nature communications

Article

https://doi.org/10.1038/s41467-025-57587-y

A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images

Received: 1 June 2024

Accepted: 21 February 2025

Published online: 10 March 2025

Check for updates

Zhaochang Yang[1,8], Ting Wei[1,8], Ying Liang[1,2], Xin Yuan[1,3,4,5], RuiTian Gao[1], Yujia Xia[1], Jie Zhou[6], Yue Zhang[1,3] & Zhangsheng Yu[1,3,4,6,7]

Computational pathology, utilizing whole slide images (WSIs) for pathological diagnosis, has advanced the development of intelligent healthcare. However, the scarcity of annotated data and histological differences hinder the general application of existing methods. Extensive histopathological data and the robustness of self-supervised models in small-scale data demonstrate promising prospects for developing foundation pathology models. Here we show BEPH (BEiT-based model Pre-training on Histopathological image), a foundation model that leverages self-supervised learning to learn meaningful representations from 11 million unlabeled histopathological images. These representations are then efficiently adapted to various tasks, including patch-level cancer diagnosis, WSI-level cancer classification, and survival prediction for multiple cancer subtypes. By leveraging the masked image modeling (MIM) pre-training approach, BEPH offers an efficient solution to enhance model performance, reduce the reliance on expert annotations, and facilitate the broader application of artificial intelligence in clinical settings. The pre-trained model is available at https://github.com/Zhcyoung/BEPH.

Histopathological image analysis has been considered as the gold standard for cancer diagnosis[1,2]. However, the conventional manual approach to diagnosis is time-consuming and burdensome for pathologists. They need to carefully examine morphological features, such as tissue shape and cell sizes, across the entire slide to provide an accurate diagnosis. Insufficient diagnostic experience can lead to missed diagnoses and misdiagnoses, significantly impacting subsequent treatment[3].

Advancements in computational pathology and artificial intelligence offer possibilities for objective diagnosis, prognosis, and treatment response prediction using gigapixel slides[4]. Particularly, deep learning-based computational pathology has shown promising prospects in various pathological tasks, including disease grading, cancer classification, and survival prediction[1,5–7]. Despite the revolutionary impact of deep learning on pathological images, many challenges still remain to be addressed.

The typical approach for deep learning analysis of pathological images usually involves loading ImageNet pre-trained weights and training the model using WSIs and specific labels. However, the inherent differences between natural images and pathological images affect the performance[8–10]. Researchers have recognized the positive impact of pre-training on pathological images, but their methods often involve supervised pre-training tailored to the specific tasks. This approach is limited by the scarcity of training data and the challenge of

[1]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. [2]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China. [3]SJTU-Yale Joint Center for Biostatistics and Data Science Organization, Shanghai Jiao Tong University, Shanghai, China. [4]Center for Biomedical Data Science, Translational Science Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [5]National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai, China. [6]School of Mathematical sciences, Shanghai Jiao Tong University, Shanghai, China. [7]Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [8]These authors contributed equally: Zhaochang Yang, Ting Wei. e-mail: yue.zhang@sjtu.edu.cn; yuzhangsheng@sjtu.edu.cn

transferring pre-trained weights to other tasks. Additionally, significant variations in morphology among different cancer types hinder the adaptation of pre-trained models to cross-cancer tasks. For more tasks, researchers need to continuously annotate data and improve upon existing methods, reducing the efficiency of practical applications[11,12]. A promising solution to address these challenges is to pre-train foundation models using a wide variety and a massive amount of pathological images[13–15].

Foundation models in computational pathology are developed by pre-training on a large volume of unsupervised digital pathology images. These models learn underlying structures and relationships in images, establishing a robust knowledge base. Subsequently, the foundation models are well fine-tuned using small labeled datasets customized for specific tasks and always achieve high performance[16,17]. There are two main training strategies: supervised learning (SL) and self-supervised learning (SSL). SL involves pre-training on labeled data, which is challenging due to the large gap between the speed of data annotation and data acquisition[18]. In contrast, SSL utilizes pretext tasks to extract information from large-scale unsupervised data, reducing reliance on labels. SSL methods include contrastive learning and masked image modeling (MIM). Contrastive learning focuses on constructing positive and negative sample pairs[19,20], while negative samples are typically obtained from other images in the dataset. However, constructing sample pairs in histopathological datasets can be challenging due to the strong resemblance between images. Chen et al. proposed a hierarchical DINO-based pre-training framework, named HIPT, which achieved excellent performance on subtype classification and survival prediction[21]. Due to DINO's bottom-up aggregation to learn image representation, this approach requires computationally costly hierarchical pre-training. Moreover, as they mentioned in the original paper, it is not simple to obtain general and discriminative representations of WSI directly.

MIM is designed to reconstruct the obscured image features and has better performance in downstream task fine-tuning compared to contrastive learning[22]. Recent studies have demonstrated the effectiveness of MIM in addressing histopathological problems, such as using Masked Autoencoders (MAE) pre-training on cervical cancer for WSI-level classification[23]. However, these methods typically focus on a single type of cancer with thousands of slides, making it challenging to address the diversity and heterogeneity in histopathology, especially when they are evaluated on data from different sources and different cancer types. Moreover, these methods necessitate interpretability and capability to precisely localize the regions used for making predictive decisions[24,25]. Therefore, there is an urgent need for foundation models that exhibit strong generalization performance, eliminate the requirement for manual region of interest (ROI) extraction, ensure data efficiency, and enhance interpretability to make computational pathology more widely applicable in clinical and research environments.

In this work, we propose an advanced SSL-based foundation model called BEiT-based model Pre-training on Histopathological Image (BEPH). We systematically evaluate its performance and versatility in various cancer detection tasks. Figure 1 provides an overview of the construction and application of BEPH. The histopathological images used for pre-training are sourced from The Cancer Genome Atlas (TCGA), encompassing 32 different types of cancer[26] (Fig. 1a, b). We have collected a total of 11.76 thousand pathology images (with 92 exclusions due to indeterminate magnification), resulting in a pre-training patch dataset of 11.77 million 224 × 224 pixel patches. This dataset is 10 times larger than the ImageNet-1K dataset (1.28 million). Detailed statistics for each cancer slide and patch are presented in Fig. 1b.

To efficiently train BEPH, we utilize advanced SSL techniques (specifically BEiTv2[27,28]) for pre-training on both natural images from ImageNet-1k and the collected pathology images (Fig. 1c). This involves initializing the network with pre-trained weights from natural images and then further pre-training on the TCGA to learn a generalized representation of pathology images[15,29].

Subsequently, we fine-tune BEPH on a wide range of challenging classification and prediction tasks (Fig. 1d). First, we consider the classic patch-level classification tasks. Second, we explore the identification of WSI levels, which provide comprehensive and accurate pathological information. This work focuses on the subtypes classification of BRCA, NSCLC, and RCC (see Supplementary Table 1 for full name of cancer). Third, we consider survival prediction tasks in BRCA, CRC, CCRCC, PRCC, LUAD, and STAD. The data used for WSI classification and survival prediction are sourced from TCGA. Hence, our evaluation of the universality of cancer analysis is based on multi-task and multi-cancer. Compared to the state-of-the-art models, including traditional transfer learning models, weakly supervised models, and foundation models including CHIEF, UNI, and GigaPath, BEPH consistently achieves superior performance and label efficiency in these tasks. Furthermore, we delve into the interpretation of BEPH's cancer detection performance through qualitative results and ablation experiments, which indicate that visualization regions of the image correspond to established knowledge in pathology. Finally, we make BEPH publicly available to encourage diverse computational pathology research, serving as a foundation for various downstream tasks.

Results

Patch-level classification

We used the BreakHis[30] dataset for binary classification (benign and malignant tumors) through five random experiments[31,32]. Unlike the image-patching strategy used in current top-performing algorithms, we downscaled the images by a factor of 3.125 to 224 × 224 pixels, resulting in a loss of a significant amount of image details (Supplementary Table 11). Even though, at the patient level and image level, BEPH achieved an average accuracy (ACC) of 94.05 ± 1.3875 and 93.65 ± 0.6730, which is 5–10% higher than the latest reported CNN models (Deep[33], SW[31], GLPB[34], RPDB[35]) and weakly supervised models (MIL-NP[32], MILCNN[32]). It also outperformed the best-reported performance of self-supervised models (MPCS-RP[36]) by 1.9% and 1.5% (Fig. 2a, b and Supplementary Tables 14 and 15). Besides, BEPH exhibits consistently higher performance at certainly magnification levels at the image level and patient level compared to other algorithms (Fig. 2c, d and Supplementary Tables 14 and 15).

To test the generalization ability of the model across different cancers, we further fine-tuned the model on the LC25000 datasets[37]. BEPH achieved an ACC of 99.99% ± 0.03 across the three lung cancer subtypes(Fig. 2e, f), which is higher than any other reported models like shallow-CNN[38], AlexNet[39], ResNet[39], VGG19[39]and EfficientNet-B0[40], as well as the self-supervised model DARC-ConvNet[41].

These results demonstrate that MIM-based pre-training allows the BEPH model to effectively generalize across multiple cancer types, achieving strong performance on diverse pathological image data. These results also suggests that BEPH is robust to variations in image magnifications (Fig. 2c, d and Supplementary Fig. 4) and cancer types, making it potentially applicable to a wide range of cancer studies.

WSI-level classification

With the advancement of technology and improvement in computational power, computational pathology research is increasingly focusing on the pathological information represented by the entirety of WSI. This detailed analysis is crucial for cancer subtyping, which provides a foundation for personalized treatment, aids in predicting patient prognosis, and contributes to understanding the disease and its underlying mechanisms. Therefore, in this section, we train a weakly supervised subtype classification model based on a self-supervised feature extractor with multiple instance learning (MIL). After randomly selecting a 10% independent test dataset, we evaluate the WSI-level

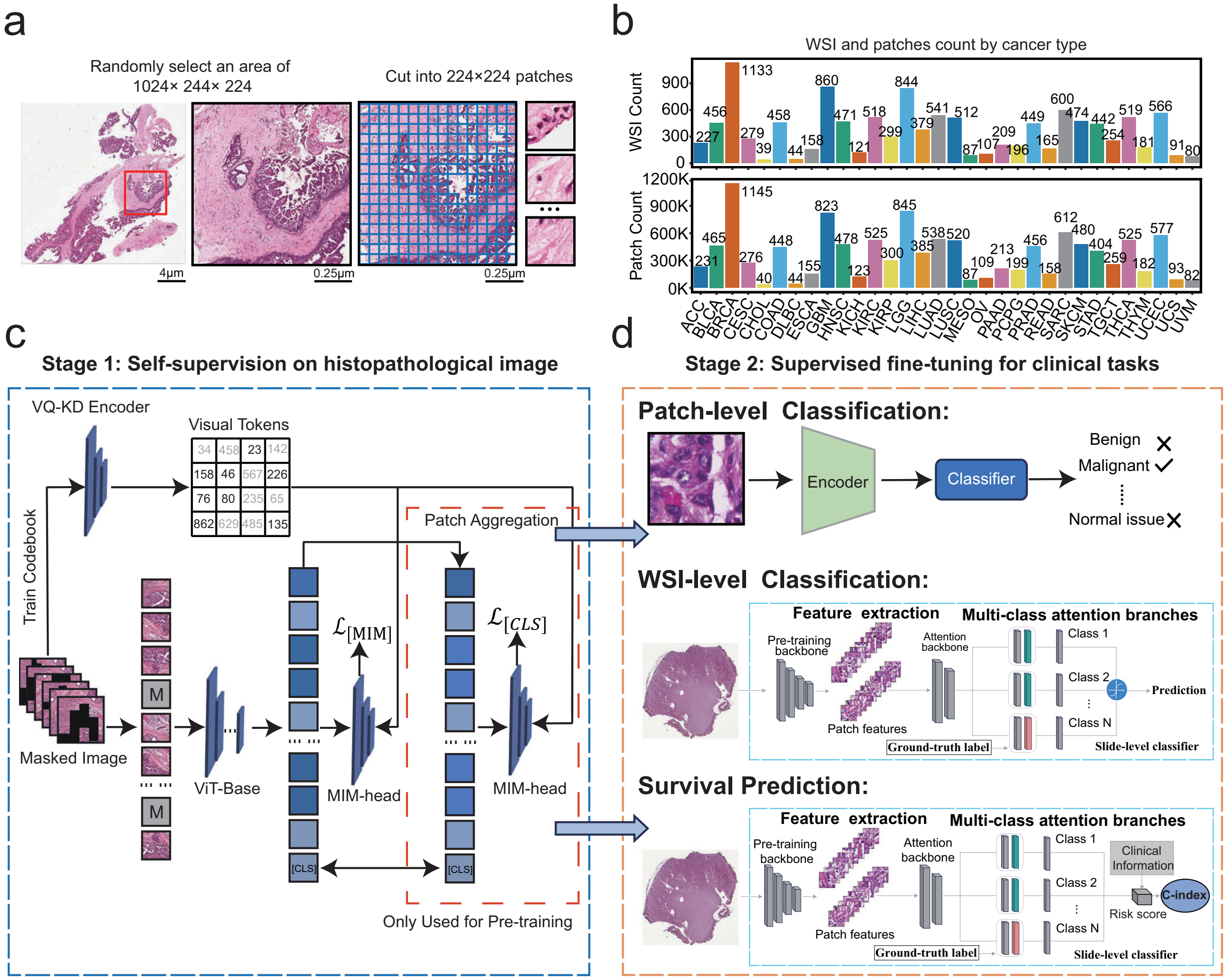
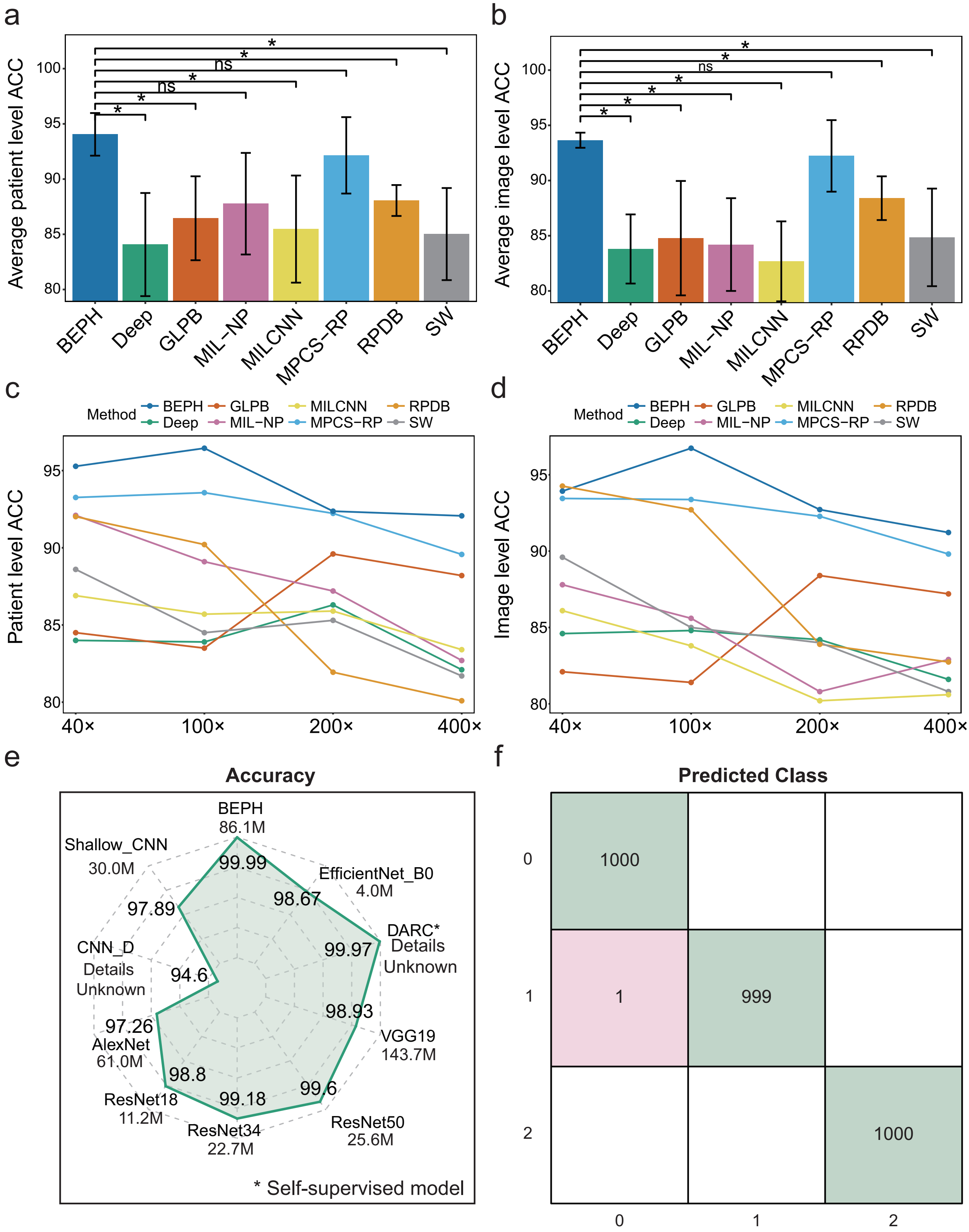


Fig. 1 | Overview of the BEPH architecture. a The segmentation of local tissue regions from the WSI and the division into patches. **b** Distributions of WSIs and patches in the collected pre-training dataset. Source data are provided with this paper. **c** The BEPH architecture for pre-training. The pre-training phase employs the MIM method for SSL of pathological image features. **d** The downstream supervised fine-tuning architectures: patch-level classification structure, WSI-level weakly supervised classification structure, and WSI-level survival prediction structure.

classification performance of three clinical diagnostic tasks through 10-fold cross-validation: renal cell carcinoma (RCC) subtypes, non-small cell lung cancer (NSCLC) subtypes, and nonspecific invasive breast cancer (BRCA) subtypes.

For the classification of the three subtypes of RCC: papillary renal cell carcinoma (PRCC), chromophobe renal cell carcinoma (CRCC), and clear cell renal cell carcinoma (CCRCC) using the public RCC WSI dataset, the model achieves a 10-fold macro-average test area under the receiver operating characteristic curve (AUC) of 0.994 ± 0.0013 (Fig. 3a, b). On the public BRCA WSI dataset, for the two subtypes: invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC), the model achieves an average test AUC of 0.946 ± 0.019 (Fig. 3d, e). For the two subtypes on the public NSCLC WSI dataset: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), the model achieves an average test AUC of 0.970 ± 0.0059 (Fig. 3g, h). These results significantly outperform the corresponding models trained with pre-training on natural images or pathology images, such as MIL[4,21] (DINO pre-trained on 256 × 256 pathology images), GCN-MIL[8,21] (DINO pre-trained on 256 × 256 pathology images), DS-MIL[21,23] (DINO pre-trained on 256 × 256 pathology images), DeepAttnMISL[21,42] (DINO pre-trained on 256 × 256 pathology images), CLAM-SB(ResNet)[4] (ResNet-50 supervised pre-trained on ImageNet-1K), CLAM-SB(DINO)[4,21] (DINO pre-trained on 256 × 256 pathology images), HIPT[21] (DINO pre-trained on 256 × 256 and 4096 × 4096 pathology images), and the latest foundation models: CLAM-SB(CHIEF)(CLAM-SB based on the pre-trained backbone of CHIEF[43]), CLAM-SB(CtransPath)(CLAM-SB based on the pre-trained backbone of CtransPath[9]). However, it performs worse than CLAM-SB (GigaPath) (CLAM-SB based on the pre-trained backbone of GigaPath) and CLAM-SB(UNI) (CLAM-SB based on the pre-trained backbone of UNI). All comparison results can be found in Fig. 3 and Supplementary Table 12.

Specially, we selected the CLAM-SB model to evaluate the impact of different pre-training methods. Overall, UNI achieves the best performance, followed by Gigapath. However, on the lung and kidney cancers, BEPH lags behind them by less than 0.7%, demonstrating competitive results. Their superior performance is reasonable, as UNI and GigaPath were trained on significantly larger datasets compared to ours. Additionally, GigaPath leverages the ViT-G backbone, which is a deeper network with more parameters. For other models, BEPH is better. On the BRCA, BEPH outperforms ResNet-50, which was trained on ImageNet-1K using SL (named ResNet-ImageNet) by 11.92%, DINO pre-trained on 256 × 256 histopathological images (named as DINO-HistoPretrain) by 8.8%. On the lung cancer, BEPH exceeds ResNet-ImageNet by 6.3%, DINO-HistoPretrain by 4.4%, CHIEF by 0.6%, and CTransPath by 1.0%. These results indicate that BEPH is a SSL model with strong stability and generalization capabilities. It effectively learns



valuable features and patterns from large-scale slides, enabling accurate classification and recognition of pathological changes in cancer.

We also tested the zero-shot capability through cross-evaluation, to ensure its reliability and usability on the different datasets and real-world scenarios. When fine-tuning the model on the CAMELYON16 dataset[44], the BEPH achieved a score of 0.9056 ± 0.0221 on the internal validation set, which is comparable to the performance of the state-of-the-art models (Supplementary Fig. 1). Furthermore, the AUC tested on the BACH dataset[45] yielded a score of 0.638 ± 0.0585, which was statistically significantly higher by 20% than other models ($P < 0.05$). Among all the self-supervised and weakly supervised models, the zero-shot capability of BEPH is the most powerful. During cross-evaluation, the AUC results were noticeably lower than those obtained on the internal validation data, indicating a significant improvement in

Fig. 2 | Performance evaluation on publicly available patches datasets. **a** Average accuracy comparison at the patient level on the BreakHis. The values are calculated based on the mean and standard deviation of classification accuracy across all magnifications. **b** As shown in (**a**), average accuracy comparison at the image level across all magnifications. Training and testing sets for each method were independently split, with significance indicated by "*" for p-values < 0.05 and non-significance by "ns", calculated using a two-sided two-sample independent t-test. **c**, **d** Performance of different methods at the patient level and image level, respectively, at magnification of 40 × , 100 × , 200 × , and 400 × . The bar chart results are from five repeated experiments using the same training and testing data ($n = 5$). **e** Performance of different algorithms on the LC25000 dataset. The numbers below each model name indicate the parameter count. "Unknown" denotes models for which the parameter count is unavailable. **f** Confusion matrix on validation set of LC25000 dataset. All data are presented as mean values ± SD. Source data are provided with this paper.

downstream task performance through fine-tuning. In addition, we plan to deploy BEPH to http://yulab-sjtu.natapp1.cc/BEPH, and currently, online fine-tuning of WSI classification tasks is available.

Label efficiency for cancer detection

The foundation models require good label efficiency, because acquiring labeled WSI data is challenging, especially for rare diseases. When the training data was reduced to 25% of its original size while keeping the test dataset unchanged, the results demonstrated that even with a significant reduction, the performance of the BEPH model remained superior to other weakly supervised models and HIPT. Although there was a slight gap compared to the latest foundation models, the difference was minimal (<3%). Additionally, we created different subsets of training data for BEPH, including 25%, 50%, 75%, and 100% of the original data (Fig. 3j). It was observed that with ~50% of the training data, BEPH's performance was comparable to that of weakly supervised models trained on 100% of the data. At this point, BEPH's performance reached levels comparable to those of self-supervised models pre-trained on significantly larger datasets, such as GigaPath (17 times larger), UNI (10 times larger), and CHIEF (6 times larger). This indicates that BEPH's MIM-based pre-training enables the model to effectively tolerate data scarcity, alleviating the challenges associated with data collection for both pre-training and downstream tasks.

Survival prediction

Clinical survival prediction models' development contributes to improving clinicians' ability to evaluate patients' prognosis. Therefore, we trained a survival risk regression model based on a self-supervised feature extractor and weakly SL, named CLAMSurvival. Compared with the weakly supervised model, BEPH ranked first in predictive capabilities among all models in each of the six different cancer types (Fig. 4a): BRCA, CRC, CCRCC, PRCC, LUAD, and STAD. The C-index values were 0.6643 (95% CI 0.6234, 0.7052), 0.6758 (95% CI 0.5863, 0.7652), 0.6685 (95% CI 0.6206, 0.7164), 0.7135 (95% CI 0.5203, 0.9068), 0.6039 (95% CI 0.5093, 0.6986), and 0.5941 (95% CI 0.5348, 0.6534), respectively. Overall, the performance improvement in C-index ranged from 1.1% to 5.5%. We conducted statistical tests using paired t-tests and Cohen's d (Even when the paired t-test results were not significant, a Cohen's d value greater than 0.8 indicates meaningful differences between methods). The test results are provided in Supplementary Table 5. Although the results of BEPH overlap with those of other methods, there are still observable differences.

Similarly, we selected the CLAMSurvival model to evaluate the impact of different pre-training methods. Compared to the latest foundation models, BEPH demonstrated excellent and consistent performance, achieving either the best or second-best rank in almost all cases (Supplementary Fig. 9 and Supplementary Table 13). Statistical analysis showed that BEPH significantly outperformed CtransPath ($p < 0.05$, Cohen's d ≥ 0.7) and GigaPath (Cohen's d > 0.5, except for CCRCC) in C-index across almost all cancer types (Supplementary Table 5). Although UNI and CHIEF outperformed BEPH in specific cancer types (UNI in CCRCC; CHIEF in BRCA and STAD), BEPH maintained the second-best rank and the highest average rank, confirming its robustness and stable performance. On average, BEPH outperforms these ResNet-ImageNet, DINO-HistoPretrain, UNI, CHIEF, CTransPath, and GigaPath, models by 7.76%, 3.25%, 2.58%, 8.74%, 3.56%, and 6.04%.

Furthermore, using ResNet-ImageNet and DINO-HistoPretrain as examples, we analyzed the ability of these methods for prognostic risk groups[46]. Figure 4b shows that BEPH is superior in distinguishing between high-risk and low-risk patients, on the cancers mentioned above. For example, on CCRCC, BEPH recognizes high-risk and low-risk patients, with a p-value smaller than 0.0001 (the p-value of 0.00071 given by DINO-HistoPretrain and the p-value of 0.0002 given by ResNet-ImageNet). On CRC, our method also brings a significant difference with the p-value of 0.016 (p-value < 0.05), while DINO-HistoPretrain and ResNet-ImageNet fail to distinguish high and low survival risks significantly (p-value = 0.21 and p-value = 0.51) (for other cancers, see Supplementary Fig. 2).

These indicate that BEPH is adept at identifying survival-related patterns and better predicts patients' survival. Furthermore, it doesn't require annotations on histopathological images or additional information such as genomes, making it more practical for real-world applications.

Whole-slide attention visualization and interpretability

To gain a deeper understanding of the exceptional performance of BEPH in downstream tasks, we conducted a qualitative analysis of BEPH's application in self-supervised pre-training and specific task decision-making. We explored the patch-level feature space learned by the BEPH. We introduced an external dataset, NCT-CRC-HE-100K[47], and used UMAP to reduce its 768-dimensional feature representation to two dimensions (only using the backbone for feature extraction without downstream training). It can be observed that patches belonging to different categories were clustered into distinct clusters in the feature space (Supplementary Table 6 and Fig. 5a). The performance of feature extraction further pre-trained on histopathology images is superior to that solely pre-trained on ImageNet-1k, and significantly better than randomly initialized parameters. This is particularly evident in the cancer-associated stroma (STR) and colorectal adenocarcinoma epithelium (TUM) categories.

Additionally, we visualized the WSI-level feature space using principal component analysis (PCA) reducing the feature vectors to two dimensions. The results show that as the training data increases, the clusters of different categories in the feature space become tighter, and show clear boundaries between different categories, indicating that an increase in training data helps the model learn features better (Fig. 3c, f, i and Supplementary Fig. 5). To interpret the fine-tuning performance, we employed attention scores to visualize the salient regions of the images in the downstream tasks. We found that even without explicitly informing the model about the presence of tumor tissue and its subtypes, just using weak SL with WSI-level labels, the attention heatmaps exhibited high consistency with the annotations of pathologists across all cancer (Fig. 5b and Supplementary Figs. 7 and 8). This indicates that the downstream fine-tuned model can accurately identify tumor and tissue regions within a limited number of iterations (20 epochs). On the other hand, we calculated the attention scores for cancerous and non-cancerous regions and then quantified the model's focus on cancerous regions by counting the number of cases where the attention scores for cancerous regions were higher than those for non-cancerous regions. Additionally, we summed the

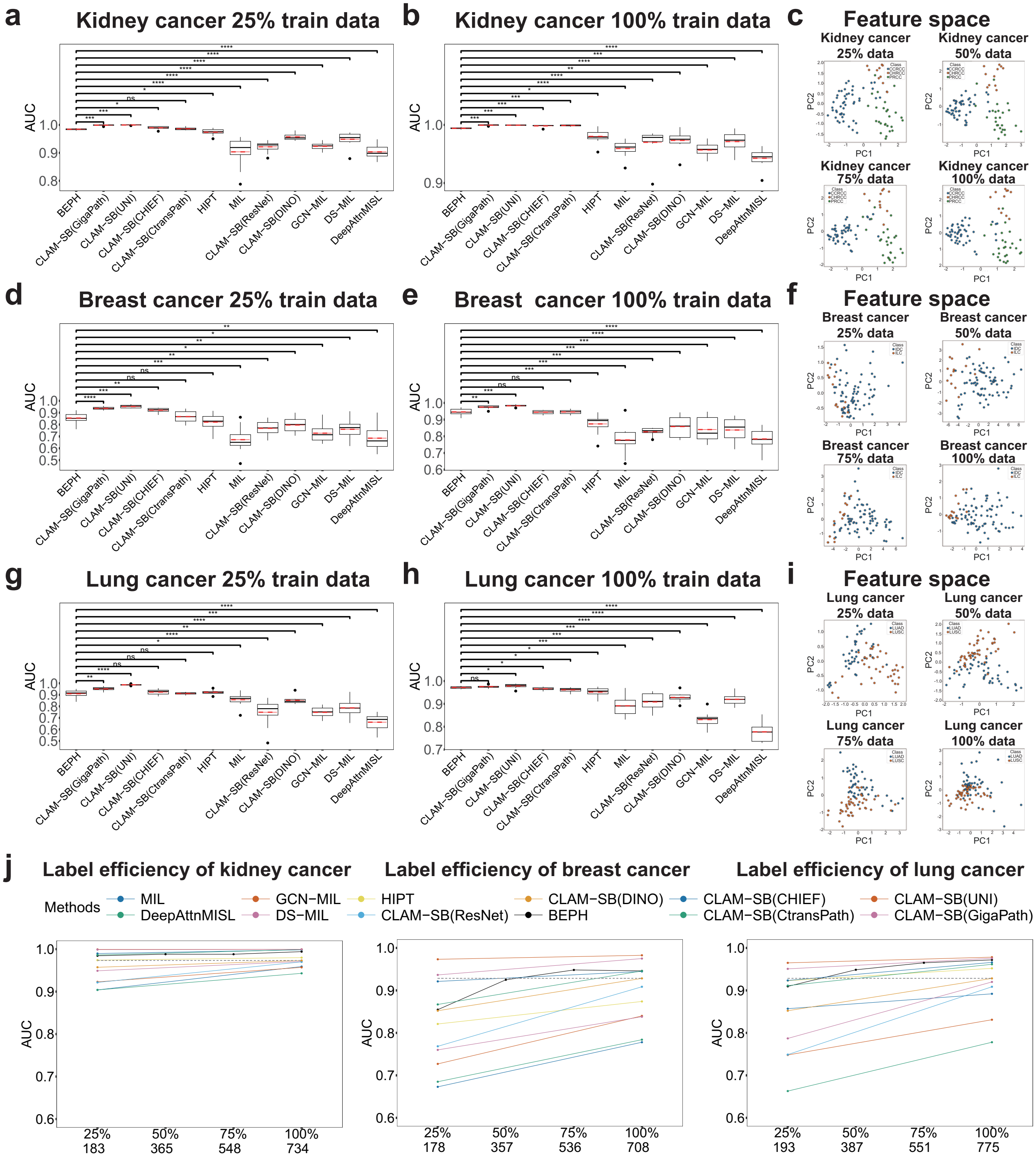
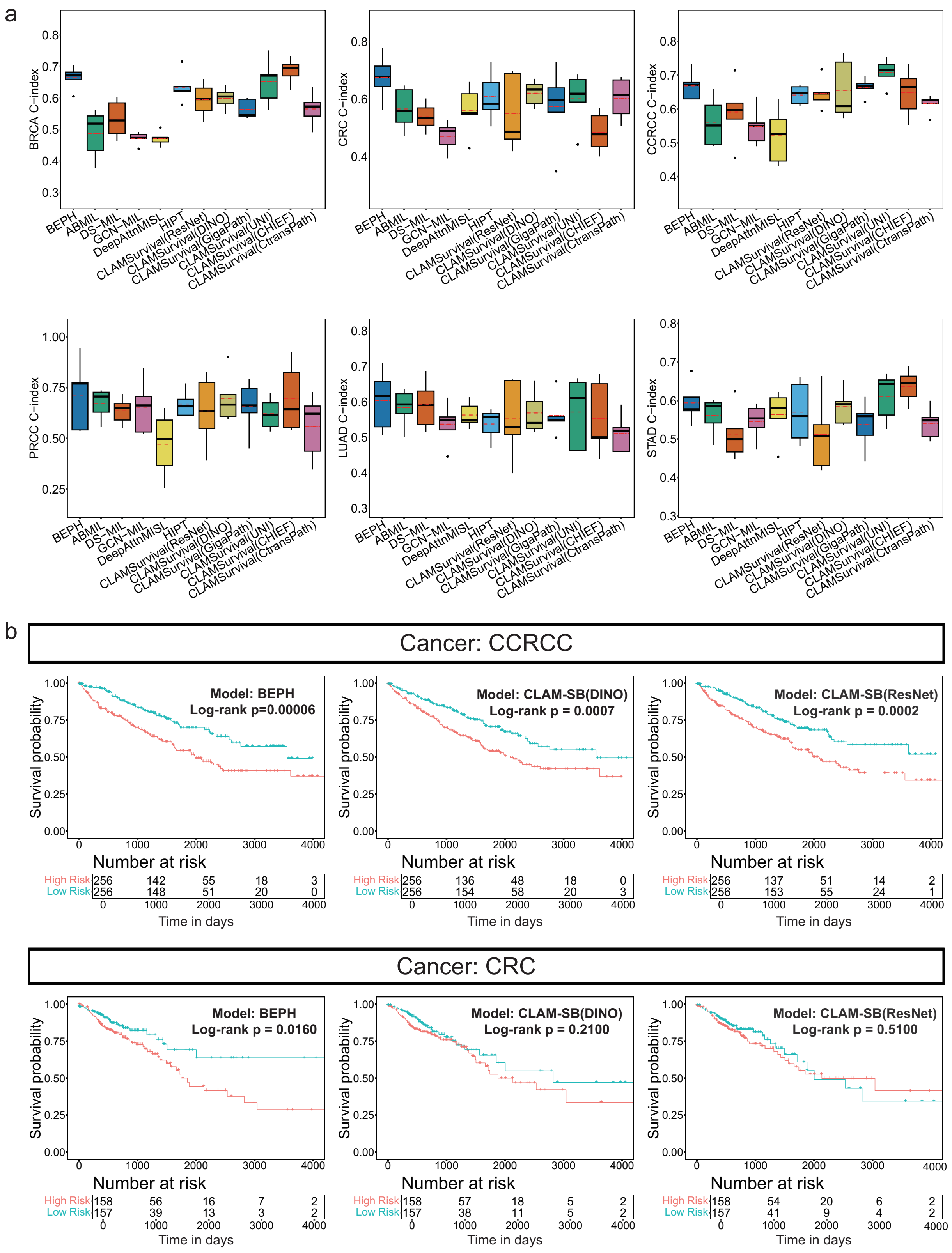


Fig. 3 | Weakly supervised classification results on the WSI-level independent test set. a, **d**, **g** Independent test results for kidney cancer, breast cancer, and lung cancer, using only 25% of the training data. **b**, **e**, **h** Independent test results for kidney cancer, breast cancer, and lung cancer using 100% of the training data. **c**, **f**, **i** Two-dimensional visualization of WSI-level feature space in BEPH, with different training data sizes. With the increase of the train dataset, the clusters are tighter, indicating that the true relationships between features are easier to capture and represent. The results in (**a**, **b**, **d**, **e**, **g**, and **h**) are obtained by selecting the best weights based on 10-fold Monte Carlo cross-validation on the validation set ($n = 10$), and then evaluated on the independent test set, with significance indicated by "*" for p-values < 0.05, "**" for p-values < 0.01, "***" for p-values < 0.001, "****" for p-values < 0.0001 and non-significance by "ns", calculated using the two-sided Wilcoxon test. The black solid line indicates the median and the red dashed line represents the mean. The whiskers extend from the box to the smallest and largest values within 1.5 times the IQR, while points outside the whiskers are considered outliers. **j** Tolerance of different pre-training models to data reduction. The *x*-axis adopts a dual-scale design, combining the percentage of samples (primary) and the training data size (secondary). These results demonstrate that the MIM pre-training model can achieve good performance after training on a limited number of labeled slides, surpassing other weakly supervised baselines. All data are presented as mean values ± SD. Source data are provided with this paper.



differences between attention scores of cancerous and non-cancerous regions (Supplementary Fig. 6). As shown in Supplementary Table 7, BEPH almost achieved the highest count and total score compared to other pre-training methods. Besides, it demonstrated statistically significant differences in attention scores between cancerous and non-cancerous regions, with a p-value less than 0.01. ResNet-ImageNet and DINO-HistoPretrain exhibit significantly reduced attention towards tumor regions, which could explain their inferior performance compared to BEPH.

Ablation test

To investigate the effectiveness of pre-training using pathology images, we conducted ablation experiments. Figure 5a and Supplementary Table 6 demonstrates that the model pre-trained with pathological

Fig. 4 | Survival Prediction. a The C-index of different methods on six cancers. BEPH exhibits a higher C-index compared to weakly supervised models and baseline models across six different cancer subtypes ($n = 5$). The black solid line indicates the median and the red dashed line represents the mean. The whiskers extend from the box to the smallest and largest values within 1.5 times the IQR, while points outside the whiskers are considered outliers. All results are from 5-fold cross validation. Two-side paired t-test (p-value < 0.5 is considered significant) and Cohen's d (Small effect size when 0.2 < Cohen's d < 0.5, Medium effect size when 0.5 ≤ Cohen's d < 0.8, Large effect size when Cohen's d ≥ 0.8) to evaluate model differences across six cancer types. All data are presented as mean values ± SD. **b** Survival curves for patient stratification are depicted using Kaplan–Meier analysis. These curves illustrate the true survival rates of patients classified as high or low risk. The number of patients who survived in each risk group is shown below the survival curves. The log-rank test calculates the p-value, indicating the significance of the difference in survival between the two groups. Source data are provided with this paper.

images exhibits superior feature extraction ability than that pre-trained solely with ImageNet-1k. To quantitatively evaluate the effect of the improvement, we performed a patch-level evaluation on BreakHis, WSI-level evaluation on NSCLC subtype, and survival prediction on BRCA, respectively. Table 1 shows that the performance of the model pre-training on pathological images is better than that of the model pre-trained only on ImageNet-1K (BEiTv2-ImageNet). This highlights the difficulty of SSL using natural images for specialized domains. Additionally, we demonstrated the importance of MIM in pre-training by reducing the mask ratio. The results show an excessively low mask ratio can decrease the difficulty of the pre-training task, thereby limiting the model's learning capacity. In some cases, this can yield performance even worse than directly using BEiTv2-ImageNet. Complete ablation test results are in Supplementary Tables (Supplementary Tables 8, 9 and 10).

Discussion

In summary, we employed the MIM approach to develop BEPH, a foundation pathology model trained on unlabeled pathology images. The contributions of this work are as follows. First, we developed a self-supervised model using over 10 thousand WSIs from 32 cancer types in TCGA. The key technical advantage lies in using MIM to train a histology-specific feature extractor on ~11 million sampled unlabeled patches. Despite the smaller scale of our pre-training data compared to CHIEF, UNI, and GigaPath, the comparable performance highlights the effectiveness of this approach. Second, BEPH adapts effectively to various tasks by learning image representations through the pretext task of reconstructing images from heavily masked patches. It improves patch-level and WSI-level cancer diagnosis, as well as survival risk prediction. The comparative analysis of ImageNet pre-trained models and contrastive self-supervised models, even when trained with limited data, shows that BEPH performs better, highlighting the importance of MIM pre-training. Additionally, after limited downstream task training, BEPH generates localized tumor heatmaps that match pathologists' annotations. Combined with the two-dimensional feature scatter plots, it is shown that MIM pre-training using tissue images enhances the model's attention to specific tissue features.

BEPH has broad applications in cancer detection tasks. When pixel annotations are available to segment cancerous tissues, BEPH can perform patch-level classification tasks. Even without pixel annotations, the model can take the entire histopathological image as input and perform weak SL for WSI classification tasks. In cases where clinical information is available, BEPH can be transformed into a survival prediction model to predict patients' outcomes.

Our work represents an important step towards establishing a general-purpose foundation pathology model. However, there are still some challenges that need to be addressed. First, the robustness of the BEPH model has been validated only in limited scenarios within this study. Further validation on multi-institutional, larger-scale, and more diverse cancer datasets is necessary to fully evaluate the model's applicability in broader clinical settings. Second, the pre-training dataset we have collected is consistent in size with datasets used in earlier studies, offering the necessary quantity and cancer diversity to ensure model performance. However, the dataset size is smaller compared to those used in many currently published models, such as the Prov-GigaPath[48], and it does not accurately reflect the data distribution in real-world scenarios. While BEPH performed better in the survival prediction task, the improved accuracy of UNI and GigaPath in WSI classification strongly suggests that dataset scale and diversity are critical drivers of model performance. Third, because of the finite computing resources and uncertain performance improvement from pre-training large-scale parameter models on BEPH, we opted for the less complex ViT-base backbone instead of ViT-Giant. However, it is worth noting that with more parameters, ViT-Giant may potentially enhance the model's representational capacity, which could be a direction for future improvements like GigaPath. Additionally, as MIM models continue to evolve, the currently used pre-training methods may constrain the potential performance of future foundation models. Finally, even if the training time and parameters are consistent with other foundation models of similar scale and the loss reaches a convergent state, we still don't have sufficient evidence that BEPH has fully learned the features of pathological images. In theory, longer training time generally leads to better performance. It remains uncertain whether further extending the training time would improve the performance.

In the future, our plans involve constructing larger-scale pre-training datasets that include a greater variety of pathological images from different sources, allowing the pre-trained models to be more extensive. Furthermore, the potential path toward achieving general artificial intelligence in pathology lies in multimodal foundation models that incorporate data from multiple domains like omics, text, and images. Exploring the integration of multimodal data holds significant research value in the field.

Methods

Processing WSIs

We established our preprocessing pipeline for the 11.76 thousand pathology patches. Following HIPT, in the first step, for each slide, the pre-training process begins with the automatic segmentation of tissue regions. The slides are loaded at a downsampled magnification (e.g., 40×). Then, a binary segmentation mask of the foreground and the background is generated by thresholding the saturation channel of the image. Specifically, we employed the histogram-based Otsu image segmentation approach. The input image is converted to grayscale, and its histogram is constructed by counting the number of pixels at each gray level (0 to 255). Then, based on a threshold t, the foreground and background regions are segmented. The optimal t is determined by maximizing the inter-class variance.

$$\sigma_B^2(t) = w_1(t) \cdot w_2(t) \cdot (\mu_1(t) - \mu_2(t))^2 \quad (1)$$

In the formula, $\sigma_B^2(t)$ is inter-class variance, $w_1(t)$ and $w_2(t)$ are the pixel ratios of foreground and background. $\mu_1(t)$ and $\mu_2(t)$ are the average grayscale values of the foreground and background.

Ultimately, we only retain the foreground regions with pixel values greater than the threshold t, which are tissue regions in the slide. Besides, considering that some slides in the TCGA dataset may contain similar pathological tissues and a slide has a large size, it is unnecessary to include all the patches for pre-training. To preserve the contextual relationship between selected patches, we locally sample image regions of 1024 × 224 × 224 (~1024 images) from each pathological image, ensuring that the sampled region has a tissue proportion

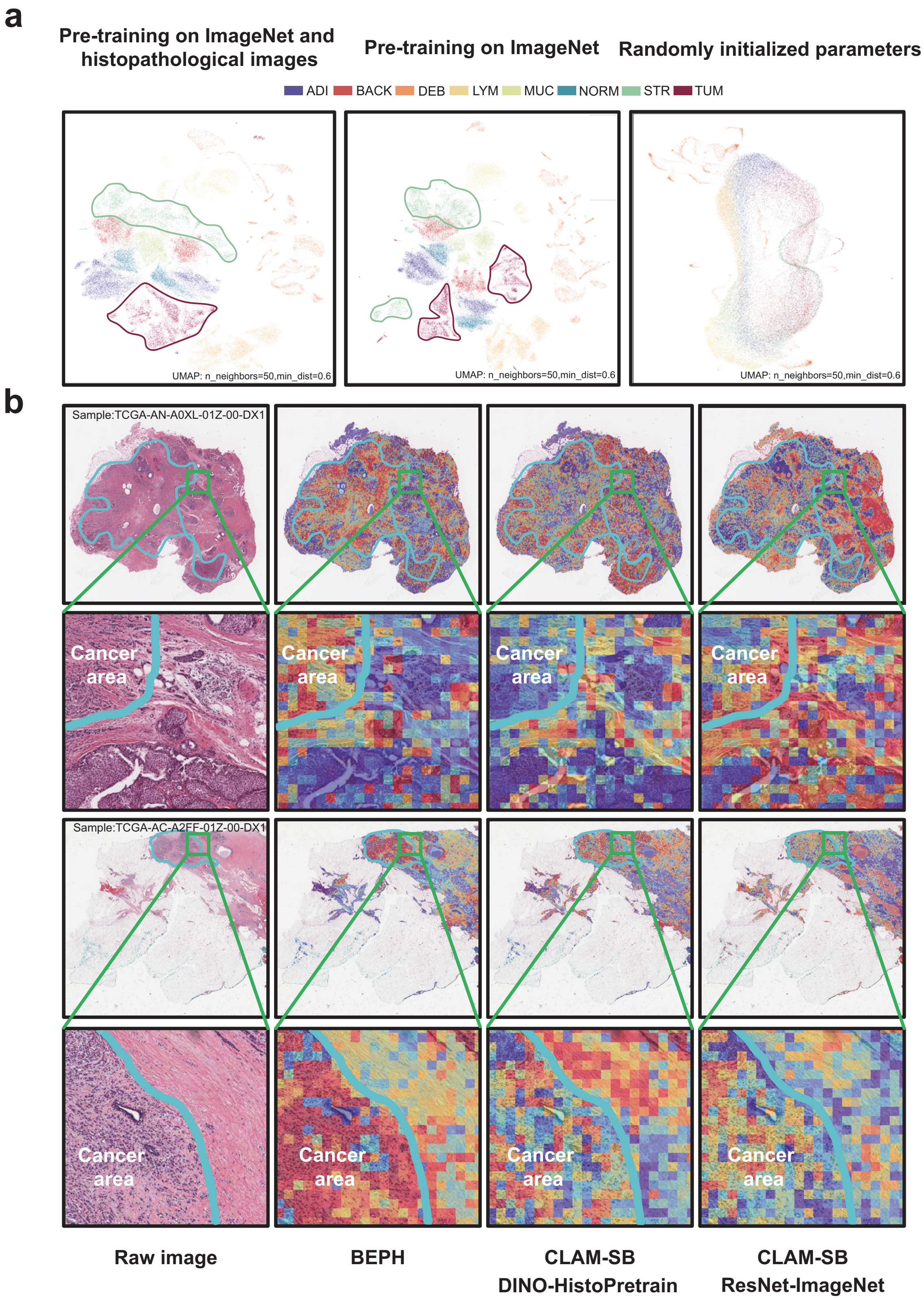

a
Pre-training on ImageNet and histopathological images
Pre-training on ImageNet
Randomly initialized parameters
ADI
BACK
DEB
LYM
MUC
NORM
STR
TUM
UMAP: n_neighbors=50,min_dist=0.6
UMAP: n_neighbors=50,min_dist=0.6
UMAP: n_neighbors=50,min_dist=0.6
b
Sample:TCGA-AN-A0XL-01Z-00-DX1
Cancer area
Cancer area
Cancer area
Cancer area
Sample:TCGA-AC-A2FF-01Z-00-DX1
Cancer area
Cancer area
Cancer area
Cancer area
Raw image
BEPH
CLAM-SB DINO-HistoPretrain
CLAM-SB ResNet-ImageNet

Fig. 5 | BEPH Model Interpretability and Visualization. a The interpretability and visualization of the pre-trained model are demonstrated after extracting features from the NCT-CRC-HE-100K dataset and applying UMAP dimensionality reduction. The closer the instances of the same class are, the better the model's feature extraction capability. **b** The figure displays automatically calculated attention heatmaps covering WSIs. The warm colors (red) in the heatmap indicate higher estimated probabilities of tumor tissue. The blue contour lines represent cancer regions annotated by pathologists, and the green boxes denote local areas containing the boundary between tumor and normal tissue. From left to right, the first column shows the ground truth, the second column displays BEPH, the third column shows DINO-HistoPretrain, and the fourth column presents ResNet-ImageNet. The odd rows depict the model's attention on the entire tumor region of the WSIs, while the even rows show the heatmaps for the corresponding local area within the box. These slides are from the TCGA-BRCA dataset.

Table 1 | Ablation experiments on the effectiveness of pre-training

	BreakHis (Mean magnifications)		WSI weakly classification		Survival prediction
	Patient-level accuracy	Image-level accuracy	NSCLC 25%	NSCLC 100%	BRCA
BEiTv2-ImageNet	93.30 ± 1.51	92.89 ± 2.07	0.8788 ± 0.0218	0.9270 ± 0.0131	0.6440 ± 0.0354
BEPH (mask 10%)	92.45 ± 1.21	91.18 ± 0.85	0.7043 ± 0.0337	0.8496 ± 0.0200	0.5400 ± 0.0439
BEPH (mask 40%)	**94.05 ± 1.39**	**93.65 ± 0.67**	**0.9095 ± 0.0332**	**0.9719 ± 0.0054**	**0.6643 ± 0.0330**

The bold text in the table indicates the best results across different tasks.

greater than 75%. These sampled image regions are then cropped into 224 × 224 tiles while maintaining a tissue proportion of 75% and above within each patch (Supplementary Fig. 3). We performed these operations on a cluster of up to 2 nodes, where each node was equipped with 32 CPU cores and 512 GB RAM, completing preprocessing in about 65 h. Tasks were parallelized, so that each node processed a set of tiles independently. In the end, we obtained an unlabeled pre-training dataset consisting of 11.77 million tiles. The distribution of patches and the slide counts are detailed in Fig. 1b. This pre-training dataset is 10 times larger than the ImageNet-1K dataset (1.28 M).

Details of BEPH pre-training

The BEPH architecture is based on a default configuration of BEiTv2, which consists of an autoencoder and an encoder similar to BERT. The MIM model in this study is based on the BEiTv2 architecture for two reasons. First, BEiT pioneered the MIM method, and BEiTv2 outperforms models like MAE and CAE[49] comprehensively. Second, both BEiT and BEiTv2 use discrete visual tokens, which are more suitable for classification tasks. Models like MAE struggle to learn features at multiple scales and finer granularity, resulting in uniform features from pathological images and limited exploration of deep semantic information. However, MAE has stronger interpretability due to its ability to reconstruct RGB pixels.

The BEPH pre-training procedure is outlined as follows: the encoder encodes the original images into visual tokens using an autoencoder, VQ-KD, and then randomly masks 40% of the tokens. Next, the token sequences are inputted into ViT-base. The pretext task for pre-training is to predict the masked visual tokens of the original image based on the unmasked image token. This approach is similar to the concept of Masked Autoencoder (MAE), but it predicts visual tokens rather than normalized pixels.

In BEiTv2, to address the discrepancy between block-level pre-training and image-level representation aggregation, a Patch Aggregation structure is added during pre-training. This structure shares the information of the [CLS] token and the parameters of the MIM head to better learn global representations.

The overall loss function minimizes the difference between the predicted tokens and the ground truth tokens (Loss_{MIM}), and the global loss ($\text{Loss}_{[\text{CLS}]}$).

In our experiments, the model parameter size is 192.55 M. The image size is 224 × 224 and the batch size is 1024. The optimizer is AdamW and the initial learning rate is 0.0015. Finally, our model was pre-trained for a total of 340 h using 8 NVIDIA A100 GPUs. The pre-training duration was based on observations of the pre-training loss, which gradually plateaued as training progressed. Additionally, the pre-training time was aligned with previous studies by Zhou et al. [15], Bao et al. [27], and Wang et al. [10], which followed similar model architectures and pre-training procedures.

Competing methods and benchmarks

Competing methods in various downstream tasks. In the patch-level classification task, we compare BEPH against both convolutional neural network (CNN) models, which are specifically designed for the datasets, and self-supervised models. CNN models such as Deep[33], SW[31], GLPB[34], RPDB[35], shallow-CNN[38], AlexNet[39], ResNet[39], VGG19[39], and EfficientNet-B0[40]. They all have a convolution architecture, with differences lying in data preprocessing, sampling techniques, and slight variations in network structure. MIL-NP[32]and MILCNN[32] organize instances (images) into bags (patients) and provide a weakly supervised framework for image recognition by combining multi-instance learning loss with deep learning networks. Self-supervised models like MPCS-RP[36] and DARC-ConvNet[41] perform label-free pre-training through specifically designed loss functions. MPCS-RP employs a paired sampling method that leverages intrinsic dataset signals (e.g., magnification factors in BreakHis) to generate meaningful contrastive views, reducing reliance on human-induced priors. DARC-ConvNet iteratively clusters learned representations, using cluster assignments as pseudo-labels to optimize network parameters.

In the WSI weakly supervised classification and survival prediction tasks, our comparative methods include various pre-trained and weakly supervised models. HIPT[21] was pre-trained on 10.68 thousand gigapixel WSIs from TCGA, employing a hierarchical image pyramid transformer with 256 × 256 and 4096 × 4096 image views. CLAM(Resnet)[4] and CLAM(DINO)[4,21] are based on the CLAM weak supervision framework: CLAM(Resnet) uses the original model structure reported by Lu et al. with an ImageNet-1k pre-trained Resnet-50 backbone, while CLAM(DINO) uses a DINO backbone pre-trained on TCGA WSIs, with weights derived from HIPT's 256 × 256 pre-trained model. CLAM(CtransPath), CLAM(UNI), CLAM(GigaPath), and CLAM(-CHIEF) use the backbone of foundation models: CtransPath, UNI, GigaPath, and CHIEF, respectively. CtransPath employed a semantically relevant contrastive learning objective, treating each image and its augmentation views as positive pairs and retrieving semantically related images as pseudo-positive pairs. UNI leverages large-scale pre-training across diverse datasets for superior generalization. GigaPath consists of a tile encoder that captures local features and a slide encoder that captures global features, utilizing extensive pre-training on gigapixel WSIs for robust representation learning. CHIEF applies weakly SL for enhanced self-supervised feature extraction. The GCN-MIL[8,21] method integrates MIL and deep GCN by using VAE-GAN for self-supervised feature extraction, along with feature selection to refine instance features, and finally aggregates bag-level features using GCN for classification. DS-MIL[21,23]achieves instance-level and bag-level classification through max-pooling and non-local operations, leveraging self-supervised contrastive learning to enhance feature representation and multi-scale feature fusion to integrate image details at different magnifications. DeepAttnMISL[21,42] extracts patches from WSIs using pre-trained models like VGG, clusters them into phenotype groups, and leverages a Siamese multi-instance fully convolutional network (MI-FCN) to learn representations that capture patient heterogeneity.

Patch-level classification

When adapting to downstream tasks, we only use the encoder of BEPH. For patch-level classification tasks, we attach a fully connected layer to the encoder's output. The encoder processes the input image and encodes it into a 768-dimensional feature vector, which is then passed to the fully connected layer to predict the probabilities of cancer categories. The category with the highest probability is defined as the final category. The number of categories determines the number of neurons in the last layer.

For BreakHis, the dataset was randomly divided into training sets (80%) and validation sets (20%) for five repeated trials. If patches were classified based on patients, they were proportionally allocated by patients. The average results of the validation sets from the five trials were taken as the final result. For LC25000, the dataset was divided into 5 folds, and the average validation results from the 5-fold cross-validation were reported as the final result.

The training object is to generate classification outputs that match the labels. Thus, the loss function is cross-entropy loss. The batch size is set to 128, and the total number of training epochs is 30. The first 20 epochs are used for learning rate warm-up (from 0.0001 to a learning rate of 0.005), followed by a cosine annealing schedule where the learning rate decreases from 0.005 to 1×10^{-6} for the remaining 10 epochs.

Performance evaluation for patch-level classification. In patch-level classification tasks, accuracy is used to measure performance, which includes two scales: image-level accuracy and patient-level accuracy (if applicable). Patient-level accuracy reflects the performance of classification at the patient level and is calculated as the average of all patient scores. The patient score represents the proportion of correctly classified images within a patients' set of images. On the other hand, image-level accuracy does not consider patient-level details. It is calculated as the proportion of correctly classified images among all images, regardless of the patients they belong to[36].

$$Patient\ Score = \frac{N_{cpi}}{N_p} \tag{2}$$

Let N_p represents the number of pathological images owned by patient p. For each patient, N_{cpi} is the number of pathological images that are correctly classified.

$$Patient\ level\ accuracy = \frac{\sum Patient\ Score}{Total\ Number\ of\ Patients} \tag{3}$$

$$Image\ level\ accuracy = \frac{N_{ci}}{N_{all}} \tag{4}$$

N_{all} represents the total number of pathological images from all patients, and N_{ci} represents the number of correctly classified pathological images among them.

WSI weakly supervised classification. For weakly supervised classification tasks, we freeze the encoder (ViT-base) and only use it as a feature extractor. The downstream structure is based on an attention-based MIL framework similar to CLAM. Using CLAM as a WSI-level classification model is based on three reasons. First, since only a small portion of the image is relevant to its class label, performing supervised classification at the patch level would introduce significant bias by labeling all patches as the same class. Second, CLAM, as an attention-based MIL model, handles imprecise annotations effectively. The plug-and-play functionality allows efficient deployment of pre-training outcomes, reducing implementation complexity. Third, BEiT, similar to large language models, typically possesses contextual learning capabilities, which can compensate for CLAM's limitation of failing to learn potential nonlinear interactions between patches. The workflow of this task is: the encoder generates high-dimensional features from the pathology images, and then the attention network aggregates patch-level information into WSI-level representations for making final diagnostic predictions.

In the subtyping classification task, we reserve 10% of each dataset as a test set and the remaining 90% for 10-fold Monte Carlo cross-validation. In each fold, the model's performance on the validation set was monitored and used for model selection, while the test set was kept separate and referenced only once after training for model evaluation. The dataset partitioning was slightly different from CLAM because CLAM fixed the validation set. To examine the model's performance on training data, both the training and validation sets should be considered as part of the training data. Although reducing the validation set may affect model selection, it can test the model's stability. All data for training, validation, and testing were performed at the max magnification.

All models are trained for 20 epochs with early stopping. Early stopping is triggered when the validation AUC does not improve from the previous highest point for consecutive 10 epochs. The learning rate is adjusted slightly based on the cancer types and the proportion of training data, typically ranging from 2e-4 to 8e-4. Other parameters remain consistent with CLAM.

Performance evaluation for WSI-level classification. The performance is evaluated using the Area Under the Curve (AUC). This metric is derived from the receiver operating characteristic curve of the classifier. For the BRCA and NSCLC subtype classification tasks, AUC is calculated between the two classes. For RCC subtype classification, AUC is calculated for each disease category separately, and then the average AUC is obtained.

WSI survival prediction

For the survival prediction task, we improved the final output layer of CLAM based on the survival model proposed by Wei et al. [50]. We named this structure CLAMSurvival (Fig. 1d). While the architectural design parallels that of weakly-supervised classification, the proposed framework introduces a linear output layer in the final MIL stage to generate risk scores and incorporates the Cox proportional hazards model for survival prediction, thereby transforming the pipeline into a survival prediction model[50,51]. It's important to note that the MIL framework is modified to incorporate clinical information.

The features of the histopathological images are extracted by the encoder. The MIL framework is then utilized to estimate the risk function $\hat{h}_\theta(x_i)$. The parameters of the network are trained based on the occurrence time (death) events to optimize the Cox likelihood function.

$$L(\theta) = \prod_{i=1}^{n} \left[\frac{\exp\left(\hat{h}_\theta(x_j)\right)}{\sum_{j \in R(X_i)} \exp\left(\hat{h}_\theta(x_j)\right)} \right]^{\delta_i} \tag{5}$$

When we minimize negative log partial likelihood, the loss function is:

$$l(\theta) = -\sum_{i=1}^{n} \delta_i \left(\hat{h}_\theta(x_i) - \log \sum_{j \in R(X_i)} \exp\left(\hat{h}_\theta(x_j)\right) \right) \tag{6}$$

δ_i is an indicator of whether the survival time is censored ($\delta_i = 0$) or observed ($\delta_i = 1$). θ is the weight of the MIL. $\mathscr{R}(X_i)$ denotes the set of individuals who are at risk for failure time of individual i.

The data was divided using 5-fold cross-validation, and both training and validation were performed at a max magnification level. The model undergoes a maximum of 20 epochs of training, and early stopping is implemented to prevent overfitting. Early stopping is triggered when the validation loss does not decrease from the previous lowest point for consecutive 5 epochs. The learning rate is adjusted slightly depending on the cancer type, ranging from 2e-4 to 2e-3. This range allows the model to find an optimal learning rate for each specific cancer type. Other parameters remain consistent with CLAM, preserving coherence in the methodology and framework.

Performance evaluation for survival prediction. For survival prediction, the Concordance Index (C-Index) is a frequently used evaluation metric. Its primary purpose is to assess the survival models' capacity to rank individual survival risks, where a higher prediction of risk is anticipated if the patient succumbs earlier. It can be written by:

$$C - Index = \frac{1}{M} \sum_{i:o_i=1} \sum_{j:t_i<t_j} I\left[\hat{y}_i(x_i) > \hat{y}_j\left(x_j\right)\right] \tag{7}$$

where M is the number of comparable pairs and $I[\cdot]$ is the indicator function. It ranges from 0.0 to 1.0. The larger the C-Index is, the better the model performance could be. The 95% confidence interval (CI) is obtained by 1.96 × standard error.

Kaplan–Meier survival analysis and prognostic risk grouping. To evaluate survival outcomes, we employed Kaplan–Meier survival analysis to estimate survival probabilities over time. Patients are categorized into either a high-risk prognostic group (short-term survivors) or a low-risk group (long-term survivors) based on their hazard predictions compared to a predefined cutoff value. This cutoff value is determined by the median of hazard predictions. To evaluate the various pre-training methods, we employ CLAMSurvival on six cancer datasets and compare the performance of BEPH, DINO (contrastive learning pre-training), and ResNet-50 (natural image supervised pre-training). The log-rank test is utilized to calculate p-values, which assess the differences in survival between the risk groups. It is important to note that the patients' hazard predictions included in the risk grouping analysis are from the validation set of the 5-fold cross-validation.

Feature space visualization and heatmap visualization

Visualization of feature extraction capabilities of pre-trained backbone. In examining the feature extraction capability of the BEPH's backbone, We processed the NCT-CRC-HE-100K dataset through the backbone and extracted 768-dimensional feature vectors from the final layer's outputs for each patch. As high-dimensional data visualization is difficult, we employed Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction to two dimensions, with the set neighborhood size of 50 and the minimum distance parameter of 0.6, ensuring the preservation of local and global structure in the feature space. When visualizing, we combine MUS and STR into one category.

WSI-level feature space visualization. To visualize the WSI-level feature space, we analyzed slides on the independent test set by leveraging models fine-tuned with different proportions of training data. For each slide, a 384-dimensional feature representation was extracted after the first fully connected layer of the classification model. This feature vector was then projected onto a two-dimensional space using PCA to facilitate visualization and comparative analysis.

Quantitative evaluation of feature space. For a quantitative evaluation of the feature space, we formulated the task as a clustering evaluation problem without applying an explicit clustering algorithm. Instead, clustering metrics were directly computed based on ground-truth labels, treating each label as a distinct cluster. The Davies–Bouldin Score[52,53], Silhouette Score[54,55], and Neighborhood Consistency[56] were calculated, providing metrics for the intra-class compactness and inter-class separability of the feature representations. This approach allowed us to quantitatively measure each model's ability to get feature embeddings, offering insight into the discriminative power of each model's learned feature space.

Model prediction interpretation through attention heatmaps. To interpret the relative importance of different regions for the model's WSI-level prediction, we utilized the attention branch corresponding to the model's predicted subtypes. Non-normalized attention scores were computed and recorded for all 224 × 224 patches extracted from the slides, before applying the softmax function to convert these scores into a probability distribution. These attention scores were then converted into percentile scores and scaled between 0 and 1.0, with 1.0 indicating the highest level of attention and 0 indicating the lowest. To visually highlight these scores, a diverging color map was applied to the normalized scores, converting them into RGB colors. These colors were overlaid on the respective spatial positions within the slides, allowing for clear visual identification of high-attention areas (displayed in red) that contribute positively to the model's prediction, and low-attention areas (displayed in blue), indicating low influence on the prediction relative to other patches.

Quantitative evaluation of attention heatmaps. While attention heatmaps generated by weakly supervised CLAM models are not designed for pixel-level annotation of ROIs, we conducted a quantitative evaluation to assess the feasibility of using these heatmaps as auxiliary annotation tools in clinical and research settings, and to verify the performance of the model's attention. For this purpose, we downloaded publicly available cancer region annotations from https://zenodo.org/records/5320076, where annotations were performed by trained observers using QuPath v0.1.2 and later converted to CSV format. Using these annotations, we segmented each slide into cancerous and non-cancerous regions, obtaining a 0–1 attention score for each 224 × 224 patch within each region type. The mean attention scores were then calculated for both cancerous and non-cancerous regions. Finally, we quantitatively evaluated the effectiveness of different pre-trained backbones by comparing the cumulative mean attention scores between cancerous and non-cancerous regions, and by counting the number of images where the mean attention score for cancerous regions exceeded that for non-cancerous regions (Supplementary Fig. 6).

Online fine-tuning in the WSI classification task

Although with the complete open source of code repositories such as GitHub, many researchers can easily access, and use advanced foundational models. there are still significant barriers for researchers or doctors who have not been exposed to artificial intelligence to deploy and reproduce code. Therefore, in order to promote automated diagnosis of various cancers, we will deploy BEPH to http://yulab-situ.natappl.cc/BEPH for use by all researchers. At present, researchers can complete online fine-tuning of WSI classification tasks by uploading WSIs data according to the tutorial.

The web interface and API layer of our service were implemented using Python's FastAPI framework, which enable efficient handling of multiple concurrent requests. FastAPI's lightweight architecture and high-performance features enable low-latency responses and support for large-scale user interactions. On the backend, we utilized PyTorch as the primary deep learning framework due to its flexibility and robust support for GPU-accelerated computations. PyTorch allowed us to design and implement custom fine-tuning pipelines that dynamically adjust model parameters based on user-provided data. These pipelines facilitate model updates in real-time, enabling adaptive fine-tuning workflows without the need for extensive re-training.

Statistical analysis

We employed statistical tests to evaluate the significance of model performance across various tasks. For the patch-level classification task, where each method was tested independently, we conducted independent two-sample *t*-tests to assess performance differences between models at both patient-level and image-level, following the approaches of Matthew et al. [57] and Zeng et al. [58]. In the WSI-level classification task, we used the Wilcoxon signed rank test, as recommended by Wei et al. [50], ensuring robustness in cases where data did not meet normality assumptions. For the survival prediction task, we applied paired t-tests along with Cohen's d to quantify effect sizes and performance differences across six cancer types, following Fan et al. [59], Tantithamthavorn et al. [60], and Hosseini et al. [61]. This combination of statistical analyses rigorously quantify the performance differences between models, facilitating a comprehensive comparison of their capabilities.

Datasets used in the experiment

The pre-training dataset. The pre-training data used in this study are available from TCGA. The TCGA diagnostic pathology images data can be obtained from the National Cancer Institute's Genomic Data Commons (https://portal.gdc.cancer.gov).

BreakHis. The BreakHis dataset consists of 2480 histopathological microscope images of benign tissue and 5429 images of malignant tissue from 82 patients. The images were captured at magnifications of 40 × , 100 × , 200 × , and 400 × . Each image in the BreakHis dataset has a size of 700 × 460 pixels and was stained using the Hematoxylin and Eosin (H&E) staining technique. In the experiment, we compressed the dataset to 224 × 224 and grouped it by patient. The dataset is available at https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/.

LC25000. The LC25000 dataset comprises 25,000 color images divided into 5 categories, with each category containing 5000 images. All images have a size of 768 × 768 pixels and are stored in the JPEG file format. The dataset consists of the following 5 classes: colon adenocarcinoma, benign colon tissue, LUAD, LUSC, and benign lung tissue. 15,000 images of lung tissue (5000 images of benign lung tissue, 5000 images of LUSC, and 5000 images of LUAD) were used in the experiment. The dataset is available at https://github.com/tampapath/lung_colon_image_set.

CAMELYON16. The CAMELYON16 dataset is a dataset for the detection of BRCA with lymph node metastasis. It consists of 270 annotated WSIs for training and another 129 slides as a held-out official test set. The dataset is collected by the Radboud University Medical Center and the University Medical Center Utrecht in the Netherlands. The dataset is available at https://camelyon16.grand-challenge.org/.

BACH. The dataset comprises the H&E stained breast histology whole-slide images in four classes: normal, benign, in situ carcinoma, and invasive carcinoma. We divide them into two classes: normal and tumor. The dataset contains 20 unlabeled normal, 10 pixel-wise labeled, and 20 non-labeled slides. We only used normal and labeled slides. The dataset is available at https://iciar2018-challenge.grand-challenge.org/Dataset/.

NCT-CRC-HE-100K. The NCT-CRC-HE-100K dataset is a collection of 100,000 non-overlapping image patches extracted from 86 human cancer tissue slices and normal tissue. The tissue slices were stained with H&E. The dataset is sourced from the NCT Biobank (National Center for Tumor Diseases) and UMM Pathology Archive (Mannheim University Medical Center). Pathologists manually annotated the tissue regions into nine tissue categories, which include: adipose tissue (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), stroma related to cancer (STR), and colorectal adenocarcinoma epithelium (TUM). The dataset is available at https://zenodo.org/records/1214456. When we visualize, we combine MUS and STR into one category.

The public TCGA-BRCA dataset. The public TCGA-BRCA dataset consists of 875 diagnostic images and corresponding labels from the TCGA-BRCA database (https://portal.gdc.cancer.gov/projects/TCGA-BRCA). Among them, 726 slides are derived from 694 cases of IDC, while 149 slides are derived from 143 cases of ILC.

The public TCGA-NSCLC dataset. In our study, the public TCGA-NSCLC dataset consists of 958 diagnostic images and corresponding labels from the TCGA-LUSC and TCGA-LUAD projects. It includes 492 LUAD (LUAD) slides from 430 cases and 466 LUSC slides from 432 cases.

The public TCGA-RCC WSI dataset. The TCGA-RCC diagnostic images and corresponding labels are available from the NIH genomic data commons (https://portal.gdc.cancer.gov). In our study, The public TCGA-RCC dataset consists of 903 diagnostic images, including TCGA-KICH, TCGA-KIRC, and TCGA-KIRP projects. It includes 118 CRCC (Chromophobe RCC) slides from 107 cases, 497 CCRCC (Clear Cell RCC) slides from 491 cases, and 288 PRCC (Papillary RCC) slides from 266 cases.

The public survival prediction WSI dataset. The subtype dataset includes IDC and ILC pathology images and clinical information from the TCGA-BRCA database, comprising 726 images from 694 patients. The CRC (Colorectal Cancer) subtype dataset includes 315 images from 315 patients, sourced from the TCGA-READ and TCGA-COAD projects. The CCRCC subtype dataset is sourced from the TCGA-KIRC project and consists of 497 images from 491 patients, while the PRCC subtype dataset is sourced from the TCGA-KIRP project and consists of 288 images from 266 patients. The LUAD dataset, a subtype of lung cancer, is sourced from the TCGA-LUAD project and consists of 492 images from 430 patients. The STAD subtype dataset is one of the gastric cancer subtypes, including 348 pathology images from 348 patients, sourced from the TCGA-STAD project.

It should be noted that our dataset size is derived from HIPT and CLAM references, and therefore the number of WSIs may not align precisely with the latest TCGA database.

Computing hardware and software

We used Python (v3.9.17) and PyTorch (v1.12.0, CUDA 11.3) (https://pytorch.org) for all experiments and analyses in the study (unless specified). For training BEPH via BEiTv2, we employed the open-source mmselfsup framework from Openmmlab (https://github.com/open-mmlab/mmselfsup) for pre-training. The training utilized 8 × 4 40 GB NVIDIA A100 GPU nodes, set up for multi-GPU, multi-node training with distributed data-parallel. All other computations for downstream experiments were conducted on single 40 GB NVIDIA A100 GPU. All WSI processing was supported by openslide-python (v1.3.0), and CLAM.

All charts are generated using R (4.2.1). The AUC is estimated using the Mann–Whitney U statistical method. The evaluation metrics algorithm implementation is provided by Scikit-learn (v1.3.2).

The online fine-tuning service is developed using a high-performance Lenovo ThinkStation ST558 tower server workstation, equipped with dual Intel Gold 6230 CPUs (40 cores, 2.1 GHz), 64 GB of RAM, dual 960 GB SSDs, three 12 TB HDDs, and an NVIDIA RTX A4000 GPU. This infrastructure was optimized to support scalable AI-driven tasks, especially for deep learning and data-intensive operations.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data analyzed in this manuscript are available from the public datasets. TCGA diagnostic pathology image data (BRCA, NSCLC, RCC, CRC, CCRCC, PRCC, LUAD, STAD) and corresponding clinical data can be obtained from the National Cancer Institute's Genomic Data Commons (https://portal.gdc.cancer.gov). BreakHis can be publicly accessed at https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/. The LC25000 dataset can be obtained from https://github.com/tampapath/lung_colon_image_set. NCT-CRC-HE-100K can be obtained from https://zenodo.org/records/1214456. The CAMELYON16 dataset is available at https://camelyon16.grand-challenge.org/. The BACH dataset is available at https://iciar2018-challenge.grand-challenge.org/Dataset/. Source data are provided with this paper.

Code availability

All code is implemented in Python, using PyTorch as the main deep-learning library. The pre-training framework relies on https://github.com/open-mmlab/mmselfsup. Our code and pre-trained weights can be obtained from https://github.com/Zhcyoung/BEPH or Zenodo[62] and can be used to replicate the experiments in this paper. All source code is released under the GNU GPLv3 free software license.

References

1. Zeiser, F. A. et al. Breast cancer intelligent analysis of histopathological data: a systematic review. *Appl. Soft Comput.* **113**, 107886 (2021).
2. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: a survey. *Med. Image Anal.* **67**, 101813 (2021).
3. Kim, I., Kang, K., Song, Y. & Kim, T.-J. Application of artificial intelligence in pathology: trends and challenges. *Diagnostics* **12**, 2794 (2022).
4. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
5. Acs, B., Rantalainen, M. & Hartman, J. Artificial intelligence as the next step towards precision pathology. *J. Intern. Med.* **288**, 62–81 (2020).
6. He, K. et al. Transformers in medical image analysis. *Intelligent Medicine* **3**, 59–78 (2023).
7. Lu, M. et al. SMILE: sparse-attention based multiple instance contrastive learning for glioma sub-type classification using pathological images. In *Proc. MICCAI Workshop on Computational Pathology* 159–169 (PMLR, 2021).
8. Zhao, Y. et al. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 4836–4845. https://doi.org/10.1109/CVPR42600.2020.00489 (IEEE, 2020).
9. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
10. Wang, X. et al. TransPath: transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021* (eds. de Bruijne, M. et al.) 186–195. https://doi.org/10.1007/978-3-030-87237-3_18 (Springer International Publishing, 2021).
11. Bagherian, M. et al. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Brief. Bioinform.* **22**, 247–269 (2021).
12. Hou, X. et al. Dual adaptive pyramid network for cross-stain histopathology image segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019* (eds. Shen, D. et al.) 101–109. https://doi.org/10.1007/978-3-030-32245-8_12 (Springer International Publishing, 2019).
13. Fei, N. et al. Towards artificial general intelligence via a multimodal foundation model. *Nat. Commun.* **13**, 3094 (2022).
14. Ding, N. et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mach. Intell.* **5**, 220–235 (2023).
15. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
16. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at http://arxiv.org/abs/2108.07258 (2022).
17. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* 1–13. https://doi.org/10.1038/s41591-024-02857-3 (2024).
18. Khan, S. et al. Transformers in vision: a survey. *ACM Comput. Surv.* **54**, 1–41 (2022).
19. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proc. 37th International Conference on Machine Learning* 1597–1607 (PMLR, 2020).
20. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)* 9729–9738 (IEEE, 2020).
21. Chen, R. J. et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16144–16155 (IEEE, 2022).
22. Xie, Z. et al. Revealing the dark secrets of masked image modeling. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14475–14485. https://doi.org/10.1109/CVPR52729.2023.01391 (IEEE, 2023).
23. Li, B., Li, Y. & Eliceiri, K. W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14313–14323. https://doi.org/10.1109/CVPR46437.2021.01409 (IEEE, 2021).
24. Coudray, N. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
25. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
26. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416.e11 (2018).
27. Bao, H., Dong, L., Piao, S. & Wei, F. BEiT: BERT Pre-Training of Image Transformers. in International Conference on Learning Representations (2022).
28. Peng, Z., Dong, L., Bao, H., Ye, Q. & Wei, F. BEiT v2: masked image modeling with vector-quantized visual tokenizers. (2022).
29. Azizi, S. et al. Big self-supervised models advance medical image classification. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 3458–3468. https://doi.org/10.1109/ICCV48922.2021.00346 (IEEE, 2021).
30. Spanhol, F. A., Oliveira, L. S., Petitjean, C. & Heutte, L. A dataset for breast cancer histopathological image classification. *IEEE Trans. Biomed. Eng.* **63**, 1455–1462 (2016).
31. Spanhol, F. A., Oliveira, L. S., Petitjean, C. & Heutte, L. Breast cancer histopathological image classification using Convolutional Neural Networks. In *Proc. International Joint Conference on Neural Networks (IJCNN)* 2560–2567. https://doi.org/10.1109/IJCNN.2016.7727519 (IEEE, 2016).
32. Sudharshan, P. J. et al. Multiple instance learning for histopathological breast cancer image classification. *Expert Syst. Appl.* **117**, 103–111 (2019).
33. Spanhol, F. A., Oliveira, L. S., Cavalin, P. R., Petitjean, C. & Heutte, L. Deep features for breast cancer histopathological image

classification. In *Proc. International Conference on Systems, Man, and Cybernetics (SMC)* 1868–1873. https://doi.org/10.1109/SMC.2017.8122889 (IEEE, 2017).
34. Mpinda Ataky, S. T., de Matos, J., Britto, A. S., Oliveira, L. E. S. & Koerich, A. L. Data augmentation for histopathological images based on Gaussian-Laplacian pyramid blending. In *Proc. International Joint Conference on Neural Networks (IJCNN)* 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206855 (IEEE, 2020).
35. Man, R., Yang, P. & Xu, B. Classification of breast cancer histopathological images using discriminative patches screened by generative adversarial networks. *IEEE Access* **8**, 155362–155377 (2020).
36. Chhipa, P. C. et al. Magnification prior: a self-supervised method for learning representations on breast cancer histopathological images. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 2716–2726. https://doi.org/10.1109/WACV56688.2023.00274 (IEEE, 2023).
37. Borkowski, A. A. et al. Lung and colon cancer histopathological image dataset (LC25000). Preprint at https://doi.org/10.48550/arXiv.1912.12142 (2019).
38. Mangal, S., Chaurasia, A. & Khajanchi, A. Convolution Neural Networks for diagnosing colon and lung cancer histopathological images. Preprint at https://doi.org/10.48550/arXiv.2009.03878 (2020).
39. Abbas, M. A., Bukhari, S. U. K., Syed, A. & Shah, S. S. H. The histopathological diagnosis of adenocarcinoma & squamous cells carcinoma of lungs by artificial intelligence: a comparative study of convolutional neural networks. 2020.05.02.20044602 Preprint at https://doi.org/10.1101/2020.05.02.20044602 (2020).
40. Mishra, S. & Agarwal, U. Lung cancer detection (LCD) from histopathological images using fine-tuned deep neural network. In *Proc. International Conference on Intelligent Computing, Communication and Information Security* (eds. Devedzic, V., Agarwal, B. & Gupta, M. K.) 249–260. https://doi.org/10.1007/978-981-99-1373-2_19 (Springer Nature, 2023).
41. Li, J. et al. DARC: deep adaptive regularized clustering for histopathological image classification. *Med. Image Anal.* **80**, 102521 (2022).
42. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N. & Huang, J. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Med. Image Anal.* **65**, 101789 (2020).
43. Wang, X. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).
44. Ehteshami Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
45. Aresta, G. et al. BACH: Grand challenge on breast cancer histology images. *Med. Image Anal.* **56**, 122–139 (2019).
46. Shao, W. et al. Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images. *IEEE Trans. Med. Imaging* **40**, 3739–3747 (2021).
47. Kather, J. N. et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLOS Med.* **16**, e1002730 (2019).
48. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
49. Chen, X. et al. Context autoencoder for self-supervised representation learning. *Int. J. Comput. Vis.* **132**, 208–223 (2024).
50. Wei, T. et al. Survival prediction of stomach cancer using expression data and deep learning models with histopathological images. *Cancer Sci.* **114**, 690–701 (2023).
51. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. USA* **115**, E2970–E2979 (2018).
52. Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 224–227 (1979).
53. Elforaici, M. E. A. et al. Cell-level GNN-based prediction of tumor regression grade in colorectal liver metastases from histopathology images. In *Proc. International Symposium on Biomedical Imaging (ISBI)* 1–5. https://doi.org/10.1109/ISBI56570.2024.10635713 (IEEE, 2024).
54. ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
55. Cisternino, F. et al. Self-supervised learning for characterising histomorphological diversity and spatial RNA expression prediction across 23 human tissue types. *Nat. Commun.* **15**, 5906 (2024).
56. Ding, C. & He, X. K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization. In *Proc. ACM symposium on Applied Computing* 584–589. https://doi.org/10.1145/967900.968021 (Association for Computing Machinery, 2004).
57. Adams, M. P., Rahmim, A. & Tang, J. Improved motor outcome prediction in Parkinson's disease applying deep learning to DaTscan SPECT images. *Comput. Biol. Med.* **132**, 104312 (2021).
58. Zeng, L.-L. et al. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine* **30**, 74–85 (2018).
59. Fan, X. et al. DNA microarrays are predictive of cancer prognosis: a re-evaluation. *Clin. Cancer Res.* **16**, 629–636 (2010).
60. Tantithamthavorn, C., McIntosh, S., Hassan, A. E. & Matsumoto, K. An empirical comparison of model validation techniques for defect prediction models. *IEEE Trans. Softw. Eng.* **43**, 1–18 (2017).
61. Hosseini, S., Turhan, B. & Mantylä, M. A benchmark study on the effectiveness of search-based data selection and feature selection for cross project defect prediction. *Inform. Softw. Technol.* **296**, 312 (2018).
62. Yang, Z. et al. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. Zhcyoung/BEPH: BEPH v1.0. https://doi.org/10.5281/ZENODO.14449988 (2024).

Acknowledgments

The computations in this paper were run on the Siyuan-1 cluster supported by the Center for High Performance Computing at Shanghai Jiao Tong University. Thanks to everyone at Yu Lab who helped with this article. This study was supported by the National Natural Science Foundation of China (ID: 12171318), Shanghai Science and Technology Commission (ID: 23XD1401900, 23DZ2290600, 24JS2810200), Shanghai Jiao Tong University STAR Grant (ID: 20190102), Medical Engineering Cross Fund of Shanghai Jiao Tong University (ID: YG2023ZD21). Shanghai Jiao Tong University "Jiaotong Star" Plan Medical Engineering Cross Research Project (20230103).

Author contributions

Z.C.Y., T.W., Z.S.Y, and Y.Z. designed the BEPH algorithm and downstream task framework. Z.C.Y. implemented the pre-training of BEPH and applied and analyzed it in WSI-Classification and Survival prediction tasks. Z.S.Y. and Y.Z. raised funds for this study. Y.L. and T.W. were responsible for dataset preprocessing and performance testing of BEPH on patch-level tasks. X.Y., R.T.G., Y.J.X., and J.Z., provided valuable suggestions and model improvement comments. Z.C.Y. and T.W. wrote the manuscript. All authors have reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57587-y.

Correspondence and requests for materials should be addressed to Yue Zhang or Zhangsheng Yu.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.