npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01411-2

Aligning knowledge concepts to whole slide images for precise histopathology image analysis

Weiqin Zhao[1,4], Ziyu Guo[1,4], Yinshuang Fan[1], Yuming Jiang[2], Maximus C. F. Yeung[3] ✉ & Lequan Yu[1] ✉

Due to the large size and lack of fine-grained annotation, Whole Slide Images (WSIs) analysis is commonly approached as a Multiple Instance Learning (MIL) problem. However, previous studies only learn from training data, posing a stark contrast to how human clinicians teach each other and reason about histopathologic entities and factors. Here, we present a novel knowledge concept-based MIL framework, named **ConcepPath,** to fill this gap. Specifically, ConcepPath utilizes GPT-4 to induce reliable disease-specific human expert concepts from medical literature and incorporate them with a group of purely learnable concepts to extract complementary knowledge from training data. In ConcepPath, WSIs are aligned to these linguistic knowledge concepts by utilizing the pathology vision-language model as the basic building component. In the application of lung cancer subtyping, breast cancer HER2 scoring, and gastric cancer immunotherapy-sensitive subtyping tasks, ConcepPath significantly outperformed previous SOTA methods, which lacked the guidance of human expert knowledge.

The analysis of histopathology images is crucial in modern medicine, particularly for cancer diagnosis and prognosis, where it serves as the gold standard. However, analyzing histopathology images is time-consuming and labor-intensive for pathologists. Digitalizing histopathology images into high-resolution whole slide images (WSIs) has ushered in a new era for computer-aided analysis[1–3]. Owing to their enormous size (e.g., 150,000 × 150,000) and the lack of fine-grained annotations, WSI analysis is typically formulated as a Multiple Instance Learning (MIL) problem, which enables weakly supervised learning from slide-level labels. MIL-based methods typically begin by extracting the feature embeddings of image patches using a pre-trained network[4–7]. The feature embeddings are then fed into an aggregation network to generate slide-level predictions. Numerous research efforts have focused on efficiently aggregating information, including using attention-based weights[8] and leveraging spatial context information[9–11]. However, most current approaches in computational histopathology learn solely from image data, contrasting with how humans teach and reason about histopathologic entities and factors, as illustrated in Fig. 1a. Although a recent innovative study[12] explores the use of language priors in few-shot weakly supervised learning for WSI analysis, it suffers from unreliable language prior generation and unsatisfactory performance under full training setups, which limits its wide application in precise WSI analysis. Thus, incorporating valuable expert knowledge for precise WSI analysis remains an unsolved yet critical challenge.

With the rapid development of multimodal learning, there has been a surge of studies on CLIP-based pathology vision-language models[13–18]. Following the principle of Contrastive Language-Image Pre-training (CLIP)[15], these models learn well-aligned representation spaces between histopathology images and text description pairs collected from medical textbooks, scientific papers, public forums, and educational videos. One major benefit is that natural language descriptions can provide more expressive, dense, and interconnected representations beyond the scope of a single categorical label, linking diverse features of histopathology sub-patch structures[13,18]. Remarkable achievements have been made in transferring the above pathology vision-language models to a wide range of downstream tasks, including patch-level histopathology image classification, segmentation, captioning, and retrieval[18]. Meanwhile, large language models (LLMs) have shown great potential in performing logic, analogy, causal reasoning, extrapolation, and evidence evaluation for medical and scientific applications[19]. Researchers have found that when treated as reasoning machines or inference engines rather than knowledge databases, LLMs are less likely to generate false statements that do not reflect scientific facts[20]. These breakthroughs provide an opportunity to extract and incorporate human expert knowledge into the WSI analysis.

[1]School of Computing and Data Science, The University of Hong Kong, Hong Kong SAR, China. [2]School of Medicine, Wake Forest University, Winston-Salem, NC, USA. [3]Department of Pathology, The University of Hong Kong, Hong Kong SAR, China. [4]These authors contributed equally: Weiqin Zhao, Ziyu Guo.
✉e-mail: mcfyeung@hku.hk; lqyu@hku.hk

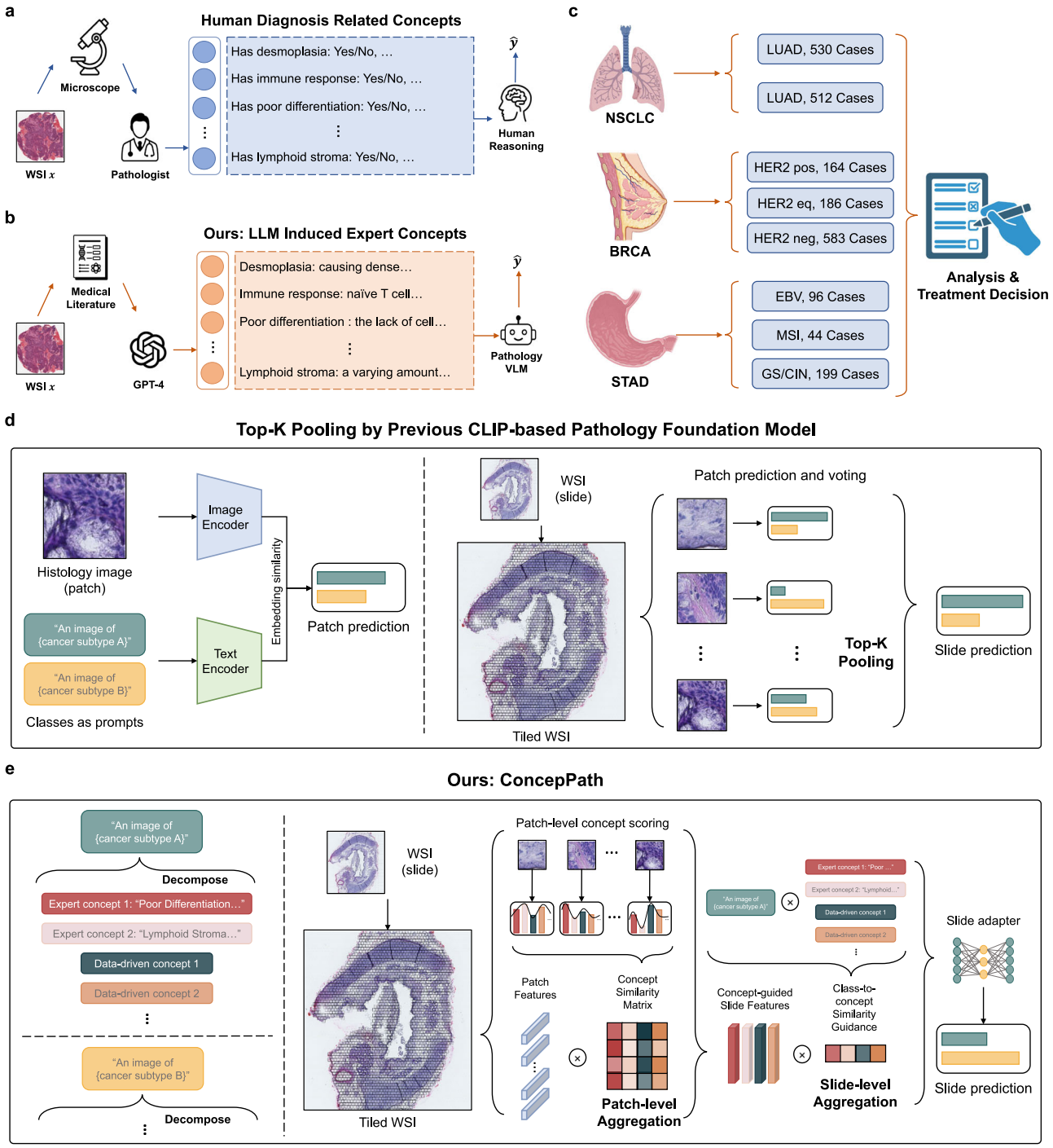


Fig. 1 | Overview of ConcepPath framework. a In real clinical processes, pathologists apply their expert knowledge to reason about histopathologic entities and factors to make a diagnosis. **b** ConcepPath utilizes a large language model like GPT-4 to induce expert concepts related to diagnosis from medical literature and integrate this knowledge into an automated WSI analysis pipeline through the CLIP-based pathology vision-language foundation model. **c** Dataset Characteristics of the evaluated tasks. **d** Left: An illustration of how the CLIP-based pathology foundation model performs patch prediction with class prompts. Right: The pipeline of how the previous CLIP-based pathology foundation model performs slide-level classification via a top-k pooling of patch predictions. **e** Left: An illustration of how ConcepPath decomposes a specific complex WSI analysis task into multiple subtasks of scoring patch-level concepts/attributes. Right: The pipeline of how ConcepPath conducts slide-level classification. Unlike the previous CLIP-based pathology foundation models' mechanism, ConcepPath leverages human prior knowledge and fully exploits the power of the CLIP-based pathology foundation model by scoring a group of expert concepts induced by GPT-4 from related medical literature and extracting complementary knowledge from training data via scoring a group of learnable data-driven concepts. The final prediction is produced with a two-stage aggregation mechanism with the above concepts.

In this paper, we present **ConcepPath**, a concept-based framework designed to make decisions by jointly leveraging the complementary human expert prior knowledge and data-driven concepts. The main idea of ConcepPath is illustrated in Fig. 1b, e. First, large language models such as GPT-4 are applied to induce reliable instance-level human expert concepts and bag-level expert class prompts from medical literature highly relevant to the given diagnostic task, as shown in Fig. 1b and Supplementary Fig. 2. It is important to note that, instead of treating GPT-4 as a knowledge database

and directly querying expert concepts, ConcepPath utilize GPT-4 as a reasoning machine to induce human expert knowledge concepts and class prompts from medical textbooks and academic papers. This strategy leads to more reliable expert concepts and class prompts generation. On the other hand, considering that extracted expert concepts may be insufficient to fully describe a disease's complexity[21,22], ConcepPath also learns complementary data-driven instance-level concepts from the training data with a group of purely learnable prompt representations, as shown in Fig. 1e. These learned concepts serve as a complement to expert concepts and play a crucial role, especially for complex and inadequately researched diagnostic tasks.

To transfer the knowledge contained in the concepts into WSI analysis, ConcepPath utilized a two-stage concept-guided hierarchical feature aggregation paradigm. With both concepts, class prompts, and instance features embedded into the well-aligned representation space via the CLIP-based pathology vision-language model as the basic building component, ConcepPath first aggregates instance features into concept-specific bag-level features under the guidance of instance-level concepts and then further aggregates the concept-specific bag-level features into the overall bag representation according to the correlations between instance-level concepts and bag-level expert class prompts. Finally, ConcepPath feeds the overall bag representation and bag-level concept embedding to slide adapters and calculates similarities between the adapted bag representations and bag-level expert class prompts embeddings for final prediction.

We validated the effectiveness of ConcepPath on five tasks (Fig. 1c): (1) lung cancer subtyping, (2) breast cancer HER2 scoring, and (3) gastric cancer immunotherapy-sensitive subtyping (including 3 binary classification tasks). ConcepPath outperformed previous state-of-the-art methods on all tasks in Fig. 2a, which shows the prominence of utilizing human expert knowledge effectively. Particularly noteworthy is the nearly 7% improvement that ConcepPath achieved in classifying Epstein–Barr virus (EBV)-positive for gastric cancer cases, demonstrating its potential as an economical and less time-consuming alternative for stratifying patients who respond to immune checkpoint inhibitor therapy.

Results

Dataset characteristics for tumor diagnosis

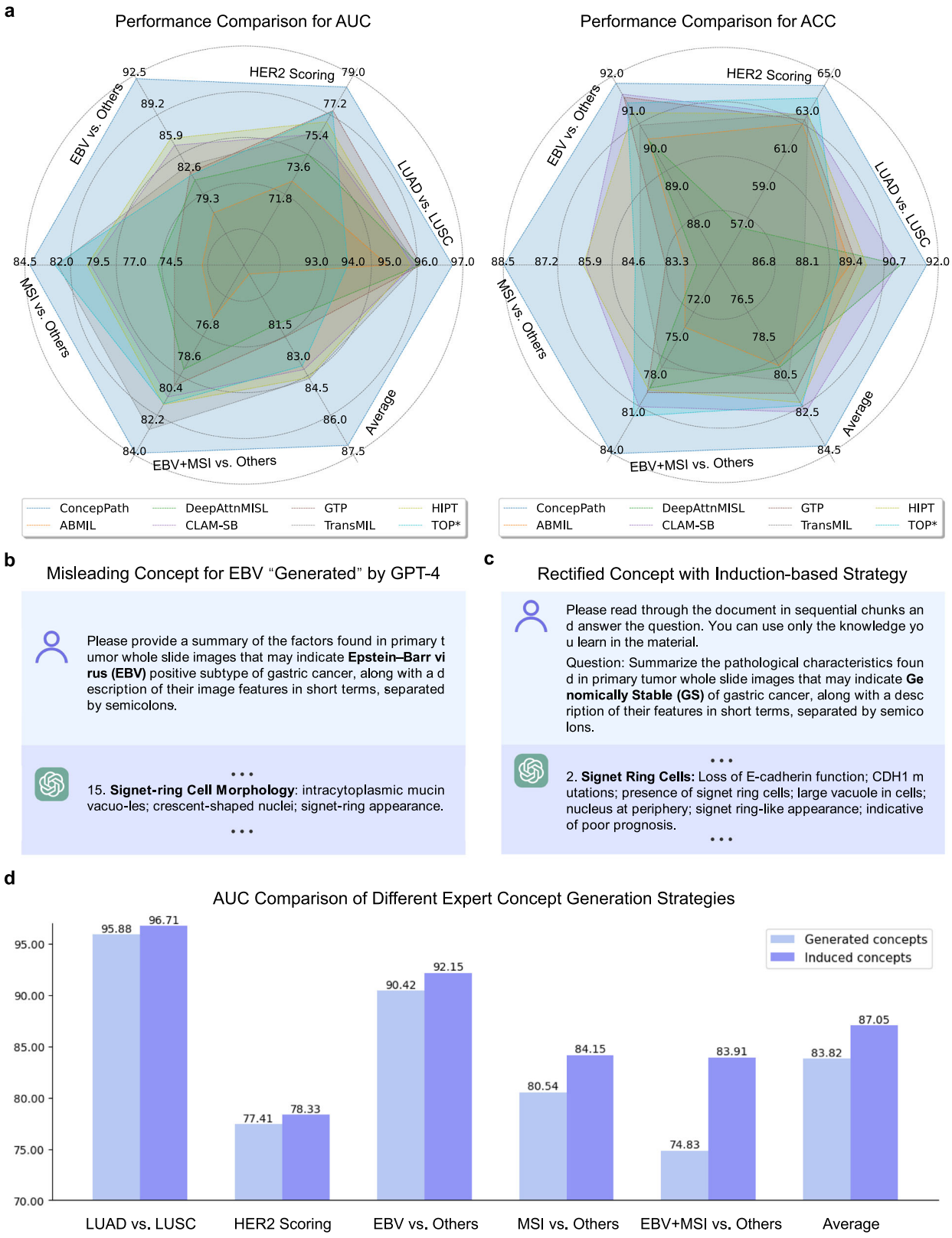
We evaluate ConcepPath on three public datasets (NSCLC, STAD and BRCA) from The Cancer Genome Atlas (TCGA) repository. The first dataset is **NSCLC**, the lung cancer project containing 1042 cases. For the tumor subtyping task on this dataset, there are 530 cases diagnosed as lung adenocarcinoma (LUAD) and 512 cases diagnosed as lung squamous cell carcinoma (LUSC). The second dataset is **BRCA**, the breast cancer project containing 933 cases. For the HER2 scoring task on this dataset, there are 164 cases diagnosed as positive, 186 cases diagnosed as equivocal and 583 cases diagnosed as negative. Human epidermal growth factor receptor 2 (HER2) plays a crucial role as both a prognostic and predictive marker, being over-expressed in approximately 15–20% of breast cancer cases. Assessing HER2 status is vital for guiding clinical treatment choices and prognostic assessments. The evaluation of HER2 status is performed through transcriptomics or immunohistochemistry (IHC) methods, including in-situ hybridization (ISH), which adds extra costs and tissue demands. Furthermore, this process is subject to variability in analysis, especially due to potential biases in manual scoring observations[23,24]. The last dataset is **STAD**, the gastric cancer project containing 268 cases. For the immunotherapy-sensitive subtyping task on this dataset, there are 26 cases diagnosed as Epstein–Barr virus (EBV)-positive, 44 cases diagnosed as Microsatellite Instability (MSI), and 199 cases diagnosed as Genomically Stable (GS) and Chromosomal Instable (CIN). EBV and MSI tumors have been reported to be highly responsive to Immune checkpoint inhibitor (ICI) therapy, which is widely used but effective only in a subset of gastric cancers[25]. However, the high costs of required diagnostic methods like immunohistochemistry and polymerase chain reaction limit the practical application of this molecular classification in treatment decisions[26]. EBV-associated and MSI gastric cancers are characterized by distinct histological traits. EBV-positive tumors often display significant lymphocyte infiltration within both the neoplastic epithelium and the stroma, frequently referred to as lymphoepithelioma-like carcinoma or gastric carcinoma with lymphoid stroma[27]. Similarly, the MSI subtype typically shows extensive lymphocytic infiltration, predominantly featuring intestinal-type histology and expansive growth patterns[28,29]. Note that in this study, we follow the setting in previous study[26] to formulate three binary classification tasks: EBV vs. Others, MSI vs. Others, and EBV +MSI vs. Others. Given the biases inherent in manual evaluations and the additional expenses associated with these assessments, particularly for the latter tasks corresponding to guiding clinical treatment choices, conducting precise analysis directly from routine H&E-stained tissue sections through deep learning techniques is of significant clinical and scientific interest. On these tasks, We utilized patient-level five-fold cross-validation for all experiments evaluating ConcepPath and other methods and reported the results of all models in the form of mean value. For evaluation metrics, the area under the curve (AUC) of receiver operating characteristic, and the accuracy (ACC) were adopted. ConcepPath could utilize different CLIP-based pathology vision-language foundation models as its basic component. As QuiltNet[17] and CONCH[14] lead to better performance in our experiments, in the following sections, we report the results of using QuiltNet as the default basic component of ConcepPath and the feature extractor of other baselines in the following sections and report the results of using CONCH in the Supplementary Materials.

ConcepPath helps in treatment decision

We evaluated ConcepPath against seven state-of-the-art (SOTA) methodologies: (1) ABMIL[8], (2) DeepAttnMISL[30], (3) CLAM-SB[31], (4) GTP[32], (5) TransMIL[33], (6) HIPT[6], and (7) TOP[12]. For TOP[12], due to the current incomplete official implementation released by the authors, we run both the author's current implementation and our re-implementation according to their paper and select the higher performance as the final result of TOP. As illustrated in Fig. 2a, ConcepPath surpasses the aforementioned methods in both AUC and ACC across all assessed tasks. Specifically, a significant performance leap for ConcepPath is observed in the BRCA and STAD datasets. For example, on the BRCA dataset, ConcepPath registers a 5.66% increase in the F1-score over the leading baseline. Similarly, on the STAD dataset, improvements of 6.23% in EBV vs. Others AUC, 1.71% in MSI vs. Others AUC, and 1.35% in EBV+MSI vs. Others AUC were noted in comparison to the best-performing baseline. These datasets involved complex histological analysis tasks such as HER2 scoring and immunotherapy-sensitive subtyping, necessitating the recognition of intricate histological features and molecular tissue characteristics. This leads us to hypothesize that for more intricate WSI analysis challenges, the fusion of prior domain knowledge with the discovery of new, diagnosis-related concepts is crucial, significantly more so than for simpler tumor/normal classification or tumor subtyping tasks. Overall, ConcepPath enhances WSI analysis capabilities, especially in tumor subtyping and immune response assessment, potentially aiding in treatment decision-making processes. Regarding TOP[12], designed to leverage language priors in few-shot weakly supervised learning for WSI analysis, it demonstrated unsatisfactory results in full training settings. This could be due to the unreliable generation of language priors, a misalignment between histopathology images and prior knowledge text, and a lack of new knowledge acquisition from the training data, which is detailed in Section Baseline Models. In addition, we report the experimental results when using CONCH as the basic component of ConcepPath and feature extractor of other baselines in Supplementary Fig. 5. Although HIPT achieved improved performance as CONCH was trained on a larger number of histology images, we noticed that ConcepPath still outperforms other baselines among most tasks and shows as a more robust method compared with others.

Comparison of different expert concept extraction strategies

The incorporation of human expert knowledge is of great significance to ConcepPath, the performance will be influenced if such prior is not imported accurately. We investigate the impact of different expert concept generation strategies and show the mean results in Fig. 2d. The "Generated



concepts" refers to the strategy used in previous works[12,34,35], which involves directly querying GPT-4 for relevant concepts without providing any expert materials. In contrast, "Induced concepts" represent our proposed strategy, which entails asking GPT-4 to induce relevant concepts from medical literature related to the target diagnostic task. As shown in Fig. 2d and Supplementary Fig. 4a, the "Induced concepts" achieve better performance across all metrics, particularly for the more challenging immunotherapy-sensitive subtyping tasks on the STAD dataset. This highlights the importance of inducing concepts from professional materials for complex WSI analysis tasks. In addition, we report the experimental results when using CONCH as the basic component of ConcepPath. As shown in Supplementary Fig. 6, we could obtain the same conclusion as above. In Fig. 2b, c,

Fig. 2 | Performance and expert concept generation comparison of ConcepPath. **a** Radar charts depicting the average AUC(Left) and ACC(Right) for the five WSI analysis tasks in the five-fold cross-validation experiment conducted on NSCLC, BRCA, and STAD datasets. "Average" denotes the average performance among all five tasks. "TOP*" represents the higher performance in the author's implementation and our implementation of TOP. ConcepPath demonstrated more accurate predictions on all five tasks since it successfully incorporates human expert prior knowledge and data-driven knowledge learned from the training data. **b**, **c** A misleading concept generated by directly querying GPT-4 (denoted as "Generated"), which our induction-based method (denoted as "Induced") successfully rectifies for the gastric immunotherapy-sensitive subtyping task. Specifically, the concept "signet-ring cells" was found in the category of Epstein–Barr virus (EBV) positive subtype in the generated concepts; however, this morphology is more commonly linked to the Genomically Stable (GS) subtype, where mutations in CDH1 and RHO genes play a pivotal role. **d** A histogram representing the AUC comparison of different expert concept generation strategies. The y-axis is the AUC(%) and the x-axis is the WSI analysis tasks and their average performance. "Induced" concepts demonstrated better performance among all tasks, especially for the more challenging immunotherapy-sensitive subtyping tasks on the STAD dataset, highlighting the importance of inducing concepts from professional materials for complex WSI analysis tasks.

we also provide one example of a misleading concept generated by directly querying GPT-4, which our induction-based strategy successfully rectifies. For the gastric immunotherapy-sensitive subtyping task, the expert concept "signet-ring cells" is more commonly linked to the Genomically Stable (GS) subtype instead of the Epstein–Barr virus (EBV) positive subtype[36]. However, as as illustrated in Fig. 2b the "Generated concepts" strategy attributes the expert concept "signet-ring cells" to the EBV-positive subtype instead of the GS subtype, which may cause confusion for the model. In contrast, as demonstrated in Fig. 2c, under the guidance of relevant medical literature, our proposed "Induced concepts" strategy accurately attributes expert concept "signet-ring cells" to the GS subtype, which provides reliable prior expert knowledge to our framework.

Effectiveness of data-driven concept learning

The integration of learnable concepts is also a key component of ConcepPath. We investigate its impact on NSCLC lung cancer subtyping, BRCA HER2 scoring, and STAD EBV vs. other classification tasks. Results are shown in Fig. 3a. The x-axis in Fig. 3a represents the number of learned concepts used for each class, where 0 refers to the only use of GPT-4 induced expert concepts in our framework. From the results, we observe a performance increase of **1.04%** in AUC for NSCLC, **1.16%** in AUC for BRCA and **3.96%** in AUC for EBV vs. Others upon integrating new concept discovery, demonstrating the effectiveness of complementing human expert knowledge with learned knowledge. Furthermore, we noticed that for more challenging and inadequately researched diagnostic tasks, a greater number of learned concepts are required. For instance, NSCLC achieved its best performance with 4 learned concepts per class, while the EBV vs. Others comparison reached its peak performance with 8 learned concepts per class. Furthermore, we identified a performance decline when the number of learned concepts was large (i.e., 12) on both datasets. This observation suggests a potential trade-off between prior expert knowledge and learned knowledge. If the number of learned data-driven concepts is excessive, the impact of prior expert knowledge may be diminished, potentially resulting in overfitting the training data. We noticed that such a trade-off still exists when using CONCH as the basic component of ConcepPath, as shown in Supplementary Fig. 7. Particularly, we also investigated the independent contributions of the expert concept and the data-driven concept in ConcepPath (Supplementary Fig. 11). We noticed that while removing the data-driven concepts will cause a bigger average performance drop, either ignoring the human expert prior or the knowledge in the training data would generally cause an obvious performance drop for most tasks on both QuiltNet-based and CONCH-based ConcepPath. These results illustrate that the success of ConcepPath lies in incorporating the complementary human expert prior and data-driven knowledge in automated WSI analysis.

Effectiveness of bag-level concept guidance

We also investigate the impact of the bag concept-guided aggregation in ConcepPath. As shown in Fig. 3b and Supplementary Fig. 4b, "w/o Bag-level guidance" bars denote directly averaging the concept-specific bag-level features into the overall bag representation without considering the correlations between the instance-level concepts and the bag-level class prompts. The model performed better with Bag-level guidance, for instance, we observed a performance drop of **2.82%** in AUC for EBV+MSI vs. Others and **4.11%** in AUC for MSI vs. Others when ignoring the relationship between the instance-level concepts and the bag-level class prompts. This demonstrates the importance of the second-stage bag-level concept-guided aggregation in our framework. The experimental results when using CONCH as the basic component of ConcepPath, the overall performance among five tasks gain notable improvement with this second-stage bag-level concept-guided aggregation, especially for the more reflective metric AUC (Supplementary Fig. 8).

Effectiveness of slide adapters

The slide adapters are proposed to learn new features and blend them with the original features of the overall bag representation and bag-level concept embedding. We also explore their effectiveness in Fig. 3b and Supplementary Fig. 4b. Particularly, we notice an obvious performance drop in AUC for EBV vs. Others, which may indicate the potential limits of the knowledge within the existing human medical research papers and the pathology vision-language model with respect to this challenging and inadequately researched diagnostic task. The experimental results when using CONCH as the basic component of ConcepPath, the overall performance among five tasks would be improved by the slide adapters, especially for the more reflective metric AUC (Supplementary Fig. 8).

Comparison of different vision-language models

The alignment of histopathology images with textual expert concepts is of great significance in our framework. We also compare the efficiency of using different CLIP-based vision-language models as the basic building component in ConcepPath to align the concepts with histopathology images. Results are shown in Fig. 3c and Supplementary Fig. 5c, where CLIP[15] is trained on a variety of image-text pairs from the internet, PLIP[16] and PathClip[37] are trained on over 200K histopathology image-text pairs, and QuiltNet[17] and CONCH[14] are trained on over 1 million histopathology image-text pairs. Comparing the performances between CLIP and PLIP, an obvious performance increase can be observed by using pathology vision-language models. Moreover, the additional performance increase brought by QuiltNet[17] and CONCH[14] further demonstrates that our framework could benefit from more accurate alignment, and thus more efficiently incorporating concept knowledge into histopathology images.

Visualization and post-hoc interpretation

Model interpretation is crucial for medical applications. ConcepPath offers post-hoc interpretation by visualizing the similarity scores between instance-level features and instance-level concepts as similarity maps on the slide, providing multi-dimensional reference information compared to previous attention map-based approaches. Some visualization examples are presented in Fig. 4, which displays four distinct expert instance-level concept similarity maps for four accurately classified lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) slides. Additionally, we include the attention maps of CLAM[31] for a comprehensive comparison. Evaluated by our expert pathologist collaborator, we note that the clinical relevance of this work lies in enabling pathologists to understand the rationale behind the model's predictions for a given WSI. The heatmaps for various expert concepts pinpoint the exact regions within WSIs that contribute to specific predictions, providing a clear and interpretable basis for the model's

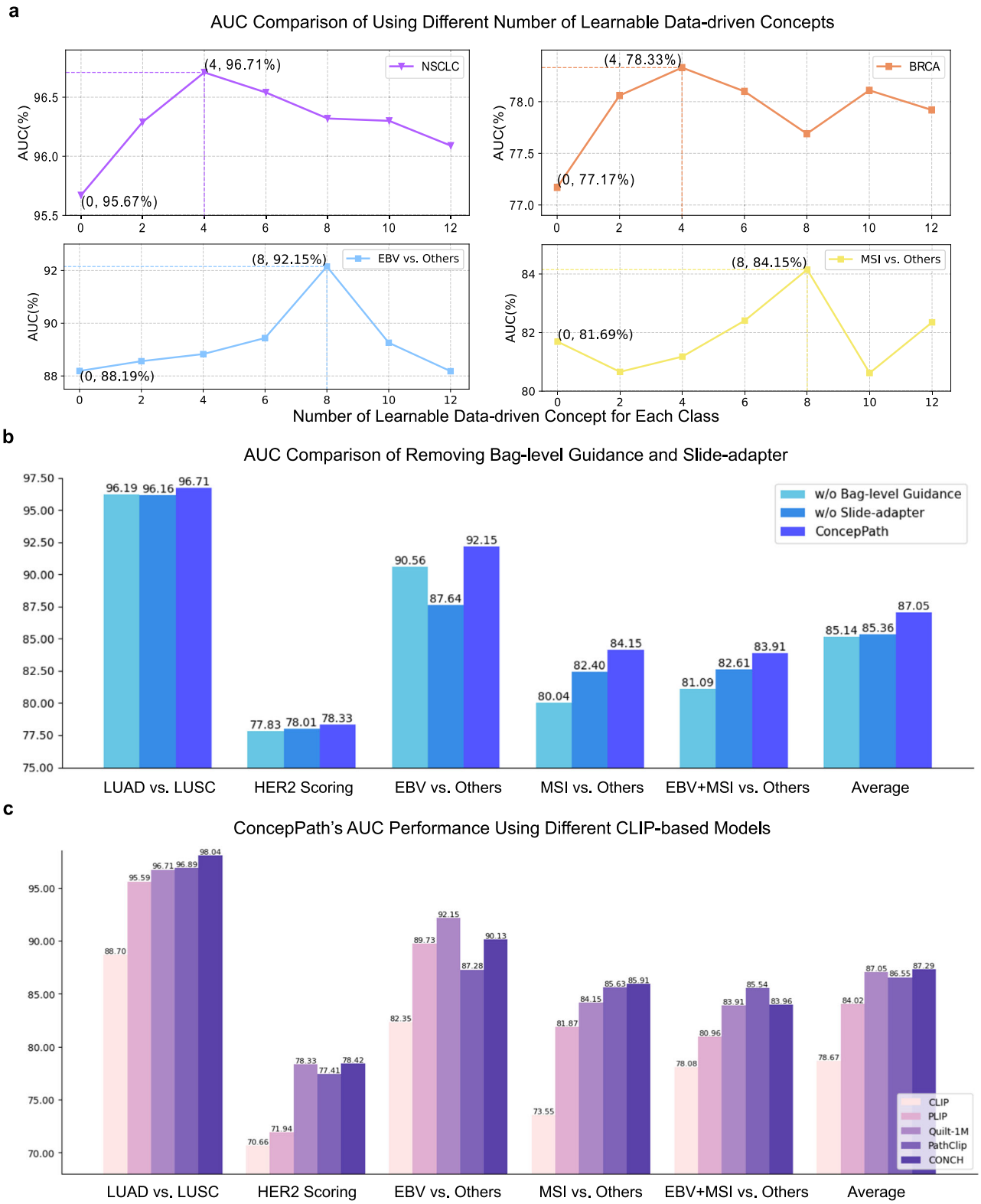


Fig. 3 | Investigation of proposed components in ConcepPath. a Line plots for investigating new data-driven concept learning, the y-axis is the AUC(%) and the x-axis is the number of learned concepts used for each class, where 0 means only using human expert concepts. Integrating data-driven knowledge learned from training data improves overall performance, and for more challenging and inadequately researched diagnostic tasks, a greater number of learned concepts are required. The performance decline when the number of learned concepts was large suggests a potential trade-off between prior expert knowledge and learned knowledge. **b** A histogram representing investigations on second-stage bag-level concept-guided aggregation and slide adapters, the y-axis is the AUC(%), and "w/o Bag-level guidance" refers to using average pooling aggregation. Both modules contributed to the improved performance. **c** A histogram representing the comparison of using different CLIP-based vision-language models as ConcepPath's basic component for aligning histopathology images and concept knowledge, and the y-axis is the AUC(%). Obvious performance increase can be observed by using pathology vision-language models, and ConcepPath could benefit from more accurate alignment if the pathology vision-language were trained on larger datasets.

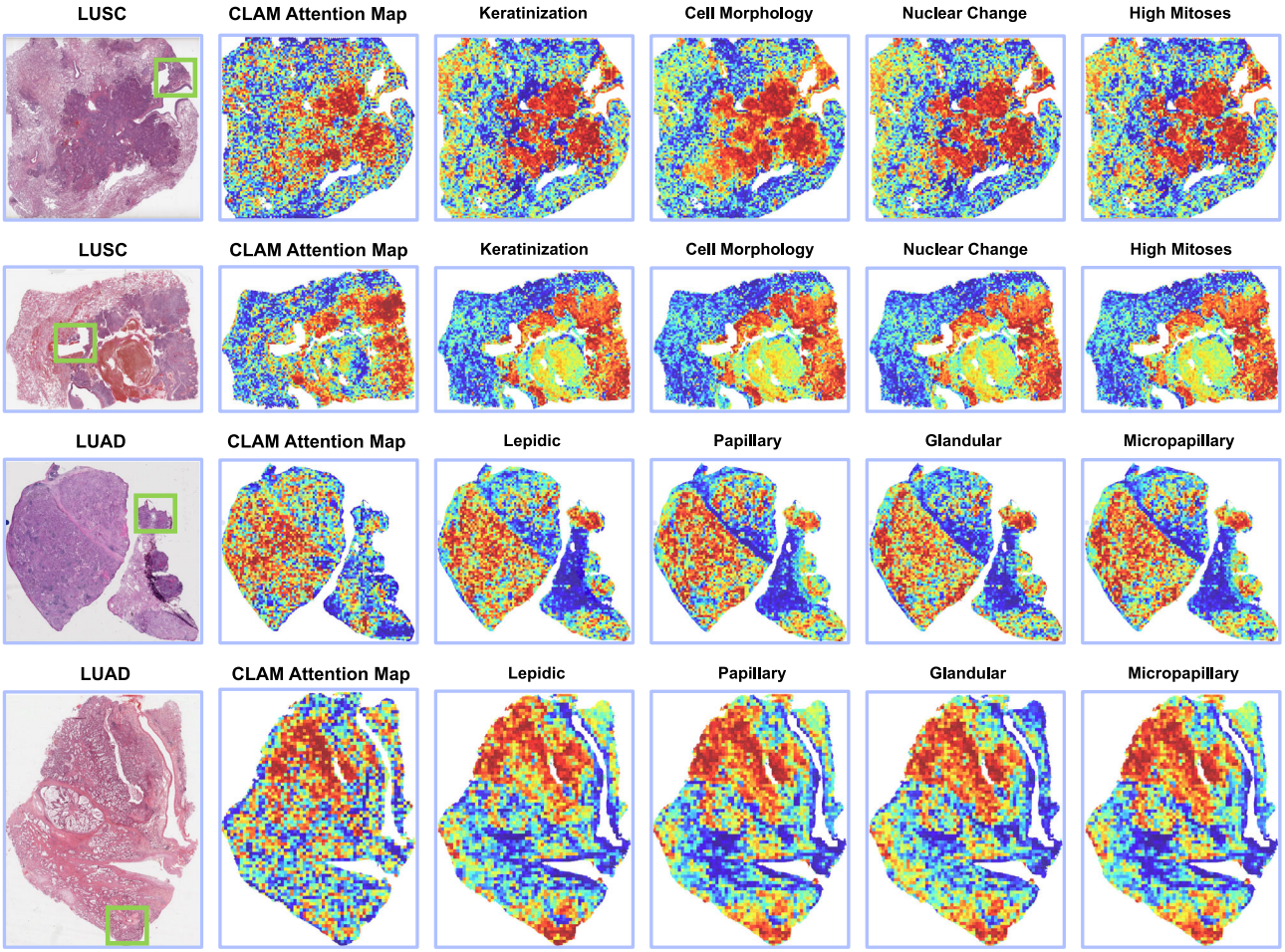


Fig. 4 | Visualizations of ConcepPath and baseline method. Instance-level expert concept similarity maps. The slides are accurately identified as the lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) subtype, respectively. In comparison to the CLAM attention map, the similarity maps of various instance-level concepts provide a more precise focus on the tumor in the green box highlighted area.

decisions. This is well-illustrated in Fig. 4 and Supplementary Fig. 12. For instance, our pathologist collaborator identified that concepts such as "lepidic", "papillary", "glandular", "micropapillary", and "solid growth" are closely associated with LUAD, while "keratinization", "cell morphology", "nuclear changes", and "high mitoses" are characteristic of LUSC. These observations align with established medical knowledge, demonstrating the interpretability and clinical validity of the model. Notably, these expert concepts exhibit high activity in tumor regions, consistent with domain expertize. Furthermore, similarity maps generated by ConcepPath provide a more precise focus on tumor regions compared to CLAM attention maps. For example, in the slide from the first row, the green box area includes additional tumor foci that are overlooked by CLAM. Similarly, benign mucous glands intermixed with inflammatory cells activate medical concepts such as inflammatory cells, associated lymphocytes, and immune cell clusters, reflecting plausible biological phenomena (Supplementary Fig. 12). Lastly, we observed minor variations in WSI regions across different expert concepts. These variations could serve as valuable supplementary references, offering multi-dimensional insights for pathologists during diagnosis.

Discussion

Despite the rapid advancement in computational pathology, the integration of valuable human expert knowledge into automated AI-assisted diagnosis remains a significant yet unresolved challenge. The advent of CLIP-based pathology vision-language foundation models and large language models (LLMs) presents a promising avenue by aligning histopathology images with human linguistic priors for more efficient WSI analysis. However, direct queries to LLMs may yield inaccurate statements that not grounded in scientific fact. Furthermore, the complexity of diseases may surpass the existing human expert knowledge, necessitating the discovery of complementary knowledge hidden within training data.

To address these issues, we introduce ConcepPath in this study, a novel framework that explicitly incorporates reliable prior expert knowledge and learns complementary concepts from training data for precise WSI analysis. To circumvent inaccuracies inherent in LLMs, ConcepPath employs GPT-4 as a reasoning engine to derive reliable instance-level expert concepts and bag-level expert class prompts from medical literature related to the target diagnostic task. Additionally, ConcepPath explores complementary data-driven instance-level concepts from the training data using a set of learnable prompt representations. Following the Multiple Instance Learning (MIL) paradigm, ConcepPath utilized a concept knowledge-guided two-stage hierarchical feature aggregation process for efficient bag-level WSI representation. Importantly, ConcepPath integrates slide adapters prior to the final prediction to address the domain shift between the pathology vision-language model's training data and downstream WSI analysis tasks.

To avoid confusion and distinguish ConcepPath from the CLIP-based pathology foundation models, we would like to emphasize that the objectives of the CLIP-based pathology foundation models and ConcepPath are in fact quite different: the CLIP-based pathology foundation models aim to align path-level histology images with their corresponding text description, while ConcepPath aims to provide precise and interpretable slide-level analysis via leveraging human expert prior knowledge and complementary knowledge extracted from training data. The CLIP-based pathology foundation models

mainly serve as a basic building component in ConcepPath in this study. We provide a more detailed comparison and illustration of ConcepPath's advantage in Section Differences from CLIP-based Models and Discussion.

The proposed ConcepPath pioneers the incorporation of human expert knowledge for efficient and accurate histopathology analysis, a topic of significant clinical importance yet largely unexplored in existing research. The comprehensive experimental results in this study affirm the superiority of ConcepPath over state-of-the-art methods and demonstrate the efficacy of the proposed components, offering new perspectives on leveraging human expert knowledge, LLMs, CLIP-based pathology vision-language foundation models, and training data for precise automated WSI analysis.

However, ConcepPath has limitations. Although it surpasses other state-of-the-art methods, it relies on additional supervised information, namely, expert knowledge induced by GPT-4, which may be insufficient or biased for rare or novel cancer types. Furthermore, the potential for divergent expert opinions on controversial topics underscores the challenge of acquiring accurate expert information - an issue we aim to address in future work. Additionally, while ConcepPath proposes learning data-driven concepts from training data, interpreting these concepts to enhance model understanding and facilitate medical discovery remains crucial. The current approach employs concept similarity maps for interpretation, but further collaboration with expert pathologists is necessary to decode the semantic information underlying different learned concepts. On the other hand, due to computational constraints, the text and image encoders remain frozen during training despite the domain shift between encoder training data and downstream task data. Exploring resource-efficient strategies for fine-tuning both encoders on downstream tasks to mitigate this domain shift will also facilitate further research. Lastly, as PathChat[38], a multimodal generative vision-language AI assistant for human pathology, emerged as a concurrent work of ConcepPath, it would be promising if we could adopt PatChat as a basic building component in ConcepPath. Specifically, PathChat could generate a response to both histology image and text input as a chatbot like GPT4V, which could provide a more flexible linkage between the histology image and human language. However, as the PathChat model is not publicly available yet, we leave integrating PathChat in ConcepPath to explore whether it can improve ConcepPath's capability as our future work.

For other future works, we will aim to refine the ConcepPath framework, specifically addressing the domain shift between the encoders' training data and the downstream task data. This enhancement will focus on adapting the model to better generalize across different datasets, thereby improving its robustness and accuracy in diverse clinical scenarios. Furthermore, we plan to quantify and mitigate the challenges associated with obtaining precise expert information tailored to specific WSI analysis tasks, enhancing the reliability of the expert knowledge integrated into the model. Additionally, we will explore the integration of graph representations into the ConcepPath framework to enrich the analysis of WSIs. By capturing the intricate spatial relationships and contextual details inherent in tissue structures, graph representations can provide a more nuanced understanding of the histopathological features. This advancement is anticipated to offer deeper insights into the tissue architecture, potentially unveiling new biomarkers and improving diagnostic accuracy.

Methods

ConcepPath overview

Figure 1 b, e presents our proposed ConcepPath framework. Specifically, as depicted in Fig. 1b, ConcepPath first utilizes the large language model, like GPT-4, to induce reliable disease-specific instance-level expert concept and bag-level expert class prompts from medical literature. On the other hand, to complement the extracted expert knowledge, ConcepPath employs a set of purely learnable instance-level concepts for complementary data-driven instance-level concepts learned from the training data. Next, ConcepPath aligns the histopathology patches and the concepts by leveraging CLIP-based pathology vision-language foundation models. Subsequently, the instance features are aggregated into the overall bag representation using a two-stage hierarchical aggregation paradigm, guided by the instance-level concept and the correlations between instance-level expert concepts and bag-level expert class prompts. Afterward, ConcepPath feeds the overall bag representation and bag-level expert class prompt embeddings to slide adapters, which serve as an additional bottleneck layer to perform residual-style feature blending with the original features. Finally, the predictions are calculated based on the similarities between the adapted bag representations and bag-level expert class prompt embeddings. For simplicity, we elaborate ConcepPath with a binary classification WSI analysis task in the formulas in the following sections, identifying class *A* and class *B*. Note that ConcepPath can also be extended to a multi-class classification setup and we conducted a 3-class classification task on the BRCA dataset.

Inducing expert concepts with LLM

To reduce the task difficulty and fully exploit the power of the CLIP-based pathology foundation models and human expert prior knowledge, ConcepPath decomposes a complex WSI analysis task into several patch-level subtasks - scoring related medical expert concepts. In this section, we detail how to induce patch-level expert concepts from human priors using a large model. As shown in Supplementary Fig. 2, in ConcepPath, given a specific WSI analysis task, medical literature was first collected using two search engines: Google and New Bing, with Google serving as the most powerful traditional search engine and New Bing serving as the recent search engine powered by large language models (LLMs). Specifically, "<key words>", "<key words>, journal" and "<key words>, paper" will be searched in both engines and also Google Scholar, for instance, the "<key words>" in lung cancer subtyping task would be "lung adenocarcinoma" and "lung squamous cell carcinoma". Then, we keep and consolidate the results of medical papers published in well-known journals such as Nature Series journals to ensure their reliability. Once collected, each medical literature is fed into large language models, we utilize GPT-4 in our study, to induce instance-level expert concepts.

To ensure the quality of the induced instance-level expert concepts, we adopt a three-step strategy: First, we ask GPT-4 to summarize the pathological factors related to the classes of the target task from the input literature, together with their descriptions. Such instance-level concepts typically correspond to potential phenotypes or clinical, pathological, and molecular characteristics that may appear in histopathology images. Then, GPT-4 is prompted to rank the summarized factors to facilitate the final manual examination of all expert concepts conducted by the users or pathologists. Finally, we require GPT-4 to re-write the descriptions of the ranked pathological factors to include more visual descriptions of the tissue for fully exploiting the power of the CLIP-based pathology vision-language foundation models. We provide a complete example of processing one medical literature related to the lung cancer subtyping task in Supplementary Fig. 10. We feed one paper each time to GPT-4 to ensure an easier and more reliable summarization process. With all collected medical literature being processed, we merged the summarized pathological factors from each paper, and manually deleted the repeated ones. The factors could not be observed from histology images to form the final instance-level expert concepts groups for each target class. Notably, as shown in Supplementary Fig. 2, the generated concepts are traceable to the users as the source literature is specified in their generation process, which further ensures ConcepPath could benefit from the human expert prior knowledge in a reliable manner.

When fed to ConcepPath, each instance-level expert concept is composed of two parts: The first part is the above-induced text description and the second part is a learnable vector, which follows the idea of the learnable prompt representation proposed in CoOp[39] to improve the overall performance of the CLIP-based models (Fig. 1e).

For the bag-level expert class prompts, ConcepPath requires GPT-4 to induce comprehensive descriptions of different target classes concerning the instance-level expert concepts induced in the previous step among each literature, then we prompt GPT-4 to merge the descriptions from all collected literature. Each bag-level expert class prompt also has two parts, similar to the instance-level expert concept (Fig. 1e). All detailed automatically collected medical literature and induced instance-level expert

concepts and bag-level expert class prompts can be found in our released code repository.

Learning complementary data-driven concepts

Given a WSI $\boldsymbol{X}$ under 20× magnification, we first apply the sliding window strategy to crop $\boldsymbol{X}$ into numerous non-overlapping image patches. Then, ConcepPath extracts the instance feature from each image patch with the image encoder from the CLIP-based pathology vision-language foundation models. In this study, unless otherwise specified, we utilized QuiltNet[17] as the default utilized CLIP-based pathology vision-language foundation model. The instance feature extraction is defined as:

$$\boldsymbol{Z} = f_{image}(\boldsymbol{X}), \tag{1}$$

where f_{image} is the image encoder, and $\boldsymbol{X}$ contains n cropped patches. $\boldsymbol{Z} \in \boldsymbol{R}^{n \times d}$ is the extracted instance features, where d represents the dimension of the features. Then, we use the text encoder from the CLIP-based pathology vision-language foundation models to obtain instance-level concept representations for each target class:

$$\boldsymbol{C}_{ins}^{A} = f_{text}(\boldsymbol{T}_{ins}^{A}), \tag{2}$$

where f_{text} is the text encoder, $\boldsymbol{T}_{ins}^{A}$ contains m instance-level concepts for target class A, and $\boldsymbol{C}_{ins}^{A} \in \boldsymbol{R}^{m \times d}$ is the instance-level concept embeddings for class A. Specifically, $\boldsymbol{T}_{ins}^{A}$ is composed of two groups:

$$\boldsymbol{T}_{ins}^{A} = \left\{ \boldsymbol{I}_{ins}^{A}, \boldsymbol{D}_{ins}^{A} \right\}, \tag{3}$$

where $\boldsymbol{I}_{ins}^{A}$ is the induced instance-level expert concepts for class A in the previous paragraph. $\boldsymbol{D}_{ins}^{A}$ is a group of learnable data-driven instance-level concepts for class A, containing a set of purely learnable prompt representations optimized with the training data during the training process. The data-driven instance-level concepts $\boldsymbol{D}_{ins}^{A}$ serve as complementary diagnostic factors to the extracted expert domain concepts $\boldsymbol{I}_{ins}^{A}$, helping to describe the whole picture of a disease. In addition, to ensure that the learned instance-level data-driven concepts extract complementary information to the instance-level domain concepts, we define a mutual distinctive loss among them as:

$$Loss_{mutual} = \sum_{\substack{i,j \in \{1,\ldots,m\} \\ i \neq j,\ cls \in \{A,B\}}} \cos\left(C_{ins,i}^{cls} \cdot C_{ins,j}^{cls} \right), \tag{4}$$

where $C_{ins,i}^{cls}$ and $C_{ins,j}^{cls}$ are instance-level concept embeddings in the corresponding class.

To avoid confusion, we emphasize that the instance-level expert concepts and the data-driven instance-level concepts in ConcepPath are quite different, and summarize their difference into the generation difference and the objective difference for easier distinguishment. For the generation difference: The expert concepts are generated from the human prior knowledge, as shown in Supplementary Fig. 1a, they are induced by GPT-4 from the medical literature which is highly related to the target WSI analysis task and collected from the Internet. In contrast, as shown in Supplementary Fig. 1b, the data-driven concepts are extracted from the WSI training dataset by ConcepPath itself and are optimized during the training process using the gradient descent algorithm. For the objective difference: The objective of involving expert concepts is to utilize human expert prior knowledge to facilitate the overall performance of automatic WSI analysis. On the other hand, the objective of extracting data-driven concepts is to learn/extract useful knowledge for automatic diagnosis from the training data with neural networks. Therefore, they might contain complementary knowledge beyond human pathologists, which is probably complementary to the prior knowledge contained in the expert concept, and thus benefit the overall performance of ConcepPath.

Hierarchical two-stage concept-guided aggregation

We elaborately apply a hierarchical two-stage aggregation paradigm to obtain the overall bag representation under the guidance of instance-level concepts and correlations among the bag-level class prompts and instance-level concepts in ConcepPath.

In the first stage, ConcepPath aggregates the extracted instance-level features $\boldsymbol{Z}$ into concept-specific bag-level features for different target classes. For example, for class A, with guidance from the instance-level concept embeddings $\boldsymbol{C}_{ins}^{A}$, the aggregation process can be formulated as:

$$\boldsymbol{W}_{ins}^{A} = \mathrm{Softmax}\left(\boldsymbol{Z} \cdot \boldsymbol{C}_{ins}^{A\ T} \right), \tag{5}$$

$$\boldsymbol{H}^{A} = \boldsymbol{W}_{ins}^{A\ T} \cdot \boldsymbol{Z}, \tag{6}$$

where $\boldsymbol{W}_{ins}^{A} \in \boldsymbol{R}^{n \times m}$ is the similarity scores between different instances and instance-level expert concepts. With $\boldsymbol{W}_{ins}^{A}$ serves as the aggregation weights, and $\boldsymbol{H}^{A} \in \boldsymbol{R}^{m \times d}$ aggregated as concept-specific bag-level features for class A, we involve multiple medical concept scoring subtasks in ConcepPath to reduce the task complexity and fully exploit the power of human prior and CLIP-based pathology foundation models.

In the second stage, to obtain overall bag-level representation for class A, ConcepPath aggregates the concept-specific bag-level features $\boldsymbol{H}^{A}$ according to the correlations among the bag-level class prompts and the instance-level concepts of class A:

$$\boldsymbol{W}_{i2b}^{A} = \mathrm{Softmax}\left(\boldsymbol{C}_{ins}^{A} \cdot \boldsymbol{C}_{bag}^{A\ T} \right), \tag{7}$$

$$\boldsymbol{F}^{A} = \boldsymbol{W}_{i2b}^{A\ T} \cdot \boldsymbol{H}^{A} + \mathrm{mean}\left(\boldsymbol{H}^{A} \right), \tag{8}$$

where $\boldsymbol{W}_{i2b}^{A} \in \boldsymbol{R}^{m \times 1}$ is the similarity scores between bag-level expert class prompt and instance-level concepts for class A, and $\boldsymbol{F}^{A} \in \boldsymbol{R}^{1 \times d}$ is the overall bag-level representation of WSI X for class A.

Inspired by clip-adapter[40], we also implement slide adapters before aligning the overall bag-level representation $\boldsymbol{F}^{A}$ and the bag-level class prompt embedding $\boldsymbol{C}_{bag}^{A}$. Specifically, the slide adapters serve as additional bottleneck layers to learn new features and perform residual-style feature blending with the original features aggregated from the pre-trained encoders' feature space. In summary, the slide adapters can be written as follows:

$$SA_v(\boldsymbol{F}^{A}) = \mathrm{LeakyReLU}\left(\boldsymbol{F}^{A\ T} \cdot \boldsymbol{W}_1^{v} \right) \cdot \boldsymbol{W}_2^{v}, \tag{9}$$

$$SA_t(\boldsymbol{C}_{bag}^{A}) = \mathrm{LeakyReLU}\left(\boldsymbol{C}_{bag}^{A\ T} \cdot \boldsymbol{W}_1^{t} \right) \cdot \boldsymbol{W}_2^{t}, \tag{10}$$

$$\boldsymbol{F}^{A\star} = \alpha SA_v(\boldsymbol{F}^{A})^{T} + (1-\alpha)\boldsymbol{F}^{A}, \tag{11}$$

$$\boldsymbol{C}_{bag}^{A\ \star} = \beta SA_t(\boldsymbol{C}_{bag}^{A})^{T} + (1-\beta)\boldsymbol{C}_{bag}^{A}, \tag{12}$$

where both $SA_v(\cdot)$ and $SA_t(\cdot)$ represent two layers of learnable linear transformations W_1^v, W_2^v, W_1^t, and W_2^t that compose the slide adapters, with α and β as adjustable hyper-parameters. Similarly, we can obtain $\boldsymbol{F}^{B\star}$ and $\boldsymbol{C}_{bag}^{B\ \star}$ for class B and any other classes. Following the CLIP method[15], the prediction probability can be computed as:

$$p(y = A|\boldsymbol{X}) = \frac{\exp\left(\cos\left(\boldsymbol{F}^{A\star}, \boldsymbol{C}_{bag}^{A\ \star} \right)/\tau \right)}{\sum_{j=A} \exp\left(\cos\left(\boldsymbol{F}^{j\star}, \boldsymbol{C}_{bag}^{j\ \star} \right)/\tau \right)}. \tag{13}$$

Here, $\cos(\cdot,\cdot)$ denotes the cosine similarity, and τ represents the temperature of the Softmax function.

Post-hoc interpretation

Model interpretation and diagnosis are crucial for medical applications. ConcepPath offers post-hoc interpretation by visualizing the similarity scores between instance-level features and instance-level concepts as similarity maps. To generate the post-hoc interpretable similarity maps, we visualize the similarity scores of the above patch-level concept scoring subtasks conducted on different patches of the corresponding slide. Highlighted regions indicate higher responses of these patches to specific medical concepts, providing detailed multi-dimensional reference information on how ConcepPath evaluates various medical factors relevant to the WSI analysis task and reaches a final diagnosis.

Differences from CLIP-based models and discussion

To avoid confusion and distinguish ConcepPath from previous CLIP-based pathology foundation models, we would like to further discuss their differences. We would like to emphasize that the objectives of the CLIP-based pathology foundation models and ConcepPath are quite different: the CLIP-based pathology foundation models aim to align path-level histology images with their corresponding text description, while ConcepPath aims to provide precise and interpretable slide-level analysis via leveraging human expert prior knowledge and complementary knowledge extracted from training data as mentioned in the above sections. Therefore, instead of proposing a new CLIP-based pathology foundation model, ConcepPath investigates how to incorporate existing CLIP-based pathology foundation models as a basic building component to link human expert knowledge and histology images, to improve slide-level WSI analysis tasks.

As shown in the left part of Fig. 1d, the current CLIP-based pathology foundation models can conduct patch-level classification tasks by calculating the similarity between the input histology image patch and different class prompts. However, due to the large size of whole slide images (WSIs), it is infeasible to directly apply current CLIP-based pathology foundation models for slide-level classification, as the default input image size of current CLIP-based pathology foundation models is usually 224×224. Therefore, in previous CLIP-based pathology foundation model works (e.g., CONCH), a top-k pooling paradigm[14] is usually adopted when dealing with slide-level classification tasks. Specifically, as shown in the right part of Fig. 1d, a WSI is first tiled into numerous image patches, which could be fed into CLIP-based pathology foundation models to obtain patch-level predictions. Then, the slide-level prediction is calculated as the top-k pooling results of the patch-level predictions. ConcepPath has several advantages/novelties over this top-k pooling paradigm, and could significantly improve slide-level classification tasks using the CLIP-based pathology foundation models as a basic building component. Specifically, as mentioned in the above sections and shown in Fig. 1d, ConcepPath's advantages/novelties mainly lie in two aspects: First, ConcepPath decomposes complex WSI analysis tasks into subtasks and utilizes human expert prior knowledge for more accurate and interpretable prediction. Second, ConcepPath extracts complementary data-driven knowledge from the training data. In particular, the above top-k pooling with only class prompts on the patch level could be referred to as "one-stage aggregation", while ConcepPath with expert and data-driven concept prompts on the patch level and class prompts on the slide level could be referred to as "two-stage hierarchical aggregation".

Although we have different objectives from the CLIP-based pathology foundation models and their top-k pooling paradigm, to ensure a more comprehensive comparison, we conducted the following experiments in Supplementary Fig. 3 to illustrate the significant advantages of ConcepPath. On all the WSI analysis tasks, we observed that ConcepPath significantly improved the performance when incorporating different CLIP-based pathology foundation models, which validated the effectiveness of our model design.

Data preprocessing

Following the default settings outlined by CLAM[31], we initiate our pipeline by cropping the requisite image patches from each digitized slide. The process begins with the automated segmentation of tissue regions. Each WSI is loaded into memory at a downsampled resolution (32× downscale) and converted from the RGB to the HSV color space to facilitate segmentation. To identify tissue regions (foreground), we generate a binary mask by thresholding the saturation channel of the HSV image, subsequent to applying median blurring to smooth the image edges. This step is complemented by morphological closing operations aimed at filling in small gaps and holes within the tissue regions. The approximate contours of these foreground objects are then delineated, filtered based on a predefined area threshold, and earmarked for downstream processing. Post-segmentation, exhaustive cropping of 448×448 patches is performed within the segmented foreground contours at 20× magnification for each slide. This meticulous process ensures that the patches are representative of the histological features relevant for subsequent analyses.

Baseline models

Multiple Instance Learning (MIL)[41] has been extensively investigated for WSI analysis due to its weakly supervised learning paradigm. Generally, previous MIL methods can be divided into two groups: (1) instance-level methods[42–46], and (2) embedding-level methods[8,31,33,47,48]. Instance-level methods first obtain instance predictions and then aggregate them into bag predictions using either average pooling or maximum pooling. In contrast, embedding-level methods initially aggregate instance features into a high-level bag representation, followed by constructing a classifier based on this bag representation for bag-level prediction. However, most existing methods exclusively learn from image data, neglecting valuable prior expert knowledge that humans utilize and consider during the diagnostic process, such as pathological and molecular factors related to the disease. As embedding-level methods generally possess better performance, in this study, we include seven baseline models in our experimental performance comparisons, implementation details are discussed below:

- ABMIL: We followed the instructions on the ABMIL Github repository: https://github.com/AMLab-Amsterdam/AttentionDeepMIL. ABMIL addresses the MIL problem by learning the Bernoulli distribution of a bag. It introduces a neural network-based, permutation-invariant aggregation operator with an attention mechanism[8].
- DeepAttnMISL: We followed the guidelines on the DeepAttnMISL Github repository: https://github.com/uta-smile/DeepAttnMISL. DeepAttnMISL employs both siamese MI-FCN and attention-based MIL pooling. This method efficiently learns imaging features from WSIs and aggregates WSI-level information to patient-level data[30].
- CLAM-SB: We followed the instructions on the CLAM Github repository: https://github.com/mahmoodlab/CLAM. CLAM utilizes attention-based learning to identify diagnostically valuable sub-regions within a slide and refines the feature space through instance-level clustering over these identified regions[31].
- GTP: We followed the guidelines on the GTP Github repository: https://github.com/vkola-lab/tmi2022. GPT combines a graph-based representation of a WSI with a vision transformer to process pathology images[32].
- TransMIL: We followed the instructions on the TransMIL Github repository: https://github.com/szc19990412/TransMIL. TransMIL leverages a Transformer-based MIL approach (TransMIL) to analyze both morphological and spatial information in WSIs[33].
- HIPT: We followed the instructions on the HIPT Github repository: https://github.com/mahmoodlab/HIPT. HIPT capitalizes on the inherent hierarchical structure of WSIs, employing two levels of self-supervised learning to learn high-resolution image representations for precise WSI analysis[6].
- TOP: As the author's official implementation of TOP seems to be incomplete according to their paper's description, we also re-implemented it to ensure a comprehensive comparison. Unlike other baselines, TOP explored integrating language prior knowledge from large language models (LLMs) and vision-language models from the natural image domain to address few-shot weakly supervised learning for WSI analysis[12]. Nonetheless, their discussion primarily focuses on the low data regime and exhibits several limitations, including unreliable prior knowledge generation from LLMs, misalignment between

histopathology images and vision-language models from the natural image domain, and unsatisfactory performance in a full training setup. In contrast, ConcepPath is specifically designed to overcome these issues, setting it apart from TOP. Specifically, as shown in Supplementary Table 1, has advantages over the limitations of TOP in three aspects - summarizing traceable expert concepts from related medical literature, fine-tuning visual-textual feature spaces using slide adapters, and extracting complementary knowledge from training WSIs with learnable data-driven concepts. In addition, we have also compared the average AUC and ACC among all five tasks of ConcepPath and TOP's original implementation among different CLIP-based pathology models in Supplementary Fig. 10 alone. We observe that ConcepPath consistently outperforms TOP's original implementation, which validates the advantages of ConcepPath over TOP's limitations.

Training details

All experiments were conducted on a workstation equipped with eight NVIDIA RTX 3090 GPUs. Unless otherwise specified, we employed the image encoder ViT-B/32 and text encoder GPT/77 of QuiltNet[17] in ConcepPath as the feature extractors for both histopathology images and textual concepts in this study. The length of learnable parts was set to 16 tokens for both instance-level and bag-level expert concepts. For the number of instance-level concepts, we utilized 26 instance-level concepts for each target class, while the number of learned instance-level concepts was tuned from {2, 4, 6, 8, 10, 12} for each target class. During the training, we fixed all weights of the CLIP-based pathology foundation models and only learned the learnable part of the instance-level expert concepts bag-level expert class prompts, and the purely learnable data-driven concepts. Note that the above-learned parts are all input for the CLIP-based pathology foundation models. For model optimization, we employed the SGD optimizer and a batch size of 2. The learning rate was set to 0.0001 for all datasets. For evaluation metrics, the area under the curve (AUC) of receiver operating characteristic, and the accuracy (ACC) were adopted. We utilized patient-level five-fold cross-validation for all experiments and reported the results of all models in the form of mean and standard deviation.

Ethics approval and consent to participate

All datasets and other materials employed in this study are previously published and publicly accessible with approved protocol and participants's informed consent. No new human research participants are involved in this study.

Data availability

All datasets and other materials employed in this study have been previously published and publicly accessible. The TCGA datasets were acquired from the Genomic Data Commons Data Portal at https://portal.gdc.cancer.gov/. Source data and the medical literature, including the extracted human expert concepts for this study, are provided alongside this paper's released code repository (https://github.com/HKU-MedAI/ConcepPath).

Code availability

The ConcepPath source code is available on GitHub (https://github.com/HKU-MedAI/ConcepPath). We also uploaded all scripts and materials to reproduce all the analyses on the same website. A tutorial Colab notebook, including the trained weights of the studied clinical tasks, is also provided.

Received: 22 April 2024; Accepted: 22 December 2024;
Published online: 30 December 2024

References

1. Wang, X. et al. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Trans. Cybern.* **50**, 3950–3962 (2019).
2. Wang, S., Yang, D. M., Rong, R., Zhan, X. & Xiao, G. Pathology image analysis using segmentation deep learning algorithms. *Am. J. Pathol.* **189**, 1686–1698 (2019).
3. 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).
4. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).
5. Riasatian, A. et al. Fine-tuning and training of DenseNet for histopathology image representation using TCGA diagnostic slides. *Med. Image Anal.* **70**, 102032 (2021).
6. Chen, R. J. et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 16144–16155 (IEEE, 2022).
7. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
8. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *Proc. 35th International Conference on Machine Learning* 2127–2136 (PMLR, 2018).
9. Hou, W. et al. H2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proc. AAAI Conference on Artificial Intelligence* 933–941 (AAAI Press, 2022).
10. Guan, Y. et al. Node-aligned graph convolutional network for whole-slide image representation and classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 18813–18823 (IEEE, 2022).
11. Chan, T. H., Cendra, F. J., Ma, L., Yin, G. & Yu, L. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 15661–15670 (IEEE, 2023).
12. Qu, L., Luo, X., Fu, K., Wang, M. & Song, Z. The rise of ai language pathologists: exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. In *2023 Advances in Neural Information Processing Systems* 36 (Curran Associates, 2023).
13. Gamper, J. & Rajpoot, N. Multiple instance captioning: Learning representations from histopathology textbooks and articles. 16549–16559 (2021).
14. Lu, M. Y. et al. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 19764–19775 (IEEE, 2023).
15. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
16. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. & Zou, J. A visual-language foundation model for pathology image analysis using medical twitter. *Nat. Med.* **29**, 2307–2316 (2023).
17. Ikezogwo, W. O. et al. Quilt-1m: one million image-text pairs for histopathology. In *2024 Advances in Neural Information Processing Systems* 36 (Curran Associates, 2024).
18. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
19. Truhn, D., Reis-Filho, J. S. & Kather, J. N. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nat. Med.* **29**, 2983–2984 (2023).
20. Sanderson, K. Gpt-4 is here: what scientists think. *Nature* **615**, 773 (2023).
21. Echle, A. et al. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br. J. Cancer* **124**, 686–696 (2021).
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**, 21 (2023).

23. Wolff, A. et al. American society of clinical oncology; college of american pathologists. recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J. Clin. Oncol.* **31**, 3997–4013 (2013).
24. Lu, W. et al. Slidegraph+: whole slide image level graphs to predict her2 status in breast cancer. *Med. Image Anal.* **80**, 102486 (2022).
25. Kelly, R. J. Immunotherapy for esophageal and gastric cancer. *Am. Soc. Clin. Oncol. Educ. book* **37**, 292–300 (2017).
26. Hinata, M. & Ushiku, T. Detecting immunotherapy-sensitive subtype in gastric cancer using histologic image-based deep learning. *Sci. Rep.* **11**, 22636 (2021).
27. Fukayama, M. et al. Thirty years of Epstein-Barr virus-associated gastric carcinoma. *Virchows Arch.* **476**, 353–365 (2020).
28. Grogg, K. L., Lohse, C. M., Pankratz, V. S., Halling, K. C. & Smyrk, T. C. Lymphocyte-rich gastric cancer: associations with Epstein-Barr virus, microsatellite instability, histology, and survival. *Mod. Pathol.* **16**, 641–651 (2003).
29. Arai, T. et al. Frequent microsatellite instability in papillary and solid-type, poorly differentiated adenocarcinomas of the stomach. *Gastric Cancer* **16**, 505–512 (2013).
30. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N. J. & Huang, J. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Med. Image Anal.* **65**, 101789 (2020).
31. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
32. Zheng, Y., Gindra, R., Betke, M., Beane, J. E. & Kolachalama, V. B. A deep learning based graph-transformer for whole slide image classification. *IEEE Transactions on Medical Imaging* **41**, 3003–3015 (IEEE, 2022).
33. Shao, Z. et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Adv. Neural Inf. Process. Syst.* **34**, 2136–2147 (2021).
34. Yang, Y. et al. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 19187–19197 (IEEE, 2023).
35. Yan, A. et al. Learning concise and descriptive attributes for visual recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 3090–3100 (IEEE, 2023).
36. Ratti, M., Lampis, A. & Hahne, J. C. Microsatellite instability in gastric cancer: molecular bases, clinical perspectives, and new treatment approaches. *Cell. Mol. Life Sci.* https://doi.org/10.1007/s00018-018-2906-9 (2018).
37. Zheng, S. et al. Benchmarking pathclip for pathology image analysis. *Journal of Imaging Informatics in Medicine* **4**, 1–17, (Springer, 2024).
38. Lu, M. Y. et al. A multimodal generative ai copilot for human pathology. *Nature* **634**, 466–473 (2024).
39. Zhou, K., Yang, J., Loy, C. C. & Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **130**, 2337–2348 (2022).
40. Gao, P. et al. Clip-adapter: better vision-language models with feature adapters. *International Journal of Computer Vision* **132**, 581–595 (Springer, 2024).
41. Maron, O. & Lozano-Pérez, T. Attention is all you need. In *2017 Advances in Neural Information Processing Systems* 30570–576 (Curran Associates, 2017).
42. Hou, L. et al. Patch-based convolutional neural network for whole slide tissue image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2424–2433 (IEEE, 2016).
43. Feng, J. & Zhou, Z.-H. Deep MIML Network. In *2017 Proceedings of the AAAI conference on artificial intelligence* 1884–1890 (AAAI Press, 2017).
44. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
45. Xu, G. et al. Camel: a weakly supervised learning framework for histopathology image segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 10682–10691 (IEEE, 2019).
46. Kanavati, F. et al. Weakly-supervised learning for lung carcinoma classification using deep learning. *Sci. Rep.* **10**, 9297 (2020).
47. Zhu, W., Lou, Q., Vang, Y. S. & Xie, X. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017* 603–611 (Springer, 2017).
48. Li, B., Li, Y. & Eliceiri, K. W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *2021 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* 14318–14328 (IEEE, 2021).

Acknowledgements

This work was partially supported by the Research Grants Council of the Hong Kong SAR, China (Project No. 27206123 and T45-401/22-N), the Hong Kong Innovation and Technology Fund (Project No. ITS/274/22), the National Natural Science Foundation of China (No. 62201483), and Guangdong Natural Science Fund (No. 2024A1515011875).

Author contributions

L.Y. conceived and supervised the study. M.Y. supervised the study and provided expert opinions on pathology concepts and data interpretation for model development. Y.J. reviewed and provided expert opinions for this study. W.Z., Z.G., and Y.F. implemented the framework and performed all data analysis. W.Z. and Z.G. wrote the manuscript with inputs from all authors. All authors reviewed and approved the final paper.

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/s41746-024-01411-2.

Correspondence and requests for materials should be addressed to Maximus C. F. Yeung or Lequan Yu.