tumor progression and correlate with patient recurrence and prognosis, CPMP was investigated for recurrence risk assessment and prognostic analysis in The Cancer Genome Atlas (TCGA) -BRCA cohort (Supplementary Fig. S9a-c). The whole-slide-level spatial heatmap inference results in the external TCGA-BRCA cohort reveal that CPMP captures the spatial localization of the tumor, which potentially associated with patient clinical outcomes (Fig. 7a, b, Supplementary Fig. S10, and Supplementary Fig. S11). We then assessed the prognostic value of CPMP using the clinical outcome endpoints. Disease-Free Interval (DFI), Progression-Free Interval (PFI), and Overall Survival (OS) were derived from the TCGA Clinical Data Resource[46]. We can observe a significant risk-group stratification on DFI, PFI, and OS prognostic indicators (log-rank test, p-value < 0.05) from Kaplan-Meier survival curves (Fig. 7c-e). Patients in the High-risk group had a worse outcome than patients predicted as Low-risk by CPMP, exhibiting statistically significant findings in DFI (a hazard ratio (HR) of 1.91, 95% confidence interval (95% CI): 1.20-3.05, p-value = 0.018), PFI (HR: 1.58, 95% CI: 1.11-2.26, p-value = 0.022), and OS (HR: 1.54, 95% CI: 1.08-2.18, p-value = 0.027). We performed a prognostic analysis of CPMP risk groups correlating with Distant Metastasis-Free Survival (DMFS) in the TCGA-BRCA dataset (Fig. 7f). Patients in the High-risk group experienced a higher probability of distant metastasis events than those classified as Low-risk (HR: 3.14, 95% CI: 1.83-5.37, p-value = 0.0014). Especially, the distinction between patients in the High-risk and Low-risk groups became more evident beyond the early survival times (~ 5 years) as observed in the Kaplan-Meier (KM) survival curves for DMFS. This reveals the capability of CPMP in identifying late distant metastasis events. We investigated 16 patients suffering from distant metastasis beyond 5 years, of whom 14 (87.5%, a mean follow-up time of ~9.44 years) were identified as High-risk by CPMP. We also evaluated the prognostic impact of the predicted risk groups on DMFS, DFI, PFI, and OS using multivariable Cox proportional hazards models (Fig. 7g and Supplementary Fig. S12a-c). The risk groups inferred by CPMP exhibited a significant prognostic impact on DMFS, with an HR of 2.76 (95% CI: 1.27-6.00, p-value = 0.01, Fig. 7g). It suggests that CPMP-derived risk group is an increased risk factor for BC distant metastasis in TCGA-BRCA patients, demonstrating the prognostic capability of CPMP. It can be implied that CPMP effectively captures valuable prognostic

features from tumor morphology in BC histopathological slides. This further highlights the potential of CPMP to be applied in clinical settings for prognostic risk assessment of BC patients.

Discussion

In this study, we presented the weakly supervised CPMP model for MP recurrence risk prediction from H&E-stained histopathological WSIs of BC patients. The predictive capability of the CPMP model was shown by inferring both the continuous risk scores and the discrete recurrence risk groups in the real-world TJMUCH-MP clinical cohort. Compared to the existing weakly supervised MIL methods, our method achieved superior AUC-efficiency performance and yielded a better alignment with genomic MP risk scores in the cohort.

To the best of our knowledge, this is the first work to explore spatially morphological tumor patterns associated with genomic MP for recurrence risk assessment and prognosis analysis in BC patients. Our model facilitated the characterization of the tumor pattern diversity in early-stage BC patients for whom the genomic MP test is applicable. It enhanced the discovery of High-specific, Low-specific, and colocalized morphological phenotypes among MP risk groups from routine histopathology slides. The intercellular interactions and quantized cellular composition patterns exhibited distinctions across diverse tumor phenotypes, which are related to genomic MP recurrence risk groups. Moreover, our study provided evidence that the CPMP model demonstrated superior generalization capabilities in identifying late distant metastasis events in the TCGA-BRCA cohort. It also exhibited significant risk-group stratification capabilities across multiple prognostic indicators. We implied that the prognostic value of CPMP arises from the bridged association between genomic MP relevant to recurrence prognosis and the tumor morphology presented in histopathological slides. We will validate this morphological prognostic biomarker in future studies.

The genomic MP test provides significant clinical value and has been proven to be an effective diagnostic tool in current clinical applications. It is recommended for invasive BC patients who are HR+/HER2-, staged T1-2, and either lymph node-negative or with 1-3 metastases[29]. This test evaluates the 10-year risk of distant metastasis and classifies patients into High or Low recurrence risk groups. High-risk patients are likely to benefit from adjuvant chemotherapy, while Low-risk patients with clinically high-risk features can safely exempt from chemotherapy. The genomic MP profiling test for recurrence risk diagnosis is performed on tumor specimens using targeted RNA

next-generation sequencing technology. However, it fails to adequately provide the spatial and morphological characteristics of tumors. To make up for the limitations of traditional MP diagnostic tests in terms of spatial analysis, we implemented a WSI-level attention heatmap visualization method in our model. This approach captures the spatial relationships and cellular interaction patterns within tumors from the perspective of the whole-slide level. It reflects that our spatially-aware, weakly supervised learning CPMP model is capable of generating a whole-slide-level attention heatmap, facilitating the visualization and discovery of spatial-morphological localization.

In addition, the time and financial costs of MP diagnostic tests are significant. This financial barrier prevents many patients from receiving precise therapeutic regimens. CPMP, based on AI technology and weakly supervised learning, underscored its remarkable capability in predicting the risk of distant metastasis for early-stage BC patients using routine histopathological slides. We experimentally evaluated that the CPMP model can infer a reliable risk score and a whole-slide-level spatial attention heatmap visualization in about 5 minutes for one WSI sized at 50,000×50,000 pixels (25 mm × 25 mm). The CPMP-based risk assessment not only rapidly delivers results but also significantly reduces the financial burden on patients. This makes it a viable complementary option for those who may not be able to afford the genomic MP diagnostic test. Additionally, it can serve as a cost-efficient AI-driven system for pre-screening among BC patients in the clinical workflow (depicted in Supplementary Fig. S13).

Our study brings favorable insights into the fields of computational pathology and biomarker discovery, but there are still certain limitations. For example, our weakly supervised CPMP model leverages general-purpose foundation models[36–41] for feature extraction, which helps reduce computational costs. However, these foundation models were pretrained on large datasets, in which the data distribution may differ from ours due to variations in slide preparation, staining, and scanning techniques. As a result, one line of future work will focus on addressing these data domain distribution challenges by fine-tuning existing self-supervised models[47,48]. While this approach has the potential to enhance predictive performance, it is essential to carefully consider the computational resources required for such experiments. The trade-off between these two factors needs to be carefully evaluated. Additionally, our model was developed based on the TJMUCH-MP cohort. It can only predict MP High and Low risk groups for early-stage BC patients. Recent studies[49,50] suggest that it may be possible to further identify the Ultra-High (H2) subgroup patients

within the High-risk group who could benefit from immunotherapy. We found that the number of Ultra-High patients in the TJMUCH-MP cohort is limited (n=11), which constrains CPMP's ability to effectively learn and automatically identify the H2 subgroup. Future investigations should also focus on examining the intercellular relationships within the tumor ecosystem and the histological morphology patterns associated with MP recurrence risk groups, utilizing extensive data from multiple institutions.

Altogether, our study revealed that AI-driven weakly supervised computational pathology technology can effectively predict risk scores for prognostic distant metastasis hints using annotation-free histopathological WSIs. Our spatially-aware transformer-based CPMP model provided valuable insights into the close connections between spatial morphology in the tumor ecosystem and gene activities under molecular biology mechanisms. This capability can further facilitate prognostic knowledge discovery, paving the way for advancements in precision treatment. These findings make it possible to be a complementary, computational, and cost-efficient diagnostic tool for assessing recurrence risk for early-stage BC patients from routine H&E slides, ultimately supporting better clinical decision-making and patient management.

Methods

TJMUCH-MP cohort curation

This study examined female patients who were diagnosed with primary breast cancer and received treatment between January 2019 and December 2023 at TJMUCH, National Clinical Research Center for Cancer. Patients who underwent MP diagnostic tests were included in this retrospective analysis. Whole-slide images were scanned using the Aperio Leica Biosystems GT450 v1.0.1 scanner at an apparent magnification of 40×. The genomic MP test was conducted by ZhenHe Genecast Biotechnology Ltd, the exclusive partner for the 70-GS assay in China, as appointed by Agendia. All participants enrolled in this study met the StGallen guidelines for the MP test[29]. Inclusion criteria comprised invasive BC that was HR+/HER2-, staged T1-2, and either lymph node-negative or with 1-3 metastases. Additionally, tumor specimens were required for genomic analysis. Exclusion criteria included HR- or HER2+ tumors and patients receiving neoadjuvant therapy. Cases involving bilateral breast cancer and inconclusive tissue sections were also excluded. Clinical information and histopathological characteristics were extracted from medical records, while adverse events, such as recurrence and death, were tracked through follow-up visits. FFPE tumor samples were collected from biopsies and excised surgical materials. Based on the genomic MP diagnostic test, continuous risk scores were yielded, and patients were categorized into High-risk and Low-risk groups using a cutoff value of 0.0 (with scores $\leqslant$ 0.0 categorized as High-risk and scores > 0.0 as Low-risk)[7]. Ultimately, a Chinese HR+/HER2- invasive BC cohort was established, comprising digital WSIs, genomic MP results, and clinicopathological features of 477 female patients (Fig. 1a-d, Supplementary Fig. S1a-b, and Supplementary Tab. S1). This cohort was utilized for subsequent model development and evaluation.

WSIs preprocessing

Histopathological WSIs typically contain billions of pixels and possess high-resolution, multi-scale attributes. However, these images include areas of white background and non-tissue regions that should be filtered out. To improve the computational efficiency of WSI analysis, we first performed data preprocessing on the H&E histopathological slides (Supplementary Fig. S1c). We utilized automated tissue segmentation techniques proposed by CLAM[32] for effective tissue detection, resulting in segmented foreground contours. Following this, we employed a non-overlapping sliding

window tiling strategy to crop image tiles measuring 256×256 pixels from the identified tissue regions at a 20× magnification. The resulting stacks of image tiles, along with their spatial coordinates, were organized and stored using the Hierarchical Data Format version 5 (HDF5)[32]. The distribution of the number of tiles obtained from each WSI varies widely, ranging from thousands to hundreds of thousands, as illustrated in Supplementary Fig. S1d.

CPMP model

Our model employs the weakly supervised MIL framework for predicting recurrence risk from histopathological WSIs (Fig. 1e). To clarify, we begin with outlining the classic MIL approach. In this framework, a dataset $\mathbb{D} = \{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^{N}$ consists of a total of N bags $\mathcal{X}_i$, each containing multiple instances $\mathcal{X}_i = \{t_j^i\}_{j=1}^{n_i}$. The primary objective of the MIL approach is to develop a function $\mathcal{F}(\cdot)$ that can predict the bag-level category $\mathcal{Y}_i$ based on the multiple instances $\{t_j^i\}_{j=1}^{n_i}$ within that bag. While the bag-level labels $\{\mathcal{Y}_i\}$ are known, the labels for the individual instances $\{t_j^i\}$ are unknown. The function is typically decomposed into three key components: 1) An instance-level transformation function that transforms instances into feature vectors. 2) An aggregation function at the instance-level that integrates the features of all instances to yield a global feature representation at the bag-level. 3) A classifier head function that predicts the bag-level results from this global feature representation. In alignment with the MIL paradigm, CPMP is composed of three components: a feature representation function $\mathcal{T}(\cdot)$: operating at the tile-level, a feature aggregation function $\mathcal{S}(\cdot)$ that integrates features across tiles within a slide, and a regression-based MLP predictor head $\mathcal{R}(\cdot)$ at the slide-level. We describe these components in detail in the following section.

- **Feature representation at the tile-level**

Feature representation, serving as the instance-level transformation function in MIL, was executed on tiles cropped from WSIs. Recently, several self-supervised models[36–41] pretrained on extensive datasets of histopathological WSIs across pan-cancer pan-tissue types became publicly available, significantly facilitating the advances in the field of computational pathology. These general-purpose foundation models succeed in capturing the diverse patterns in pathology images and hold promise for integration into clinical diagnostical workflows, particularly in applications with limited data availability. We explored several models, including UNI[38], Prov-GigaPath[36], Virchow[37], H-optimus-0[39], Phikon[41], and CTransPath[40], for tile-level feature extraction (Supplementary Fig. S1e). We first

resized the 256×256 tiles to 224×224 pixels before inputting them into feature encoders for forward computation. This process allows us to represent each tile with low-dimensional feature vectors: 1024-dimensional (UNI encoder), 1536-dimensional (Prov-GigaPath), 2560-dimensional (Virchow), 1536-dimensional (H-optimus-0), and 768-dimensional (Phikon and CTransPath). Low-dimensional feature representations at the tile level make it feasible to reduce computational complexity and aggregate diverse patterns from all tiles within a slide.

- **Feature aggregation**

Based on the feature embeddings of local tiles, we employed the Transformer[51] architecture for feature aggregation to derive a global feature representation at the slide-level. The self-attention module in the Transformer is crucial, as it can capture long-range dependencies across tiles. However, the global self-attention mechanism incurs a high computational cost, particularly when the number of tiles in a single slide can reach up to 100,000. This excessive computation and redundancy limits the applicability of the self-attention mechanism in our high-resolution scenarios. To this end, we implemented an agent-attention mechanism[30] to integrate the embedding sequences by introducing an agent token (Fig. 1f, g). This modification transforms the traditional self-attention $(\mathcal{Q},\mathcal{K},\mathcal{V})$ into the agent-attention paradigm $(\mathcal{A},\mathcal{Q},\mathcal{K},\mathcal{V})$. balancing between the need for global representation capabilities and the requirement for computational efficiency.

Specifically, for the input tile features $x = \{\mathcal{T}(t_j)\}_{j=1}^{n}$, $x \in \mathbb{R}^{n\times c}$, the general self-attention module can be formulated as follows:

$$\begin{aligned} \mathcal{Q} &= w_Q(x) \\ \mathcal{K} &= w_K(x) \\ \mathcal{V} &= w_V(x) \end{aligned} \tag{1}$$

$$SAttn(\mathcal{Q}_i,\mathcal{K},\mathcal{V}) = \sum_{m=1}^{n} \frac{Sim(\mathcal{Q}_i,\mathcal{K}_m)}{\sum_{j=1}^{n} Sim\left(\mathcal{Q}_i,\mathcal{K}_j\right)} \mathcal{V}_m \tag{2}$$

where n and c are the number of tile embeddings in one slide and the dimension of tile embedding, respectively. $w_Q, w_K, w_V \in \mathbb{R}^{c\times d}$ denote respective linear transformation layers for Query, Key, and Value in the Transformer block, where d is the output dimension of linear layers. $Sim(\cdot)$ represents the similarity function, commonly $Sim(\mathcal{Q},\mathcal{K}) = exp\left(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d}}\right)$ in the Softmax attention. We abbreviate the similarity function to $\sigma(\cdot)$ and the Softmax attention can be rephrased as follows:

$$SAttn(\mathcal{Q},\mathcal{K},\mathcal{V}) \triangleq \sigma(\mathcal{Q}\mathcal{K}^T)\mathcal{V} \tag{3}$$

Softmax attention computes a query-key product, resulting in the computation complexity of $\mathcal{O}(n^2)$. It is noteworthy that we do not consider multi-heads here.

In the agent-attention paradigm, we added an agent token $\mathcal{A}$ to the foregoing attention triplet $(\mathcal{Q}, \mathcal{K}, \mathcal{V})$, This modification results in the agent-attention paradigm of $(\mathcal{A}, \mathcal{Q}, \mathcal{K}, \mathcal{V})$, which we denote as $AAttn(\mathcal{A}, \mathcal{Q}, \mathcal{K}, \mathcal{V})$. To be specific, the agent-attention paradigm consists of a stacked state of two Softmax attention operations $SAttn(\cdot)$. In the first phase, referred to as 1st-aggregation, the agent token $\mathcal{A}$ is treated as a Query $\mathcal{Q}$ in the first Softmax attention operation $SAttn(\mathcal{A}, \mathcal{K}, \mathcal{V})$. Information is aggregated into the Agent-Value $\mathcal{V}_{agent}$ from the Value $\mathcal{V}$, with the attention matrix computed based on Agent $\mathcal{A}$ and Key $\mathcal{K}$. The second Softmax attention operation $SAttn(\mathcal{Q}, \mathcal{A}, \mathcal{V}_{agent})$ is applied in the second phase, known as 2nd-broadcast. Here, the Agent-Value $\mathcal{V}_{agent}$ serves as the output from the first Softmax attention. This process yields the final output from the agent-attention mechanism (Fig. 1g).

$$V_{agent} = SAttn(\mathcal{A}, \mathcal{K}, \mathcal{V}) \tag{4}$$

$$AAttn(\mathcal{A}, \mathcal{Q}, \mathcal{K}, \mathcal{V}) = SAttn(\mathcal{Q}, \mathcal{A}, \mathcal{V}_{agent}) = \underbrace{SAttn\,(Q, A, \underbrace{SAttn(\mathcal{A}, \mathcal{K}, \mathcal{V})}_{\text{1st-aggregation}})}_{\text{2nd-broadcast}} \tag{5}$$

The pairwise Query-Key product computation is replaced by the Agent-Key and Query-Agent product, and the information flow is connected and preserved through the Agent. Notably, the number of the agent token k is a very small constant, which theoretically reduces the computational complexity of the agent-attention to $\mathcal{O}(nd)$, where $d \ll n$.

Using the agent attention mechanism described above, we employed the agent-attention transformer for instance aggregation (Fig. 1f). We first transformed n tile embeddings into a dimension of 256 through a linear projection layer, followed by the non-linear activation function ReLU. To incorporate the spatial positional information of the tiles, we employed a 2-dimensional absolute position encoding method[51,52] to encode the tile coordinates. The resulting 256-dimensional positional embeddings were then added to the 256-dimensional tile embeddings. Next, a class token $x_{cls} \in \mathbb{R}^{1\times 256}$ was initialized and concatenated with the spatially-aware tile embeddings, yielding a fused representation $\{x, x_{cls}\} \in \mathbb{R}^{(n+1)\times 256}$. We configured an agent-attention transformer block consisting of one layer with four heads for global information perception and aggregation. The output dimensions of the latent projection layers for Query, Key, and Value were set to 256, while

the Feed Forward hidden linear projection layers produced an output dimension of 512.

- **A regression-based MLP predictor head**

Based on the agent-attention transformer block, the class token x_{cls} was utilized as the aggregated global representation at the slide-level. This token was then passed into the MLP predictor head, which included a LayerNorm normalization followed by a linear layer. Given that the patient's MP risk is represented as a continuous score, binarizing these values may result in information loss. Recent studies[53,54] provide evidence that regression-based approaches are superior to classification-based methods for predicting continuous biomarkers. Therefore, we adopted a regression-based strategy to output a continuous numerical value as the risk prediction result. The linear projection layer transformed the feature embedding of the class token x_{cls} into a single-neuron logit output, which was then passed through a Sigmoid activation function to yield a continuous risk value between 0 and 1. We employed the mean square error loss as the overall objective function $\mathcal{L}(\cdot)$ to optimize CPMP.

$$\mathcal{L}(\cdot) = \frac{1}{B}\sum_{i=1}^{B}\left[\lambda_1 \cdot \left(Y_i - \hat{Y}_i\right)^2 + \lambda_2 \cdot \left(\mathbb{1}_{>0.5}(Y_i) - \hat{Y}_i\right)^2\right] \tag{6}$$

$$\{\hat{\theta}_S\} \leftarrow \arg min_{\{\theta_S\}}\mathcal{L}(\cdot) \tag{7}$$

where B is the batch size during training, Y_i represents the slide-level continuous MP risk score, and $\hat{Y}_i$ signifies the predicted continuous risk probability value for a patient. The parameters λ_1 and λ_2 are scale factors that balance the contributions of different loss components in the objective function. The indicator function $\mathbb{1}_{>0.5}(\cdot)$ returns 1 when $Y_i > 0.5$, and 0 when $Y_i \leq 0.5$. The optimization process is represented as $\arg min_{\{\theta_S\}}\mathcal{L}(\cdot)$, indicating that the optimization algorithm seeks to find the optimized parameters $\{\theta_S\}$ which minimize the objective loss $\mathcal{L}(\cdot)$. Here, $\{\theta_S\}$ refers to the set of optimized parameters in our model.

Spatial attention heatmap visualization at the whole-slide level

The whole-slide-level attention heatmap visualization was conducted to explore spatial localization in tumors. It was implemented by mapping attention scores of tiles to their corresponding locations within WSIs. Specifically, we first performed data preprocessing on each WSI to detect tissue areas. The identified foreground tissue regions were tessellated into 256×256 tiles with an overlapped rate of 0.1 at a 20× objective magnification. Subsequently, we leveraged the self-supervised foundation

model UNI for feature extraction from these tiles, resulting in 1024-dimensional feature vectors. The extracted tile embeddings were then passed through CPMP for inference, producing slide-level prediction results alongside tile-level attention scores that contributed to the final risk prediction. The attention scores at the tile-level were calculated using the attention rollout method[55], based on our agent-attention matrix $M_{AAttn}(\cdot)$ between tiles.

$$M_{AAttn}(\mathcal{A},\mathcal{Q},\mathcal{K}) = \sigma\left(\mathcal{Q}\mathcal{A}^T\right)\cdot\sigma\left(\mathcal{A}\mathcal{K}^T\right) \tag{8}$$

These attention scores were normalized to a range between 0 and 1, serving as attention weights. Finally, the normalized attention scores were arranged according to their corresponding spatial coordinates within the WSI, allowing for the visualization of an attention heatmap. This heatmap can be optionally overlaid on the original H&E-stained WSI, providing insights into the spatial distribution of attention that aligns with tumor areas.

Cell identification and intercellular graph construction

To gain further insights into the cellular interactions within the tumor ecosystem of MP High and Low risk groups, we first extracted the high-attention ROIs highlighted by the spatial attention heatmap at the whole-slide level. Subsequently, we applied HoverNet[42] for simultaneous segmentation and classification of nuclei. We utilized the pretrained model weights trained on the public PanNuke dataset. The identified cells were classified into one of five cell categories: neoplastic (tumor), inflammatory, necrotic (dead), connective (stroma), and non-neoplastic epithelial cells. The spatial coordinates of the centroids and contours of these cells were located in the ROIs (illustrated in Supplementary Fig. S6a).

We defined the density of tumor cells (TcD) as the number of identified neoplastic cells per square millimeter. To further analyze the tumor ecosystem, we selected the three main cellular components[43] (tumor, inflammatory, and stromal cells) to construct topological graphs among the detected cell nuclei. The sc-MTOP framework[44] was employed for profiling graph-based features at the single-cell level within the tumor microenvironment. We focused on the topological features of tumor cells interacting with other cells, specifically examined the connectivity graphs of tumor-tumor, tumor-inflammatory and tumor-stroma interactions. Topological features related to graph-edge length, especially MeanEdgeLength[44], were calculated to characterize the spatial intercellular relationships among different cell types within tumors. This metric represents the average edge

lengths of each tumor cell interacting with other tumor, stromal, and inflammatory cells (illustrated in Supplementary Fig. S6b).

Characterization of phenotypic diversity

To investigate phenotypes potentially associated with MP recurrence risk groups, we first chose the model demonstrating optimal performance from the experiments based on a patient-level 5-Time 5-Fold cross-validation scheme. The histopathological WSIs of the testing cases were then tessellated into 256×256 tiles, which were characterized using the UNI foundation model for feature representation. These tiles were subsequently fed into the CPMP model for inference, yielding WSI-level risk scores. In addition, we identified and saved the top 100 tiles with the highest attention scores from the slide, which were deemed representative of local regions within the tumor. Each instance in this tile collection was also processed through CPMP to infer tile-level risk scores. The intermediate vector (cls_token) between the transformer block and the MLP classifier was detached, resulting in a 256-dimensional embedding vector for each tile instance.

The Scanpy toolkit was employed for further analysis of the embedding data. All tile embeddings were combined into an AnnData object, which contained a data matrix of the tile embeddings along with the additional metadata. The metadata included the spatial coordinates of the tiles within WSIs, genomic MP risk results corresponding to the slide from which the tile originated, WSI-level risk scores, and tile-level risk scores from the CPMP model. The nearest neighbors distance matrix and a neighborhood graph of tile embeddings were computed using the scanpy.pp.neighbors() function, employing cosine similarity as the distance measurement metric. The minimum size for local neighborhoods was set to 50 to balance between local information and global views. Based on the constructed neighborhood graph, Leiden clustering implemented by the scanpy.tl.leiden() function was performed for subcluster community detection on tile embeddings. Finally, we applied UMAP dimensionality reduction to visualize the tile embeddings and their subclusters in two-dimensional space via the scanpy.tl.umap() function. The low dimensional embedding was initialized using coarse-grained connectivity structures generated by the scanpy.tl.paga() function, which employs the partition-based graph abstraction (PAGA) algorithm.

Based on the Leiden clustering on the tile embeddings, we identified subclusters representing local morphological patterns. The following analytic steps were undertaken for these subclusters. First, we extracted the respective submaps for Low-risk and High-risk groups from the global

subclusters, based on the MP risk group of the WSI from which the tile originated. We only preserved those subclusters whose number of tiles exceeds 25% of that in the corresponding global subcluster in specific risk group. Next, we determined the intersection of subcluster ID sets from both the High- and Low-risk groups to identify subclusters that present in both MP risk categories, termed as colocalized subclusters. Subcluster IDs exclusive to the High-risk group were identified as High-specific subclusters, while those unique to the Low-risk group were designated as Low-specific subclusters. Subsequently, we applied HoverNet to all tiles within the High-specific, Low-specific, and colocalized subclusters for nuclei segmentation and classification. This process identified five cell types (neoplastic, inflammatory, necrotic, connective, and non-neoplastic epithelial cells), as well as their corresponding spatial coordinates within the tiles. Finally, we quantified the cellular composition for each subcluster by calculating the average number of cells for each cell type across all tiles within the subcluster. We then applied Min-max normalization to standardize these quantitative features of the cellular composition and conducted hierarchical clustering on all subclusters using the Ward variance minimization algorithm and the Euclidean distance metric.

Prognostic analysis on TCGA-BRCA

We applied the CPMP model to the external TCGA-BRCA cohort for recurrence risk assessment and prognostic analysis. TCGA-BRCA cohort includes 1,056 patients with digital diagnostic slides as well as clinical follow-up information (Supplementary Fig. S9a). All digital slides (n = 1,133) in the TCGA-BRCA cohort were first preprocessed and tessellated into local tiles. Tile-level feature representations were generated using UNI, and they were fed into the trained CPMP model for slide-level distant metastasis risk score inference. The prediction results of multiple slides from the same patient were averaged into the patient-level risk values. The continuous prediction risk scores were categorized into High-risk and Low-risk groups based on a cutoff value determined by the trained model. Spatial attention heatmap inference at the whole-slide level was also performed to visualize how regions with distinct tumor morphological patterns contribute to patient clinical outcomes. The prognostic value of the CPMP model within the TCGA-BRCA cohort was assessed using the Kaplan-Meier survival curves and the multivariable Cox proportional hazards model. The clinical outcome endpoints including Disease-Free Interval, Progression-Free Interval, and Overall Survival were derived from the TCGA Clinical Data Resource[46].

Comparative methods

We compared the CPMP model with existing state-of-the-art weakly supervised MIL methods in WSIs classification. Brief introductions of the comparative methods are summarized as follows:

1. ABMIL[31], Attention-based MIL method, currently a MIL baseline for WSIs classification in the field of CPATH.
2. CLAM[32], an instance-level clustering-constrained-attention MIL algorithm for pathological subtyping. It relied on instance-level clustering to refine the feature space in MIL.
3. The full Transformer-based architecture model applied for end-to-end biomarker prediction from WSIs[33] (denoted as Wagner et al.). We halved the architecture to a smaller scale with 4 attention heads due to data limitation in our experiments.
4. TransMIL[34], a correlated Transformer-based MIL framework for pathology diagnosis on WSIs. It designed a PPEG for position encoding to capture both the locally contextual information and the correlation between regions.

To verify the effectiveness of CPMP in using annotation-free WSIs, we conducted the experiments on noisy tissue filtering, tile sampling, and clinical-features-only logistic regression. In the experiment of tissue filtering, we aimed to investigate CPMP's sensitivity to noisy tiles. We performed zero-shot classification to determine tissue types for all tiles using the visual-language foundation model PLIP[35] (illustrated in Supplementary Fig. S3a). Tiles were identified as one of eight tissue types: tumor, adipose, stroma, immune infiltrates, gland, necrosis or hemorrhage, background or black, and non (Supplementary Fig. S3b). Tiles classified as “background or black” and “non-tissue” types were filtered out. In contrast, tiles identified as tumor, adipose, stroma, immune infiltrates, gland, and necrosis or hemorrhage tissue types were preserved and utilized for training the other model. This model with noisy tile filtering was compared with the CPMP model. To evaluate the scale flexibility of CPMP on the number of local tiles in WSIs, we conducted random sampling from the set of local tiles in each WSI, employing sample sizes of 1,000, 5,000, and 10,000 tiles as subsets for model training. The trained models were further evaluated on multiple subsets of tiles with variant sizes.

A clinical-based logistic regression model designed to predict patients' MP recurrence risk from clinicopathological features was experimented for comparison. This clinical-based model was constructed using a Logistic Regression classifier, incorporating a set of eight clinical variables: 'Surgery Option', 'T', 'N', 'AJCC', 'Histological grading', 'TILs', 'PR', and 'Ki67'. The clinical-based

model underwent training, validation, and independent testing on the same patient cases, utilizing a patient-level 5-Time 5-Fold cross-validation strategy, consistent with the approach used in the CPMP model. Detailed hyperparameter settings for the Logistic Regression classifier are provided in the Supplementary Notes. Among the eight clinical variables, 'T', 'N', 'AJCC', and 'Histological grading' are categorical variables while 'TILs', 'PR', and 'Ki67' are continuous variables representing intensity percentages or positive proportions, with values ranging from 0 to 100. The 'Surgery Option' variable was categorized by assigning numerical values to different surgical options, with the specific categorization rules outlined in Supplementary Tab. S5.

We also performed ablation studies to analyze modules in our spatially-aware, regression-based weakly supervised transformer architecture. The design outline and the experimental configurations utilized for the ablation models were illustrated in Supplementary Fig. S4a. Specifically, we first analyzed the transformer-based aggregation function with self-attention or agent-attention mechanisms. Regression-based mean square error (MSE) loss strategy was compared with classification-based cross-entropy (CE) loss. Positional coordinate information embedding (PE) was introduced into tile embeddings for spatial awareness in the CPMP model. PE, No-PE, and PPEG-based experimental configurations were thus adopted for comparison. In the ablation study of tile-level feature representation, we explored six self-supervised general-purpose foundation models, UNI[38], Prov-GigaPath[36], Virchow[37], H-optimus-0[39], Phikon[41], and CTransPath[40], for tile-level feature extraction and experimental comparison.

Implementation details

CPMP was implemented using Python 3.8 and the PyTorch 2.2 deep learning framework. All models, including comparative methods, were conducted on one workstation equipped with two 24GB NVIDIA GPUs (GeForce RTX 4090). The CUDA version is 12.2 and the GPU driver version is 535.171. Adam algorithm was employed for parameters optimization. The momentum factor was 0.9, and the learning rate was initially set to 0.0001. LinearLR scheduler aimed to the learning rate decays was operated and set to a limit with 0.00001 when the number reaches epoch 500. The maximum number of epochs was set to 1000, and an early stopping strategy was utilized. The early stop is determined when the loss value of the validation set no longer decreases for 50 consecutive epochs. A gradient accumulation strategy was also adopted to address the issue of an inconsistent number of patches across WSIs. The Gradient accumulation size was set to 32, which is equivalent to

using a minibatch size of 32. The weights of all layers were initialized with the Kaiming uniform strategy. The optimized model with the lowest loss metric value on the validation set was recorded and utilized for the following evaluation on the test set. WSIs were processed by OpenSlide Toolkit v3.4.1.

For fair evaluation in case of data bias, we employed a patient-level 5-Time 5-Fold cross-validation scheme in the TJMUCH-MP cohort. In the experiment of each Time, we randomly divided 20% of the samples for testing, while the remaining samples were evenly divided into 5 non-overlapping subsets (Folds). Slides from each patient were only assigned to one of these sets to prevent data leakage. We then iteratively trained the model on 4 subsets (4 Folds) while using the remaining subset for validation. This process was repeated for each Fold, ensuring that each Fold served as the validation subset exactly once. The trained models corresponding to the 5 Folds were evaluated on the independent testing cases, and the results were averaged. This experiment was repeated for 5 random Times and the average results over the 5-Time experiments are reported. The cross-validation strategy is illustrated in Fig. 2a.

Statistical analysis

All statistical analyses were performed using Python 3.8 and R 4.2.2. The nonparametric Spearman rank-order correlation coefficient (Spearman's r) and corresponding *p*-values were calculated to measure the relationship between two continuous variables. Spearman's r correlation coefficient varies between -1 and +1 with 0 indicating no correlation. Correlations of +1 or -1 imply an exact positive or negative relationship. Multiple metrics, including the ROC analysis with AUC value, the PRC with the AUPRC value, the balanced accuracy, as well as the Precision, Recall, and F1 score were employed to evaluated the performance of our approach and comparative methods on MP risk groups. Continuous variables were compared between independent groups using the nonparametric Mann-Whitney U rank test. Survival differences were compared between groups using the log-rank test and Cox proportional hazards models. All tests were two-sided, and a p-value < 0.05 was considered statistically significant.

Data availability

The TCGA data (slides and clinicopathological information) used in this study are available at https://portal.gdc.cancer.gov/. The follow-up data are available at TCGA Clinical Data Resource[46].

The TJMUCH-MP cohort is accessible upon request. The data can only be used under the condition that the request is for non-profit, purely academic research purposes, and the requesting researchers must provide valid ethics approval from their institution. The data generated in this study for the creation of the figures are provided in the Source Data file.

Code availability

The source codes of CPMP are provided in Supplementary Software and available under an open-source license at https://github.com/lyotvincent/CPMP.

Author contributions

H.L. curated the clinical cohort and designed the clinical investigation. J.L. conceived and designed the study. C.Y. developed the approach and conducted the experiments. L.L. and X.Q. assisted with clinical analysis of tumor spatial location experiment results. L.L., X.Q., Y.O., and Z.H. collected the histopathological slides and preprocessed the clinical data. Z.R., and W.X. assisted with image preprocessing and ablation experiments. C.Y. and J.L. prepared the manuscript with input from all authors. All authors read and approved the final manuscript.

Conflict of interest statement

No competing interest is declared.

Role of funding source

This work was supported by the National Key Research and Development Program of China (2020YFA0908700 and 2020YFA0908702) and the National Natural Science Foundation of China (62272246).

Ethics approval

This retrospective study was approved by the Ethics Committee of Tianjin medical university cancer institute and hospital (bc20240004). Informed consent was obtained from all participants in the cohort for all data acquisition. For TCGA, no formal ethics approval was required for a retrospective study of anonymous samples.

References

1. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **74**, 229–263 (2024).
2. Pedersen, R. N. *et al.* The Incidence of Breast Cancer Recurrence 10-32 Years After Primary Diagnosis. *JNCI: Journal of the National Cancer Institute* **114**, 391–399 (2022).
3. Kim, M. Y. Breast Cancer Metastasis. in *Translational Research in Breast Cancer* (eds. Noh, D.-Y., Han, W. & Toi, M.) 183–204 (Springer, Singapore, 2021). doi:10.1007/978-981-32-9620-6_9.
4. Giuliano, A. E. *et al.* Breast Cancer—Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. doi:10.3322/caac.21393.
5. van ’t Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
6. Glas, A. M. *et al.* Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics* **7**, 278 (2006).
7. Cardoso, F. *et al.* 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *New England Journal of Medicine* **375**, 717–729 (2016).
8. Piccart, M. *et al.* 70-gene signature as an aid for treatment decisions in early breast cancer: updated results of the phase 3 randomised MINDACT trial with an exploratory analysis by age. *The Lancet Oncology* **22**, 476–488 (2021).
9. Brandão, M., Pondé, N. & Piccart-Gebhart, M. Mammaprint™: a comprehensive review. *Future Oncology* **15**, 207–224 (2019).
10. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat Cancer* **3**, 1026–1038 (2022).
11. Bahadir, C. D. *et al.* Artificial intelligence applications in histopathology. *Nat Rev Electr Eng* **1**, 93–108 (2024).
12. Song, A. H. *et al.* Artificial intelligence for digital and computational pathology. *Nat Rev Bioeng* 1–20 (2023) doi:10.1038/s44222-023-00096-8.
13. Hosseini, M. S. *et al.* Computational Pathology: A Survey Review and The Way Forward. Preprint at https://doi.org/10.48550/arXiv.2304.05482 (2023).
14. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer* **1**, 800–810 (2020).
15. Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer* **1**, 789–799 (2020).
16. Tsai, P.-C. *et al.* Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nat Commun* **14**, 2102 (2023).
17. Ding, K., Zhou, M., Wang, H., Zhang, S. & Metaxas, D. N. Spatially aware graph neural networks and cross-level molecular profile prediction in colon cancer histopathology: a retrospective multi-cohort study. *The Lancet Digital Health* **4**, e787–e795 (2022).
18. Saillard, C. *et al.* Pacpaint: a histology-based deep learning model uncovers the extensive intratumor molecular heterogeneity of pancreatic adenocarcinoma. *Nat Commun* **14**, 3459 (2023).
19. Jin, D. *et al.* Teacher-student collaborated multiple instance learning for pan-cancer PDL1 expression

prediction from histopathology slides. *Nat Commun* **15**, 3063 (2024).

20. Juan Ramon, A. *et al.* Development and deployment of a histopathology-based deep learning algorithm for patient prescreening in a clinical trial. *Nat Commun* **15**, 4690 (2024).
21. Lazard, T. *et al.* Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images. *CR Med* **3**, (2022).
22. Mandair, D., Reis-Filho, J. S. & Ashworth, A. Biological insights and novel biomarker discovery through deep learning approaches in breast cancer histopathology. *npj Breast Cancer* **9**, 1–11 (2023).
23. Jaber, M. I. *et al.* A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Research* **22**, 12 (2020).
24. Yan, C. *et al.* PhiHER2: phenotype-informed weakly supervised model for HER2 status prediction from pathological images. *Bioinformatics* **40**, i79–i90 (2024).
25. Boehm, K. M. *et al.* Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. *Nat Commun* **16**, 2106 (2025).
26. Wahab, N. *et al.* AI-enabled routine H&E image based prognostic marker for early-stage luminal breast cancer. *npj Precis. Onc.* **7**, 1–13 (2023).
27. Chen, Y. *et al.* Computational pathology improves risk stratification of a multi-gene assay for early stage ER+ breast cancer. *npj Breast Cancer* **9**, 1–10 (2023).
28. Howard, F. M. *et al.* Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *npj Breast Cancer* **9**, 1–6 (2023).
29. Goldhirsch, A. *et al.* Progress and promise: highlights of the international expert consensus on the primary therapy of early breast cancer 2007. *Annals of Oncology* **18**, 1133–1144 (2007).
30. Han, D. *et al.* Agent Attention: On the Integration of Softmax and Linear Attention. Preprint at http://arxiv.org/abs/2312.08874 (2023).
31. Ilse, M., Tomczak, J. & Welling, M. Attention-based Deep Multiple Instance Learning. in *Proceedings of the 35th International Conference on Machine Learning* 2127–2136 (PMLR, 2018).
32. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng* **5**, 555–570 (2021).
33. Wagner, S. J. *et al.* Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell* **41**, 1650-1661.e4 (2023).
34. Shao, Z. *et al.* TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. in *Advances in Neural Information Processing Systems* vol. 34 2136–2147 (Curran Associates, Inc., 2021).
35. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical Twitter. *Nat Med* **29**, 2307–2316 (2023).
36. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* 1–8 (2024) doi:10.1038/s41586-024-07441-w.
37. Vorontsov, E. *et al.* A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat Med* 1–12 (2024) doi:10.1038/s41591-024-03141-0.
38. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat Med* **30**, 850–

862 (2024).

39. releases/models/h-optimus/v0 at main · bioptimus/releases. *GitHub* https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.
40. Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis* **81**, 102559 (2022).
41. Filiot, A. *et al.* Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. Preprint at https://doi.org/10.1101/2023.07.21.23292757 (2023).
42. Graham, S. *et al.* Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis* **58**, 101563 (2019).
43. Wagner, J. *et al.* A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell* **177**, 1330-1345.e18 (2019).
44. Zhao, S. *et al.* scMTOP: Single-cell morphological and topological atlas reveals the ecosystem diversity of human breast cancer. *Nat Commun* **14**, 6796 (2023).
45. Lüönd, F., Tiede, S. & Christofori, G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *Br J Cancer* **125**, 164–175 (2021).
46. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).
47. Oquab, M. *et al.* DINOv2: Learning Robust Visual Features without Supervision. Preprint at https://doi.org/10.48550/arXiv.2304.07193 (2023).
48. Darcet, T., Oquab, M., Mairal, J. & Bojanowski, P. Vision Transformers Need Registers. Preprint at https://doi.org/10.48550/arXiv.2309.16588 (2024).
49. Cobain, E. F. *et al.* Elucidating the immune active state of HR+HER2- MammaPrint High 2 early breast cancer. *JCO* **42**, 506–506 (2024).
50. Pusztai, L. *et al.* Durvalumab with olaparib and paclitaxel for high-risk HER2-negative stage II/III breast cancer: Results from the adaptively randomized I-SPY2 trial. *Cancer Cell* **39**, 989-998.e5 (2021).
51. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2017).
52. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2021).
53. El Nahhas, O. S. M. *et al.* Regression-based Deep-Learning predicts molecular biomarkers from pathology slides. *Nat Commun* **15**, 1253 (2024).
54. Hoang, D.-T. *et al.* A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. *Nat Cancer* 1–13 (2024) doi:10.1038/s43018-024-00793-2.
55. Abnar, S. & Zuidema, W. Quantifying Attention Flow in Transformers. Preprint at https://doi.org/10.48550/arXiv.2005.00928 (2020).

Figures: Fig. 1 to Fig. 7

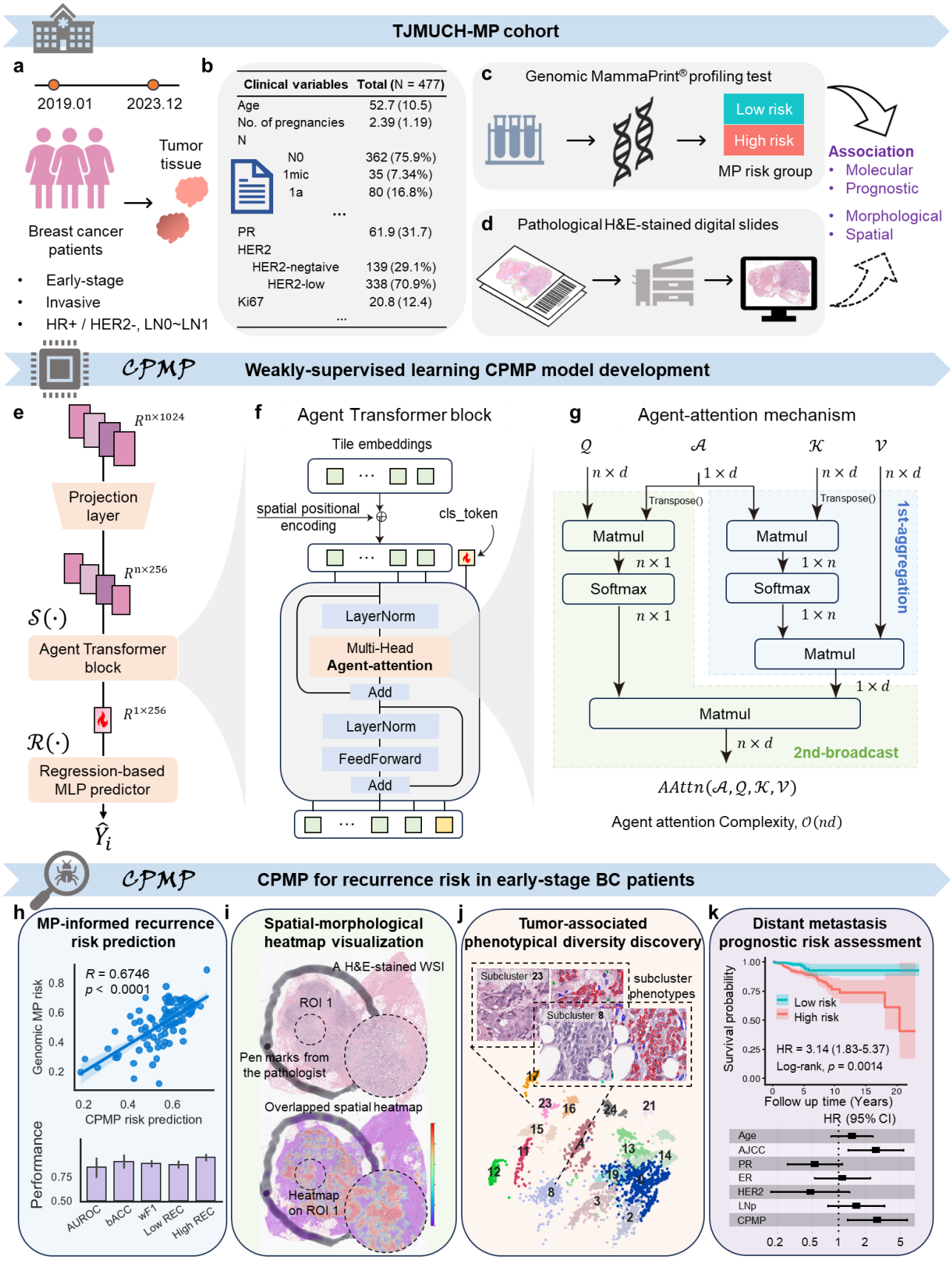

Fig. 1. Overview of CPMP and study design.

a. Data curation criteria for the clinical TJMUCH-MP cohort. The patients in the cohort were diagnosed with and treated for HR+/HER2- primary BC between January 2019 and December 2023 at Tianjin Medical University Cancer Institute and Hospital. **b-d**. The TJMUCH-MP cohort was comprised of clinicopathologic features (**b**), genomic MP test risk results (**c**) and pathological digital WSIs (**d**) of 477 patients. **e.** Depiction of proposed weakly-supervised transformer architecture CPMP based on multi-instance learning paradigm. It first performed linear projection on the feature embeddings of local tiles. Then it employed the Agent Transformer architecture for feature aggregation to obtain a slide-level global feature representation, followed by a regression-based MLP predictor head that predicts the slide-level results for the case. **f.** The Agent Transformer block adopted in the CPMP model. We introduced a 2-dimensional absolute coordinate encoding method to incorporate the spatial positional information and added the resulting positional embeddings to the projected tile embeddings. Then we initialized a class token (cls_token) and concatenated it with the spatial-aware tile embeddings. This Agent Transformer block consists of 1 layer with 4 heads for global information aggregation. **g.** The Agent-attention mechanism utilized in the Agent Transformer block. The Agent-attention paradigm is a stack state of two Softmax attention operations, including the first phase of 1st-aggregation (colored in light blue) and the second phase of 2nd-broadcast (colored in light green). **h-k.** Flowchart of the study design: (**h**) MP-informed recurrence risk prediction, (**i**) Spatial-morphological attention heatmap visualization, (**j**) Tumor-associated phenotypical diversity discovery, and (**k**) Distant metastasis prognostic risk assessment. These also summarize the clinical value of CPMP. BC=Breast Cancer, MP=MammaPrint, HR+ =Hormone Receptor positive, HER2- =Human Epidermal growth factor Receptor 2 negative, LN0=Lymph Nodes negative, LN1=The number of Lymph Nodes positive $\leqslant 3$, MLP= MultiLayer Perceptron.

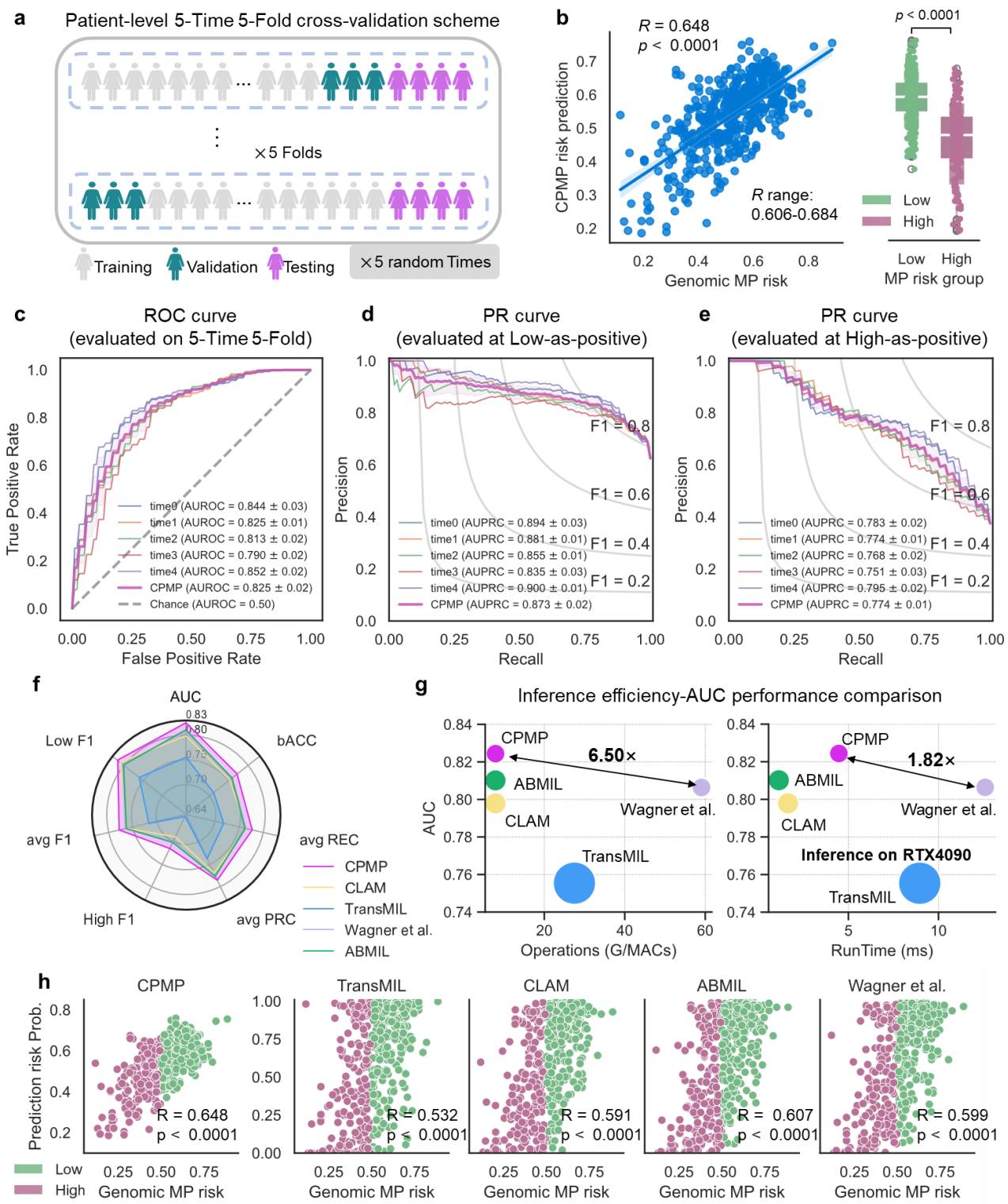

Fig. 2. Performance overview of CPMP and its comparison with other methods in predicting MP recurrence risk.

a. Schematic diagram of the patient-level 5-Time 5-Fold cross-validation strategy. It was employed to mitigate data bias and underly scanning effects. **b.** Correlation between genomic MP risk score value and predicted risk probability from CPMP for each patient in the test set. Fitted linear regression line is also shown. The Spearman

correlation analysis was performed. **c.** The ROC curves of the CPMP model for MP recurrence risk group prediction on the test sets across 5-Time experiments. **d**, **e.** The PR curves of the CPMP model for MP recurrence risk group prediction on test sets across 5-Time experiments. The model was evaluated in the modes of Low-as-positive (**d**) and High-as-positive (**e**). Low-as-positive is the condition that we aim to focus on Low-risk patients to avoid overtreatment. High-as-positive is to observe the High-risk patients who may be at risk from recurrence. **f**. Comparison of results of CPMP, CLAM, TransMIL, Wagner et al., and ABMIL models for MP recurrence risk prediction in terms of systematic metrics of AUROC, balanced accuracy, weighted average recall, precision, and F1 scores. **g.** Inference Efficiency-AUC performance for CPMP and comparative methods. The horizontal axis represents the efficiency evaluation index (Operations and Runtime), and the vertical axis is the AUC index, while the number of parameters for each model is denoted by the size of dots. We measured the computational operation and runtime cost of inference on a workstation with two RTX 4090 GPUs. **h.** Correlations between molecular MP risk values and recurrence risk prediction probabilities of patients in the test set for comparative methods. The patient points are separated into High-risk (colored in Lilac) and Low-risk (colored in Spring Green) groups. The Spearman correlation analysis was performed. ROC=Receiver Operating Characteristic, PR=Precision-Recall, AUC=Area Under the receiver operating Characteristic.

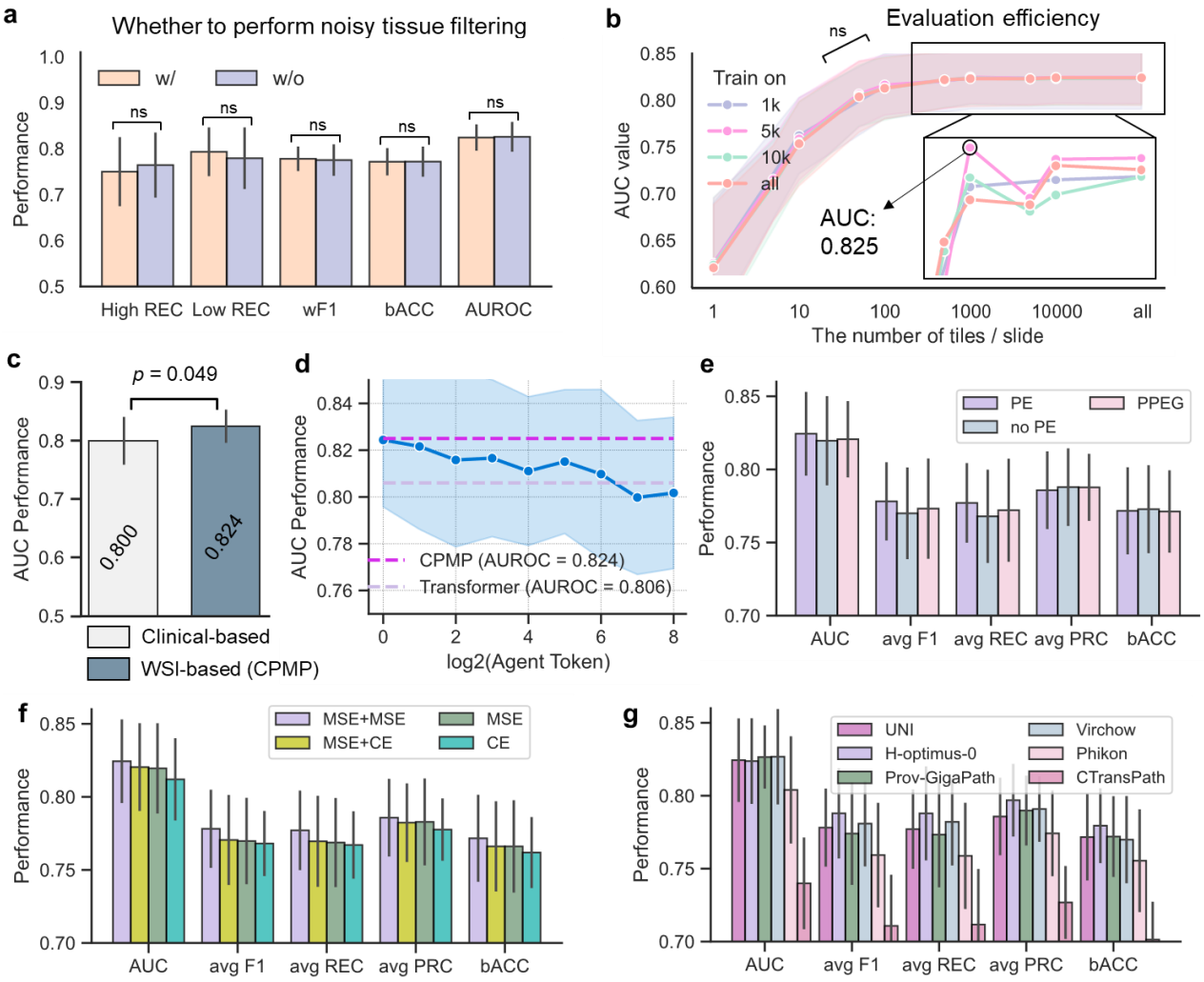

Fig. 3. Effectiveness overview of CPMP in leveraging annotation-free WSIs.

a. Comparison of performance of CPMP and the model with noisy tissue-type filtering in terms of 5 evaluation metrics. The Wilcoxon signed-rank test analysis was performed. **b.** AUROC values of models trained on different sets with the specific number of tiles random sampled from each slide for evaluation. The AUROC values were averaged across 5-Time experiments. **c.** Comparison of AUROC performance of CPMP and the clinical-based logistic regression model trained on clinicopathological features only. The Wilcoxon signed-rank test analysis was performed. **d.** AUROC performance trend evaluation of the agent-transformer models with variant number of the agent tokens. The light blue shading represents the standard deviation interval of performance across 5-Time 5-Fold experiments while the dashes are AUROC performance lines of CPMP and self-attention transformer. **e.** Performance evaluation of the experimental models equipped with different positional encoding configurations including PE, No-PE, and PPEG. **f.** Performance evaluation of the experimental models equipped with different loss functions. CE: Classification-based Cross-Entropy loss, MSE: Regression-based Mean-Square-Error loss.

MSE+CE: Regression-based MSE loss and CE loss with soft label, MSE+MSE: Regression-based dual MSE loss strategy. **g.** Performance evaluation of the experimental models equipped with different feature extraction foundation models including UNI, Virchow, H-optimus-0, Phikon, Prov-GigaPath, and CTransPath. w/=With, w/o=Without, ns=Statistically no significance, AUC=Area Under the receiver operating Characteristic, PE=Positional coordinates information Embeddings, PPEG=Pyramid Position Encoding Generator, CE=Cross-Entropy, MSE=Mean Square Error.

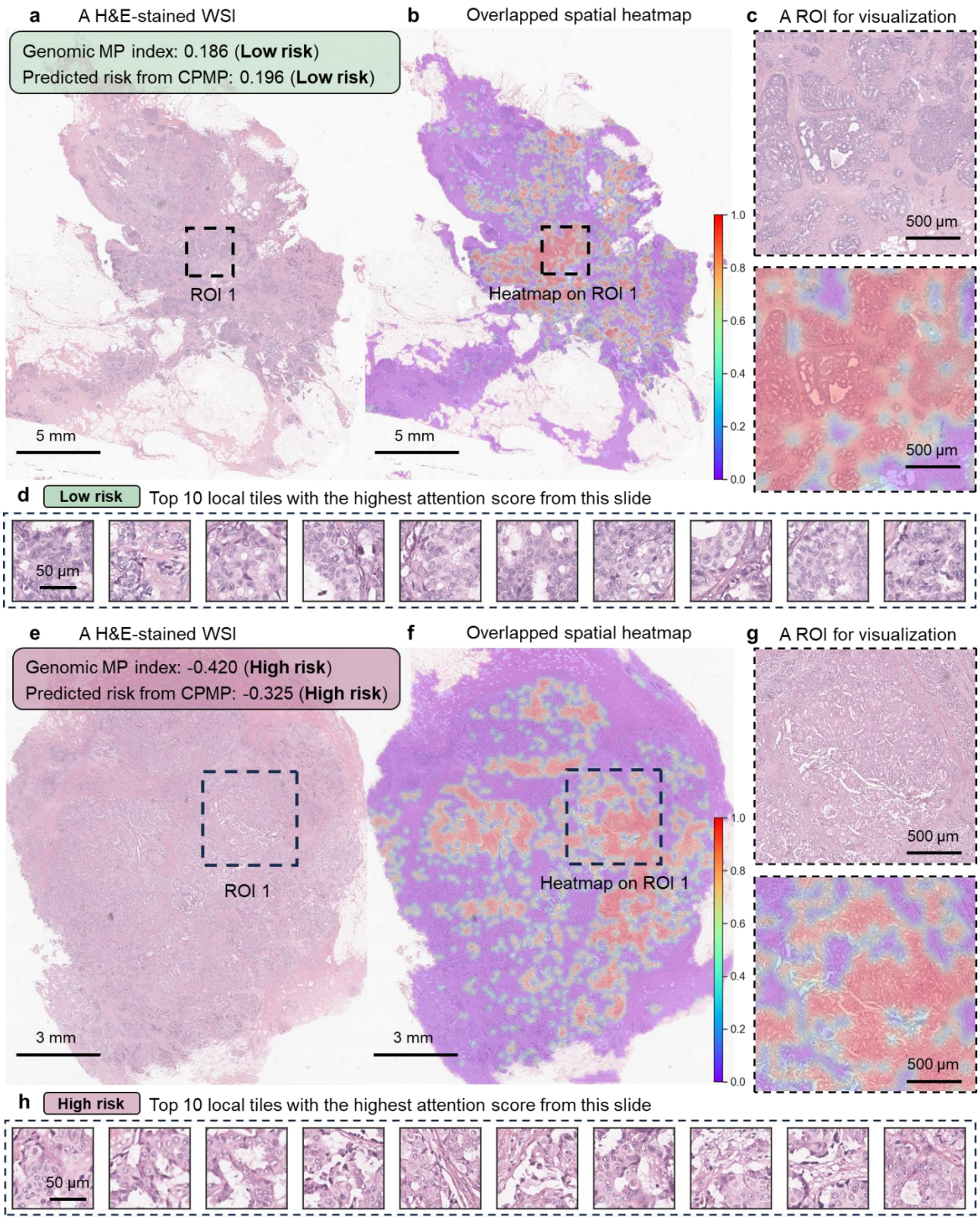

Fig. 4. Heatmap visualization from CPMP revealing the spatial localization of tumors at the whole-slide level.

a, e. Example H&E-stained WSIs of patients with genomic MP Low-risk (**a**) and High-risk (**e**) in the TJMUCH-MP cohort. Genomic MP diagnostic risk index and corresponding risk prediction results from CPMP are marked in the

box. Scale bar: 5 mm / 3 mm. **b, f.** The overlapped spatial heatmap visualization for the corresponding Low-risk (**b**) and High-risk (**f**) cases at the whole-slide level. The tile attention scores that contributed to the final patient-level risk prediction were yielded from the agent-attention matrix of local tiles using the attention rollout method. Spatial heatmap results were then generated by mapping the attention scores of all local tiles to their corresponding locations within the WSIs and they were superposed onto the histopathological WSIs for interpretability. Please see Methods section. A colorbar was added on the lower right of the heatmap. Scale bar: 5 mm / 3 mm. **c**, **g.** Fine-grained pathological ROIs extracted from the WSIs and the corresponding spatial heatmap visualization results. Scale bar: 500 μm. **d, h.** Top 10 local tiles with the highest attention score from each slide for both MP Low-risk (**d**) and High-risk (**h**) cases. Scale bar: 50 μm. WSI=Whole Slide Image, ROI=Region of Interest, MP=MammaPrint diagnostic test.

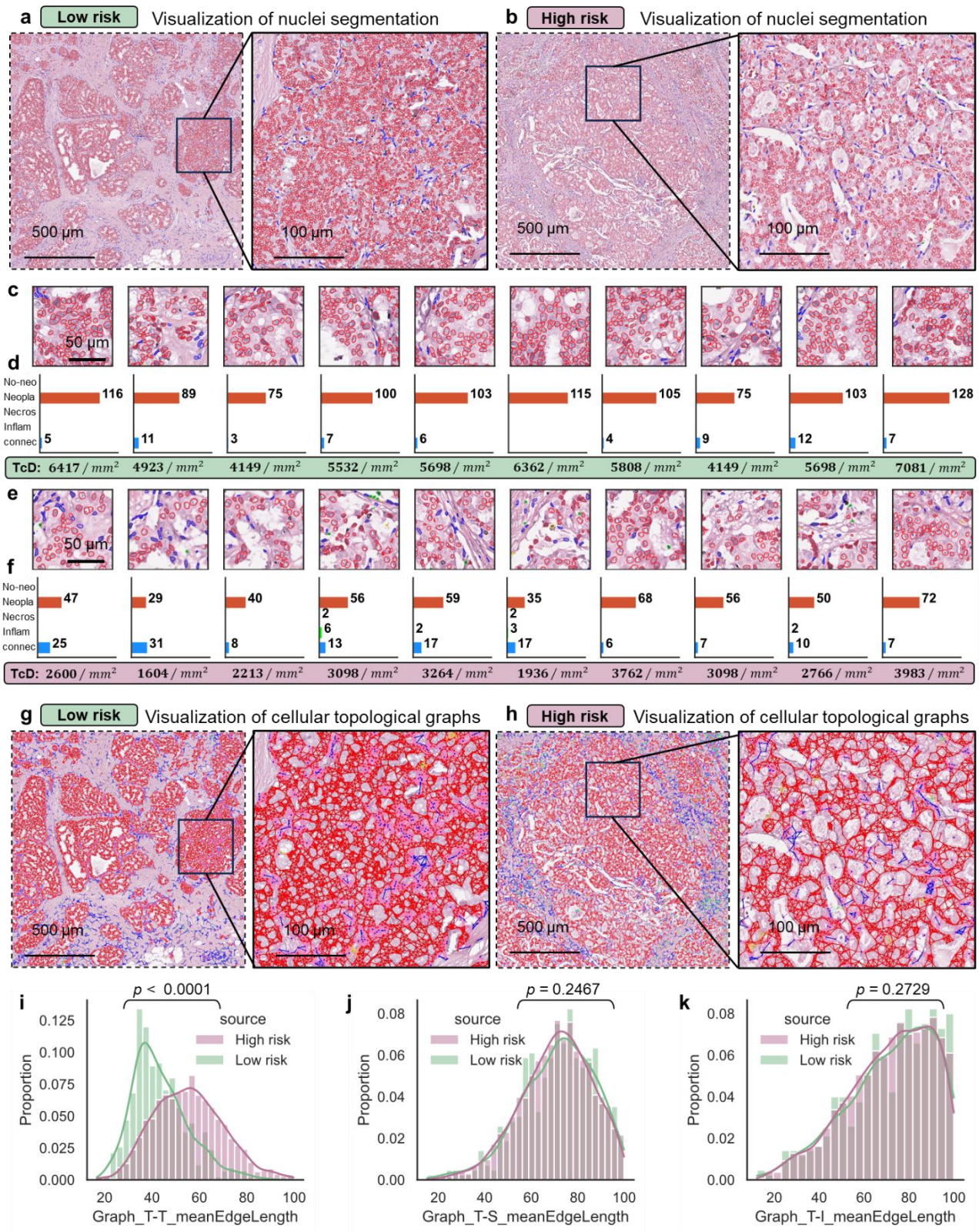

Fig. 5. Intercellular interaction patterns within the tumor ecosystem.

a, b. Visualization of nuclei segmentation results for MP Low-risk (**a**), High-risk (**b**) ROIs and their corresponding fine-grained regions. HoverNet was applied for simultaneous nuclei segmentation and classification. Each cell nucleus was classified as one of 5 cell types (neoplastic, inflammatory, necrotic, connective, and non-neoplastic

epithelial cells). The contours of identified nuclei were located and drawn on the fine-grained pathological ROIs. Scale bar: 500 μm and 100 μm. **c, e.** Nuclei segmentation results of top 10 local tiles with the highest attention scores for MP Low-risk (**c**) and High-risk (**e**). Scale bar: 50 μm. **d**, **f.** The cellular composition quantitation of the five cell types in each local tile region. The number of cells for each type are labelled on the right side of the plot and the TcD values representing the density of tumor cells are marked in the lower part of the plot. **g, h**. Visualization of cellular interaction topological graphs in the tumor ecosystem. Three main cellular components (neoplastic, inflammatory, and connective stromal cells) within the tumor ecosystem were retained to construct topological graphs among these detected cell nuclei. MP Low-risk (**g**), High-risk (**h**). Scale bar: 500 μm and 100 μm. **i-k**. The feature distribution of graph-edge-length-related MeanEdgeLength (the average edge lengths of a tumor cell interacting with tumor cells (**i**, Graph_T-T_MeanEdgeLength), stromal cells (**j**, Graph_T-S_MeanEdgeLength), and inflammatory cells (**k**, Graph_T-I_MeanEdgeLength)) at the single-cell level for both MP risk groups. The sc-MTOP framework was leveraged for profiling the graph-based features of tumor cells interacted with other cells. *p*-value was calculated by the nonparametric Mann-Whitney U rank test. WSI=Whole Slide Image, ROI=Region of Interest, MP=MammaPrint diagnostic test, TcD=Tumor cells Density.

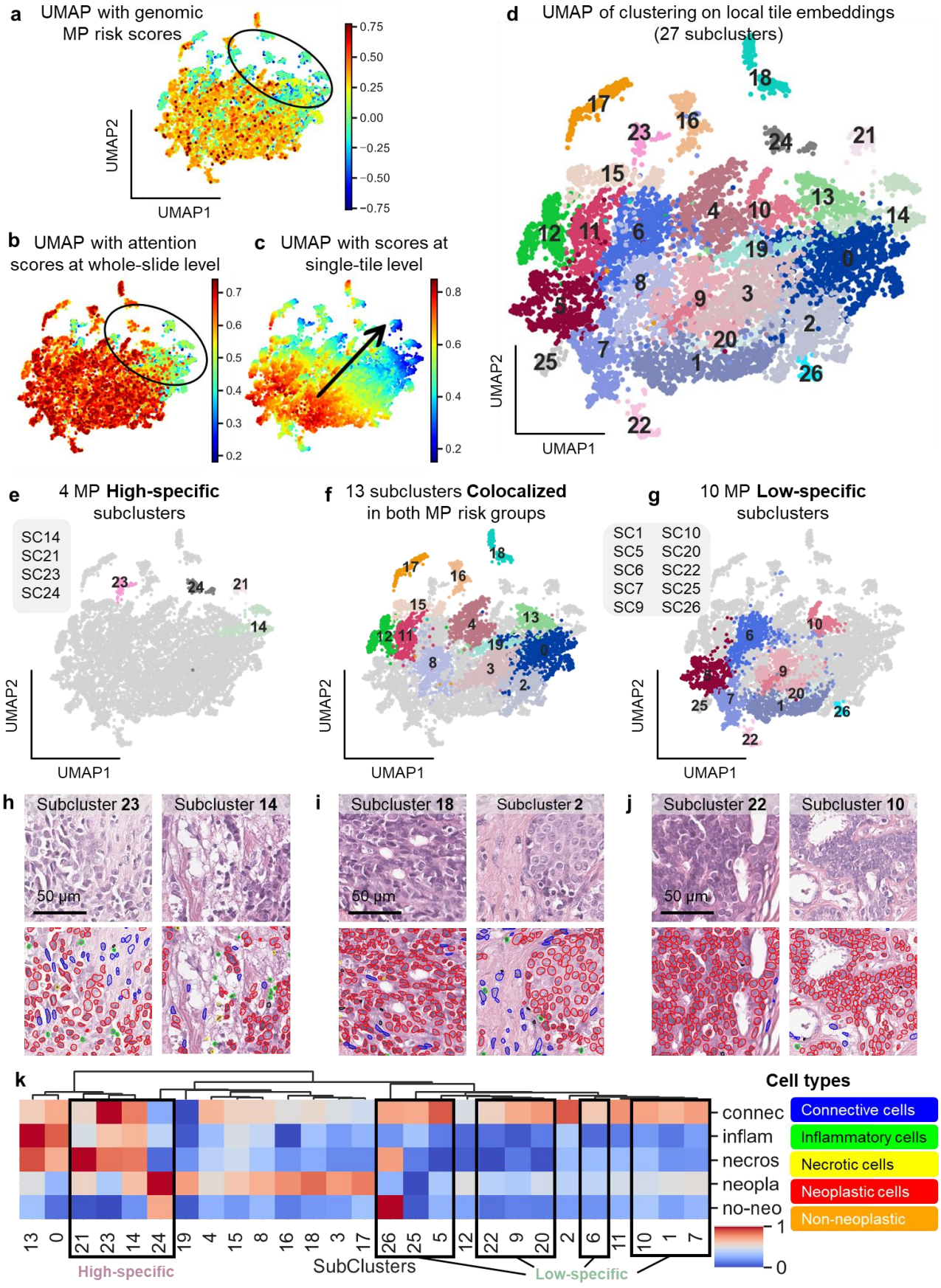

Fig. 6. The characterization of tumor morphological phenotypes discovered by CPMP.

a-c. UMAP plot of learned embeddings of local tiles sampled from WSIs in the test set. The UMAP plots were colored by: (**a**) genomic MP risk scores, (**b**) local attention scores from CPMP at the whole-slide level, (**c**) prediction probability scores from CPMP at the single-tile level. For each WSI, top 100 tiles with the highest attention scores were identified as representative local regions to create a collection of tiles. Each tile in this collection was then fed into CPMP for feature embedding extraction. **d.** UMAP plot of unsupervised clustering results of learned embeddings of local tiles. Clusters are labeled with inferred subcluster ids (subclusters were named as SC0-SC26 and assigned colors for visualization). **e-g**. UMAP subplots of High-specific (**e**), Low-specific (**f**), and colocalized (**g**) subclusters of embeddings of local tiles. The genomic MP Low-risk or High-risk groups were assigned to each subcluster and subclusters associated with a specific risk group were detached for UMAP subplots. **h-j**. Depiction of representative local tiles and the corresponding nuclei segmentation results in High-specific (**h**), Low-specific (**i**), and colocalized (**j**) subclusters demonstrating the phenotypic characteristics of MP risk groups. **k.** Comparison of the cellular composition quantitation of the five cell types (connective, inflammatory, necrotic, neoplastic, and non-neoplastic epithelial cells) among 27 subclusters. Min-max normalization was conducted to scale the quantitative features of the cellular composition and hierarchical clustering was performed on 27 subclusters using the Ward variance minimization algorithm and the euclidean distance metric. Distinct cellular composition of subclusters were observed in the cluster-heatmap, revealing its association with the High-specific, Low-specific, and colocalized subcluster phenotypes discovered by CPMP. UMAP=Uniform Manifold Approximation and Projection, MP=MammaPrint diagnostic test.

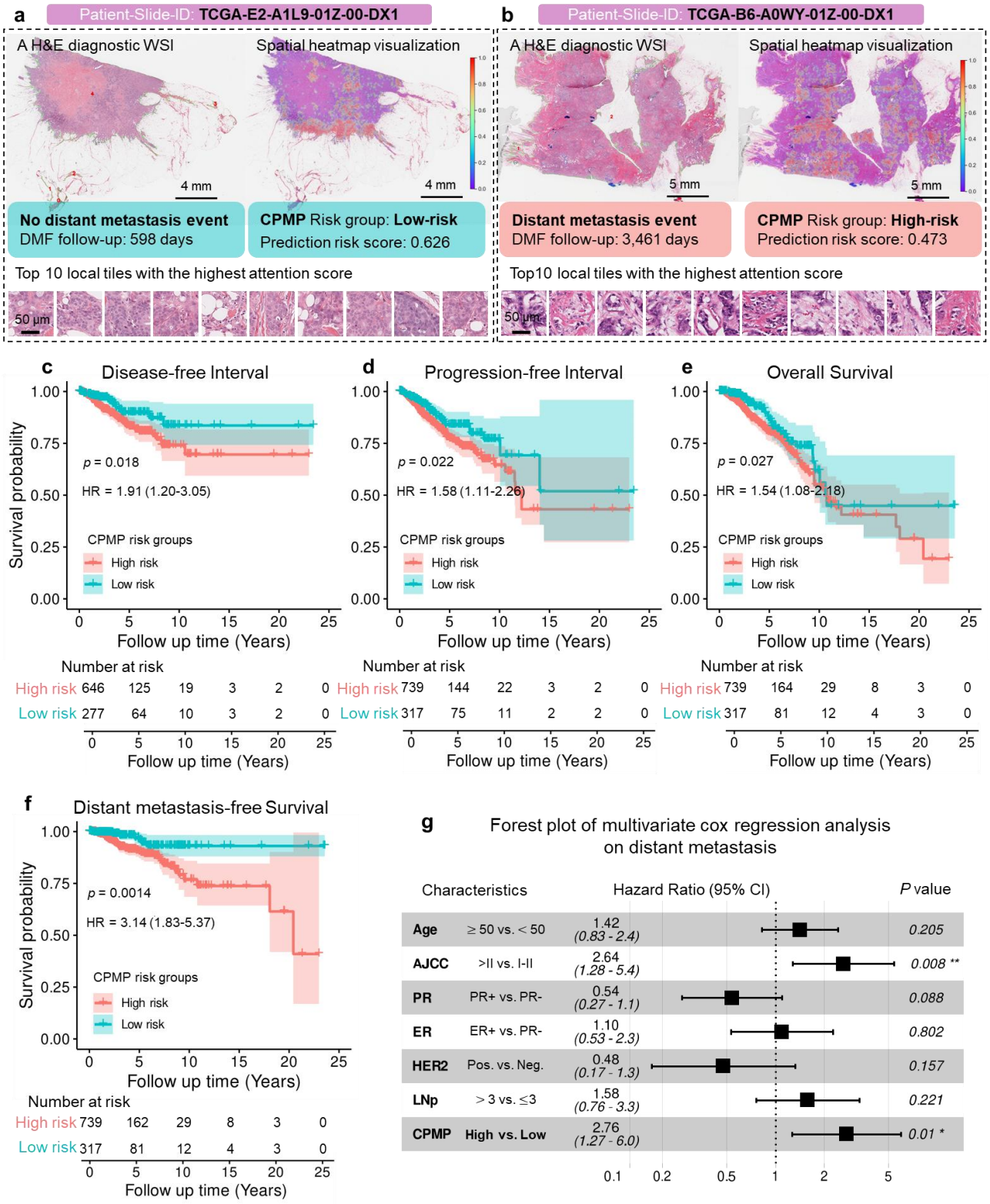

Fig. 7. Prognostic risk stratification performance of CPMP on the external TCGA-BRCA data.

a, b. Whole-slide-level spatial attention heatmap visualization results for tumor morphological patterns associated with prognostic events in the TCGA-BRCA dataset. The H&E-stained diagnostic WSIs of patients with (**b,** Scale bar: 5 mm) or without (**a,** Scale bar: 4 mm) distant metastasis event, the whole-slide-level overlapped spatial

heatmap visualization derived from CPMP, and top 10 local tiles (Scale bar: 50 μm) with the highest attention score revealing the tumor morphological patterns of the slide are all shown. **c-f**. Kaplan–Meier curves for Disease-free Interval (**c**), Progression-free Interval (**d**), Overall Survival (**e**), and Distant metastasis-free Survival (**f**) of patient groups with High-risk and Low-risk in the external TCGA-BRCA dataset. Risk groups were directly predicted by CPMP trained in the TJMUCH-MP cohort. HR and 95% CI were calculated by the Cox proportional hazard model while *p*-value was calculated using the two-sided log-rank test. **g.** Forest plot of multivariate cox regression analysis including the covariates age, AJCC pathologic tumor stage, PR status, ER status, HER2 status, LNp, and CPMP risk group and their association with distant metastasis-free survival. HR and 95% CI were calculated by the Cox proportional hazard model and *p*-value was calculated using the two-sided log-rank test. (* indicate statistical significance, * $p<0.05$, ** $p<0.01$). WSI=Whole Slide Image, HR=Hazard Ratio, DMF= Distant Metastasis-Free, CI=Confidence Intervals, AJCC=The American Joint Committee on Cancer, PR=Progesterone Receptor, ER=Estrogen Receptor, HER2=Human Epidermal growth factor Receptor 2, LNp=The number of Lymph Nodes positive, HR+ =Hormone Receptor positive, LN0=Lymph Nodes negative, LN1=The number of Lymph Nodes positive≤3.