Medical Image Analysis 101 (2025) 103437

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Self-interactive learning: Fusion and evolution of multi-scale histomorphology features for molecular traits prediction in computational pathology

Yang Hu [a,c,*], Korsuk Sirinukunwattana [b,c], Bin Li [b,c], Kezia Gaitskell [d,g], Enric Domingo [f], Willem Bonnaffé [c,e], Marta Wojciechowska [a,c], Ruby Wood [b,c], Nasullah Khalid Alham [b,e], Stefano Malacrino [e], Dan J Woodcock [c,e], Clare Verrill [e,g,k], Ahmed Ahmed [h,i,k], Jens Rittscher [a,b,c,j,k,**]

[a] Nuffield Department of Medicine, University of Oxford, Oxford, UK
[b] Department of Engineering Science, University of Oxford, Oxford, UK
[c] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK
[d] Nuffield Division of Clinical Laboratory Sciences, Radcliffe Department of Medicine, University of Oxford, Oxford, UK
[e] Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK
[f] Department of Oncology, University of Oxford, Oxford, UK
[g] Department of Cellular Pathology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[h] MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK
[i] Nuffield Department of Womens and Reproductive Health, University of Oxford, Oxford, UK
[j] Ludwig Institute for Cancer Research, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK
[k] Oxford National Institute for Health Research (NIHR) Biomedical Research Centre, Oxford, UK

ARTICLE INFO

Dataset link: https://portal.gdc.cancer.gov/

MSC:
41A05
41A10
65D05
65D17

Keywords:
Computational Pathology
Whole slide images (WSI)
Cancer morpho-molecular subtyping
Self-interactive Deep Learning

ABSTRACT

Predicting disease-related molecular traits from histomorphology brings great opportunities for precision medicine. Despite the rich information present in histopathological images, extracting fine-grained molecular features from standard whole slide images (WSI) is non-trivial. The task is further complicated by the lack of annotations for subtyping and contextual histomorphological features that might span multiple scales. This work proposes a novel multiple-instance learning (MIL) framework capable of WSI-based cancer morpho-molecular subtyping by fusion of different-scale features. Our method, debuting as Inter-MIL, follows a weakly-supervised scheme. It enables the training of the patch-level encoder for WSI in a task-aware optimisation procedure, a step normally not modelled in most existing MIL-based WSI analysis frameworks. We demonstrate that optimising the patch-level encoder is crucial to achieving high-quality fine-grained and tissue-level subtyping results and offers a significant improvement over task-agnostic encoders. Our approach deploys a pseudo-label propagation strategy to update the patch encoder iteratively, allowing discriminative subtype features to be learned. This mechanism also empowers extracting fine-grained attention within image tiles (the small patches), a task largely ignored in most existing weakly supervised-based frameworks. With Inter-MIL, we carried out four challenging cancer molecular subtyping tasks in the context of ovarian, colorectal, lung, and breast cancer. Extensive evaluation results show that Inter-MIL is a robust framework for cancer morpho-molecular subtyping with superior performance compared to several recently proposed methods, in small dataset scenarios where the number of available training slides is less than 100. The iterative optimisation mechanism of Inter-MIL significantly improves the quality of the image features learned by the patch embedded and generally directs the attention map to areas that better align with experts' interpretation, leading to the identification of more reliable histopathology biomarkers. Moreover, an external validation cohort is used to verify the robustness of Inter-MIL on molecular trait prediction.

* Corresponding author at: Nuffield Department of Medicine, University of Oxford, Oxford, UK.
** Corresponding author at: Department of Engineering Science, University of Oxford, Oxford, UK.
E-mail addresses: yang.hu@ndm.ox.ac.uk (Y. Hu), jens.rittscher@eng.ox.ac.uk (J. Rittscher).

https://doi.org/10.1016/j.media.2024.103437
Received 5 April 2024; Received in revised form 6 October 2024; Accepted 9 December 2024
Available online 3 January 2025

1. Introduction

In up-to-date developments of computational pathology, with powerful deep neural networks (DNNs), whole slide image (WSI)-based morphological subtyping has become an emerging tool with great potential for future use in the clinic and in drug discovery. The core concept of morpho-molecular subtyping is to infer biologically relevant molecular traits directly from the morphological features presented in hematoxylin and eosin (H&E) histopathological samples, thus circumventing the need for expensive and time-consuming molecular assays (Sirinukunwattana et al., 2021). The successful deployment of such methods can have a profound impact on cancer treatment, offering a cost-effective solution for personalised medicine that leverages the richness of the information contained in WSIs. Such techniques promise to alleviate the dependency on time-consuming and potentially costly gene sequencing (Shendure et al., 2017; Gao et al., 2019), enabling a quicker initiation of treatment for patients with distinct cancer biomarkers.

When modelling WSIs, employing DNNs like Convolutional Neural Networks (CNNs) is a standard strategy, however, applying them directly to WSIs at full resolution is fraught with challenges, as WSIs typically possess pixels in the millions and beyond the limits of current standard graphics processing units (GPUs). Hence, the commonly applied strategies for WSI processing usually involve splitting the image into small tiles and/or pre-compressing the image tiles into feature vectors to reduce the size of the input (Campanella et al., 2019; Li et al., 2021c; Hashimoto et al., 2020; Li et al., 2021a; Lu et al., 2021b; Kalra et al., 2021; Lipkova et al., 2022b).

Another obstacle in applying deep learning on WSIs is the difficulty in obtaining reliable annotations. Informative disease-relevant histomorphological features can be rare and subtle occurrences. Extensive tile-level annotations may incur extremely high labour costs and are often impractical to acquire (Brunt, 2010; Hekselman and Yeger-Lotem, 2020). On the other hand, weakly-supervised learning-based methods model the tile-to-slide correlations to achieve slide-level predictions without the need for the presence of tile-level annotations at training. Particularly, Multiple Instance Learning (MIL), has received significant attention for its effectiveness in tackling the complexities of WSI analysis. Since weakly supervised labels are meaningless for a considerable number of background tiles in whole slide images (WSI), the advantages of Multiple Instance Learning (MIL) — such as task-driven perception of overall morphology and dynamic selection of informative tiles — are not attainable with subset-based methods like pre-selecting of essential regions or representative tile sets (Barker et al., 2016; Kalra et al., 2020). MIL primarily involves tile-level encoders for feature compression of tiles and slide-level aggregators to integrate all tile-level features within a single WSI.

In addition, predicting molecular traits from histomorphology poses some specific challenges that are less relevant in other computational pathology tasks like tumour detection. Firstly, collecting gene sequencing-supported subtype annotations requires strict quality control, is expensive, and often, different molecular subtypes can exhibit visually similar phenotypes on H&E slides, making molecular subtypes not as distinguishable as other histopathological classification tasks (Bilal et al., 2021; Yang et al., 2022; Tomita et al., 2022). As a result, successful recent research for molecular subtyping normally requires a considerable amount of training samples (Sirinukunwattana et al., 2021; Hong et al., 2021; Niehues et al., 2023). Some methods leverage multi-omics data (Couture et al., 2018; Tsai et al., 2023), immunohistochemistry (IHC) stained images targeting proteins associated with bespoke phenotypes (Lu et al., 2022; Niyas et al., 2023; Huang et al., 2023), or manual pixel-level region of interest (ROI) annotations (Huang et al., 2023) for training. Given the cost and effort necessary to provide such additional information, well-annotated patient cohorts that are suitable for developing models for molecular subtyping are often small. This often necessitates that models search for highly heterogeneous discriminative features helpful for molecular typing across multiple scales, from cellular level to slide level (Gao et al., 2022).

In this paper, we present a novel MIL-based approach for WSI-based morpho-molecular subtyping. Unlike most existing frameworks that utilise pretrained tile-level encoders, our work, named Inter-MIL, enables end-to-end training of the tile-level encoder jointly with the slide-level feature aggregator, allowing more task-specific discriminative features to be learned at the tile level. To optimise the tile-level encoder, we employ a pseudo label propagating strategy which captures the interaction between tiles and slide-level labels. Notably, this strategy also improves the aggregator responsible for summarising the tile-level features, leading to better global tissue features. Visualisation reveals that during the iterative optimisation, the aggregator's attention is generally directed to better align with regions that experts find significant. At the same time, the tile-level image features become more discriminative. Moreover, using gradient-based methods to examine encoder activations, we extract finer-grained attentions that capture cellular features within individual tiles, surpassing the capabilities of a task-agnostic pretrained encoder. With the proposed Inter-MIL, the tile-level features are extracted in a task-relevance fashion.

The main contributions in this study are summarised with the help of Fig. 1:

(1) *Novel Inter-MIL framework* - We design an iterative optimisation strategy through communication between features with local- and large-scale granularity. Our proposed method — Inter-MIL improves the learning efficiency of MIL on small histopathological datasets. Inter-MIL also introduces optimising steps at various feature scales for both the tile-level encoder and slide-level aggregator.

(2) *More representative features* - Inter-MIL searches for representative features of molecular subtypes from multiple scales, allowing the identification of cytopathological features as well as improving the search for coarse-grained features.

(3) *Inter-MIL features improve discrimination* - Inter-MIL reshapes the tile-level feature space, making the visual features of different molecular subtypes more distinguishable, thereby reducing the difficulty of subtyping new samples.

(4) *Validation of multiple molecular subtyping tasks* - We consider four very different molecular subtyping tasks, including the prediction of high epithelial–mesenchymal transition (EMT) in serous epithelial ovarian cancer (SOC) (Hu et al., 2020, 2021), the prediction of Kirsten rat sarcoma viral oncogene (KRAS) mutation status in colon (Abraham et al., 2021) and lung cancer (Coudray et al., 2018; Jain and Massoud, 2020), epidermal growth factor receptor (EGFR) mutation status in lung cancer (Coudray et al., 2018; Jung et al., 2022), and human epidermal growth factor receptor 2 (HER2) amplification in breast invasive cancer (Binder et al., 2021).

2. Related works

2.1. Whole slide image analysis

As one of the most critical data carriers in digital pathology, we have witnessed rapid adoption of WSI in both the clinic and research as the digitisation of physical histology samples enables automatic software analysis, advanced data management, and remote viewing & conferencing (Hoque et al., 2024; Kassab et al., 2024). A large volume of research has been focused on analysing WSIs using DNNs. Various methodologies for improving diagnostic accuracy, prognostication, and identifying ambiguous and high-risk cases prioritised for detailed molecular testing and immunohistochemistry have been proposed (Chen et al., 2022c; Lipkova et al., 2022a; Bilal et al., 2021; Kers et al., 2022; Cen et al., 2024; Yan et al., 2023).

Although this paper focuses on supervised learning for WSIs, depending on the available supervisory information in various tasks, researchers usually adopt different DNN methods for WSI analysis.

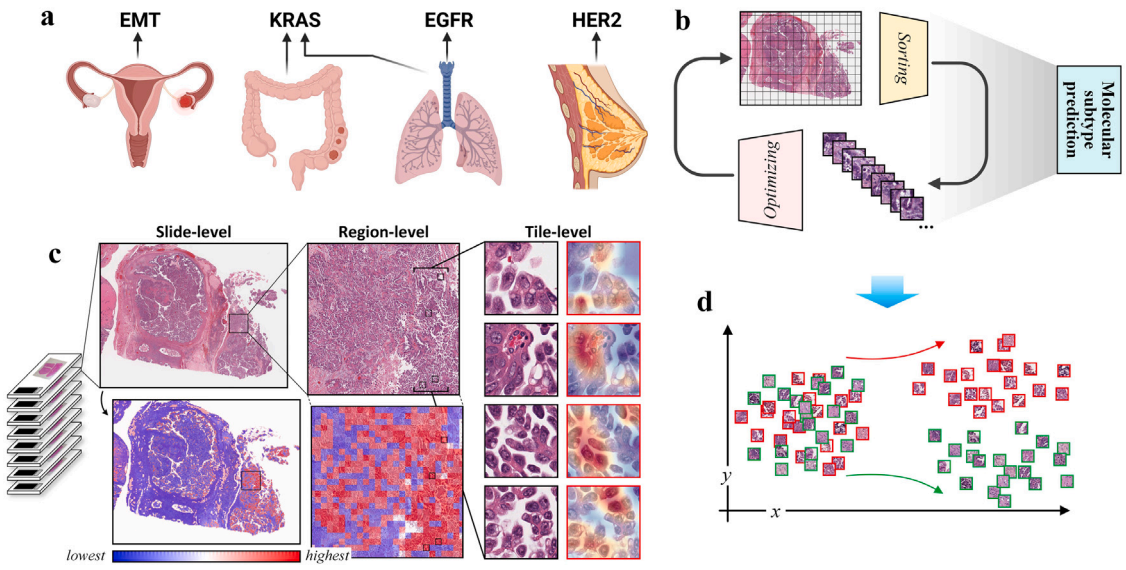

Fig. 1. Molecular trait prediction and feature investigation: task, method, and vision. a, Task: Prediction of 4 molecular traits on datasets of 4 cancer types. b, Method: Our novel Inter-MIL approach to drive self-interaction between global biopsy WSI features and fine-grained tile-level features. c, Vision: From left to right, presented at slide-level, region-level, and fine-grained tile-level attention interpretation of models, where the Grad-CAM tool (Selvaraju et al., 2017) provides an attention interpretation that highlights individual nuclei. d, Vision: The proposed Inter-MIL approach is expected to provide a more discriminative feature space for informative tiles from all slides.

Some basic tasks on WSIs at the tile-level and pixel-level, such as malignant tissue identification (Chen et al., 2022a), cells or nuclei segmentation (Li et al., 2022; Pan et al., 2023), etc., can assist pathologists in quickly locating lesion areas on WSIs. Such research often relies on labour-intensive, fine-grained, and high-quality annotations. Some studies utilise multi-stage learning to generate pseudo-masks or employ attention feedback from tile-level classification (Han et al., 2022; Guo et al., 2023; Li et al., 2023), thereby reducing the dependency on high-quality pixel-level annotations in cell segmentation tasks. Some research combats the impact of unstable annotation quality by adopting a novel evaluation framework for various DNN frameworks (Springenberg et al., 2023). In advanced tasks like tumour grading and immunotyping, expert annotations on Regions of Interest (ROIs) aid models in filtering noise (Yan et al., 2023; Godson et al., 2024). However, rich, high-quality supervisory annotations are costly and subject to pathologists' bias (Aubreville et al., 2023). Thus, many WSI analysis tasks apply weak supervision mode, enabling models to analyse entire WSIs with single slide-level annotations.

As mentioned, one family of weakly supervised learning methods frequently used in WSI analysis is MIL. Here, tiles cropped from a WSI are considered individual instances. The WSI is considered to be a bag containing these instances (Ilse et al., 2018). The original formulation of MIL for binary classification deploys a basic independent and identically distributed (i.i.d.) assumption for the instances, meaning that the label of each instance is assumed to be independently drawn from a Bernoulli distribution (Campanella et al., 2019). Recent efforts have also included modelling the correlation between the instances, using architectures such as graph models or self-attention (Lee et al., 2022; Zhao et al., 2020; Chen et al., 2022b). Moreover, MIL's performance varies in different tasks; it is more stable in tasks like survival prediction (Lu et al., 2020) or distinguishing heterogeneous phenotypes, like LUDA vs. LUSC in lung cancer (Cao et al., 2023), than in molecular subtyping, which can be challenging due to unclear histomorphology distinctions (Niehues et al., 2023).

2.2. Multiple instance learning

Mainstream MIL approaches used for analysing WSIs can be roughly separated into two categories. Instance-based methods optimise the subtyping likelihood of a subset of tile images that are representative of the whole slide (Coudray et al., 2018; Campanella et al., 2019). In these methods, the randomness in the initial selection of representative tiles may lead to difficulties in optimisation convergence. Additionally, it relies on assigning pseudo-labels to representative tile-level features for slide-level modelling, which could inadvertently ignore the global morphological features.

Another category of more complex methods is embedding-based, where the learning is generally separated into two stages: (1) Encoding: embedding tiles into an abstract feature space and (2) Aggregating: summarising the tile embeddings for a slide-level embedding and then scoring the slide-level embedding. The goal of the encoding phase is to obtain compressed representations of tiles, which is usually performed using a pretrained neural network. The pretraining tasks can be ImageNet classification (Deng et al., 2009), self-supervised learning (Uegami et al., 2022; Krishnan et al., 2022), or an easier task related to the target task (Kalra et al., 2021). The aggregator fuses the tiles embeddings of a WSI to produce a global representation and performs the final prediction. Due to the fact that informative regions may only occupy a small portion of WSI, attention-based pooling is often used to select potentially representative tiles while suppressing the contribution of other, noisy regions (Ilse et al., 2018; Lu et al., 2021a). The attention-based pooling scheme is also frequently used in conjunction with multi-resolution representations, clustering, self-attention layers (e.g. Transformers Azad et al., 2024a), and graph-based models, enabling the model to integrate contextual information or prior knowledge regarding tissue morphology (Lu et al., 2021b; Li et al., 2021a; Hashimoto et al., 2020; Li et al., 2021b; Shao et al., 2021; Zhao et al., 2022; Lu et al., 2022; Azad et al., 2024b; Xing et al., 2024). Improved MIL variants focus on superior tile-level feature encoding by employing strategies such as fine-grained pre-learning on "benign vs. malignant" tiles (Zhao et al., 2024), utilising cell segmentation for dispersed attention (Kapse et al., 2024), and incorporating external genomic knowledge via distillation models (Xing et al., 2024). Multi-task MIL naturally integrates multi-perspective supervisory information within a unified learning framework, thereby enhancing MIL's comprehensive understanding of pathological phenotypes (Alsaafin et al., 2023). Moreover, some improved MIL optimises attention representation with advanced pooling methods at the aggregation stage (Oner et al., 2023).

Algorithm 1 Construction of tile-level fine-grained feature pool (for each slide)

Input: Ranked tiles: $\{x_{d_1}, x_{d_2}, \ldots, x_{d_L}\}$; Sampling numbers: k^1, k^2, n; Sampling range: K, N, for each slide with different number L of tiles.
Output: Tile-level pool: S; Negative tile-level pool: S^{neg}

$\{x_{d_i}\}_{i=1}^{K} \leftarrow \{x_{d_1}, x_{d_2}, \ldots, x_{d_L}\}[:K]$, $\{x_{d_j}\}_{j=K+1}^{L} \leftarrow \{x_{d_1}, x_{d_2}, \ldots, x_{d_L}\}[K+1:]$;
$S^{top} \leftarrow Random\left(\{x_{d_i}\}_{i=1}^{K}, k^1\right)$, $S^{sup} \leftarrow Random\left(\{x_{d_j}\}_{j=K+1}^{L}, k^2\right)$;
$S^{pos} \leftarrow S^{top} \cup S^{sup}$, $S^{neg} \leftarrow None$;
if $N > 0$ **then**
    $\{x_{d_i}\}_{i=(L-N)+1}^{L} \leftarrow \{x_{d_1}, x_{d_2}, \ldots, x_{d_L}\}[(L-N)+1:]$
    $S^{neg} \leftarrow Random\left(\{x_{d_i}\}_{i=(L-N)+1}^{L}, n\right)$
end ifreturn S^{pos}, S^{neg}

Algorithm 2 Histopathology subtyping workflow of Inter-MIL framework

– Train stage:

Input: training slides: $S_{train} = \{s_1, s_2, \ldots, s_{train}\}$; number of training epochs in each round for aggregator $f_{mlp}(\cdot) = \{att(\cdot), f_{cls}(\cdot)\}$ and encoder $f_{cnn}(\cdot)$: ep_{mlp} (specially, $ep_{mlp} \leftarrow ep_{mlp}^{init}$ for the first round), ep_{cnn}; convergence point: $\mathcal{L}_{final}$.
Output: trained aggregator and encoder: $f_{mlp}^{t}(\cdot)$, $f_{cnn}^{t}(\cdot)$.

while in round t, $\mathcal{L}_{s \in S_{train}} \geq \mathcal{L}_{final}$ **do**
    for each $s \in S_{train}$ **do**
        load embedding: $E_s^{t-1} \leftarrow f_{cnn}^{t-1}(s)$
    end for
    for each $ep \in [1 : ep_{mlp}]$ **do**
        $f_{mlp}^{t}(\cdot) \leftarrow$ train on $\{E_{s_1}^{t-1}, E_{s_2}^{t-1}, \ldots, E_{s_{train}}^{t-1}\}$, with forward-func. Eq. (1) and loss $\mathcal{L}_{s \in S_{train}}(y_s, y_{cls})$.
    end for
    S^{pos}, $S^{neg} \leftarrow$ construct from $s \in S_{train}$, according to Algorithm 1.
    for each $ep \in [1 : ep_{cnn}]$ **do**
        if S^{neg} is *None* **then**
            $f_{cnn}^{t}(\cdot) \leftarrow$ train as Eq. (6)
        else
            $f_{cnn}^{t}(\cdot) \leftarrow$ train as Eq. (7)
        end if
    end for
    update: $\mathcal{L}_{s \in S_{train}}$, t
end whilereturn $f_{mlp}^{t}(\cdot)$, $f_{cnn}^{t}(\cdot)$

– Test stage:

Input: testing slides: $S_{test} = \{s_1, s_2, \ldots, s_{test}\}$; aggregator and encoder: $f_{mlp}^{t}(\cdot)$, $f_{cnn}^{t}(\cdot)$.
Output: prediction results: $Y_{cls} = \{y_{cls}^{s_1}, y_{cls}^{s_2}, \ldots, y_{cls}^{s_{test}}\}$.

for $s_p \in S_{test}$ **do**
    load embedding: $E_{s_p}^{t} \leftarrow f_{cnn}^{t}(s_p)$
    $y_{cls}^{s_p} \leftarrow \mathrm{softmax}(f_{mlp}^{t}(E_{s_p}^{t}))$
end forreturn $Y_{cls} = \{y_{cls}^{s_1}, y_{cls}^{s_2}, \ldots, y_{cls}^{s_{test}}\}$

However, the majority of these methods do not optimise the tile-level encoder with respect to the prediction task in a closed-loop and instead resolve to pretrained encoders obtained from a proxy task agnostic to the downstream prediction. This strategy limits the aggregator's ability to perceive fine-grained information (Zhang et al., 2022). Consequently, most histopathological image analysis frameworks utilise large cohorts with hundreds or thousands of WSIs for training, compensating for sub-optimal fine-grained tile-level features (Coudray et al., 2018; Campanella et al., 2019; Lu et al., 2021b,a; Lipkova et al., 2022b; Li et al., 2021a).

3. Materials and methods

3.1. Datasets and processing

Experimental datasets. In this section, we describe the datasets used to validate the proposed approach. We conducted four subtyping tasks: (1) ***OV-EMT:*** Approximately 20% of serous ovarian cancers (SOCs) are classified as EMT-high tumours, which are associated with poor survival (Hu et al., 2021). Here, we analysed 70 WSIs from TCGA-OV dataset with a binary EMT status (38 EMT-high vs. 32 EMT-low); (2) ***COLU-KRAS:*** Mutations in the KRAS gene are often associated with different cancer types, including lung and colorectal cancer (Abraham et al., 2021; Jain and Massoud, 2020). The presence of KRAS mutations in colorectal cancer can have implications for treatment decisions (Lievre et al., 2006). Here we present a combined cohort of 112 WSIs with KRAS mutation status (44 mutated vs. 68 wild-type) from TCGA-COAD and TCGA-LUAD datasets; (3) ***LU-EGFR:*** Detection of EGFR mutations is now a standard part of the diagnostic workup for patients with non-small cell lung cancer (NSCLC), as it helps guide treatment decisions (Li et al., 2013). Here we utilised 261 WSIs from TCGA-LUAD dataset for subtyping EGFR mutation status (75 mutated vs. 186 wild-type); (4) ***BR-HER2:*** HER2 is a protein that is overexpressed in approximately 15%–20% of breast cancers. HER2-positive breast cancers tend to be more aggressive and less responsive to hormone treatments compared to HER2-negative cancers (Gianni et al., 2010). 415 WSIs from the TCGA-BRCA dataset where HER2 status was determined based on fluorescence amplification in situ hybridisation (FISH) expression (77 positives vs. 338 negatives) were used. For annotations, EMT status used in *OV-EMT* is available in Hu et al. (2020)

Algorithm 3 Contrastive training for instance-bag aggregator

Input: aggregator: $f_{mlp}(\cdot)$; the set of embeddings of train slides: $S_{\mathrm{E}}^{train} = \{\mathrm{E}_1, \mathrm{E}_2, \ldots, \mathrm{E}_{train}\}$; number of sampled contrastive pairs N_{pt}, the size of sampled embeddings bag d_{pt}; number of pre-training epochs ep_{pt}.

Output: pretrained aggregator: $f_{mlp}^{pt}(\cdot)$.

for each $ep \in [1 : ep_{pt}]$ **do**
  $S_{pt}^{q} \leftarrow \{\mathrm{E}_1^{\mathrm{q}}, \ldots, \mathrm{E}_{\mathrm{N_{pt}}/2}^{\mathrm{q}}\}$, $\mathrm{E}_{\mathrm{i}}^{\mathrm{q}} \leftarrow rand(\mathrm{E_i}, d_{pt})$ from any $\mathrm{E_i} \in S_{\mathrm{E}}^{train}$.
  $S_{pt}^{k+} \leftarrow \{\mathrm{E}_1^{\mathrm{k+}}, \ldots, \mathrm{E}_{\mathrm{N_{pt}}/2}^{\mathrm{k+}}\}$, $\mathrm{E}_{\mathrm{j}}^{\mathrm{k+}} \leftarrow rand(\mathrm{E_j}, d_{pt})$ from $\mathrm{E_j}$ which belongs same subtype with $\mathrm{E}_{\mathrm{j}}^{\mathrm{q}}$.
  $S_{pt}^{k-} \leftarrow \{\mathrm{E}_1^{\mathrm{k-}}, \ldots, \mathrm{E}_{\mathrm{N_{pt}}/2}^{\mathrm{k-}}\}$, $\mathrm{E}_{\mathrm{h}}^{\mathrm{k-}} \leftarrow rand(\mathrm{E_h}, d_{pt})$ from $\mathrm{E_h}$ which belongs different subtype with $\mathrm{E}_{\mathrm{h}}^{\mathrm{q}}$.
  $S^{q\cdot k} \leftarrow \{(\mathrm{E}_1^{\mathrm{q}}, \mathrm{E}_1^{\mathrm{k+}}), \ldots, (\mathrm{E}_{\mathrm{N_{pt}}/2}^{\mathrm{q}}, \mathrm{E}_{\mathrm{N_{pt}}/2}^{\mathrm{k+}}), (\mathrm{E}_1^{\mathrm{q}}, \mathrm{E}_1^{\mathrm{k-}}), \ldots, (\mathrm{E}_{\mathrm{N_{pt}}/2}^{\mathrm{q}}, \mathrm{E}_{\mathrm{N_{pt}}/2}^{\mathrm{k-}})\}$
  $f_{mlp}^{pt} \leftarrow$ train on $S^{q\cdot k}$, with loss $\mathcal{L}_{pt}$ as Eq. (8).
end forreturn $f_{mlp}^{pt}(\cdot)$

and Hu et al. (2021) while the subtype labels for the other tasks are available in the TCGA repository (Tomczak et al., 2015).

In addition to the TCGA cohorts, samples of patients with newly diagnosed advanced colorectal cancer taken FOCUS clinical trial (part of S:CORT project) (Seymour et al., 2007; Malla et al., 2021; Sirinukunwattana et al., 2021) were used, which contains 666 slides of resection specimens from 362 patients. For all samples, histology slides matched RNA sequencing results are available. In our study, a subset of FOCUS was used for external validation. In which, we selected 200 WSIs from the first 100 patients, with KRAS mutation status as the label. For the patients, the distribution of the KRAS status is 56 mutated vs. 44 wild-type. Additional information on the external validation dataset and preprocessing details, including the data acquisition protocol and population characteristics of the FOCUS cohort, is in the supplementary material.

The proximity the tissue used for generating H&E slides and RNA sequencing is one critical quality metric that needs to be taken into consideration. The FOCUS cohort has been generated under a very strict quality assurance protocol. However, it is known that the correlation between H&E slides and RNA sequencing in the TCGA cohort is more variable.

Cohort curation. Here we specify which slides were selected from each of the cohorts and how we addressed the problem of missing information. The slide sets and metadata are collected from the TCGA repository (Tomczak et al., 2015). We use only the digitised Formalin-Fixed Paraffin-Embedded (FFPE) slides, as it is the gold standard for histopathological diagnosis. We exclude the following cases due to technical artefacts: 1. In TCGA-OV cohort, we exclude 37 WSIs without available EMT-score; 2. For TCGA-COAD cohort, we exclude 413 WSIs out of 459 and for TCGA-LUAD cohort, we exclude 475 WSIs out of 541 as their labels on KRAS status are labelled as 'Not Available' or 'Unknown'. The remaining 46 WSIs from TCGA-COAD and 66 WSIs from TCGA-LUAD are combined as the experimental dataset for task *COLU-KRAS*; 3. In TCGA-LUAD cohort, we exclude 280 WSIs out of 541 with labels on EGFR as 'Not Available' or 'Unknown'; 4. In TCGA-BRCA cohort, we exclude 718 WSIs out of 1133 with labels on HER2 as 'Not Evaluated'.

For each slide, we extract 256 × 256 tiles without overlap at 40× magnification (0.25 μm per pixel) from tissue regions for TCGA cohorts and 20× (0.5 μm per pixel) for FOCUS. These morphomolecular subtyping tasks are challenging not only because of the nature of the problem but also because of the small cohort size as well as a combination of different tissue types (*COLU-KRAS*), and severe class imbalance (*LU-EGFR* and *BR-HER2*). To ensure enough training samples, we split the data into training/test sets with the following ratio: 70%/30% for *OV-EMT* and *COLU-KRAS*, 50%/50% for *LU-EGFR* and *BR-HER2*. For *FOCUS-KRAS* cohort, 2 evaluation modes are utilised, (1) Re-train mode: the cohort is split to training/test sets with a ratio 70%/30%; (2) External-test mode: we apply all samples for testing.

Next, we comment on our annotation protocol. In the *OV-EMT* task, the continuous EMT scores, which range between 0 and 1, are derived from the results of sequencing analyses previously conducted by our collaborators (see Hu et al. (2020)), and the median EMT score of the entire cohort is used to distinguish between EMT-low/high. For the other three tasks, all annotations can be found in the clinical records of TCGA, and they are discrete labels: 'YES/NO' for KRAS and EGFR status in TCGA-COAD and TCGA-LUAD, and 'Positive/Negative' for HER2 status in TCGA-BRCA.

Whole slide image preprocessing. The preprocessing of WSIs involves the following steps: (1) Removing grey or white background; (2) Removing blue and green contaminated areas; (3) Removing markers of the black, red, green and blue pens. All these preprocessing steps are conducted with OpenCV tools (pypi.org/project/opencv-python/). Then, we divide WSIs into tiles, after calculating the valid tissue proportion for each tile, we discard those with a proportion of tissue less than 70%.

3.2. Overall framework of inter-MIL

In this paper, a self-interactive multi-instance learning (Inter-MIL) approach is proposed, which utilises two modules to model fine-grained tile-level features and global slide-level features, respectively. These two modules interact with each other to achieve mutual optimisation.

As mentioned, the proposed Inter-MIL framework consists of two learnable neural network modules: Module-1, the instance-bag aggregator (aggregator) based on Gated Attention Pooling network, and Module-2, the trainable tile-level feature encoder. These modules are optimised alternatively until convergence conditions are met, as illustrated by Fig. 2, which depicts the details in the main framework of Inter-MIL method as well as supporting modules.

To begin with, given a WSI with L tiles $\{x_i\}_{i=1}^{L}$ and the tile-level feature encoder $f_{cnn}(\cdot)$, we let $\mathrm{E} = \{E_i\}_{i=1}^{L}$ be the set of tile embeddings, such that $E_i = f_{cnn}(x_i)$. The dimension of tile embedding is $E_i \in \mathbb{R}^{512}$. In Module-1, the AttPool or Gated-AttPool based aggregator (Ilse et al., 2018; Lu et al., 2021a) receives tile embeddings E and outputs a classification result y_{cls}:

$$y_{cls} = \mathrm{softmax}\left(f_{cls}\left(\sum_{i=1}^{L} att\left(E_i\right) \cdot E_i\right)\right), \tag{1}$$

where $f_{cls}(\cdot)$ is the output layer for classification, the attention score $att(E_i) \in [0, 1]$ reflects the contribution of the ith tile to the classification, and $\sum_i^L att(E_i) = 1$. We train an aggregator by optimising the slide-level prediction, which can be formalised as follows:

$$\theta_{f_{cls}^{t}} = \theta_{f_{cls}^{t-1}} + \nabla \mathcal{L}_{E \in S_{train}}\left(y_E, y_{cls}\right), \tag{2}$$

where t and $t-1$ indicate the current and previous training loops, the tile embedding set E comes from training slide set S_{train}, and y_E denotes its subtype label.

Here, Inter-MIL aims to train fine-grained histological features on representative tiles in multiple iterations, which helps to optimise the WSI embedding set E for the aggregator. In Module-2, to refresh tile embeddings E^{t-1} to E^t for the tth training loop, we use k representative

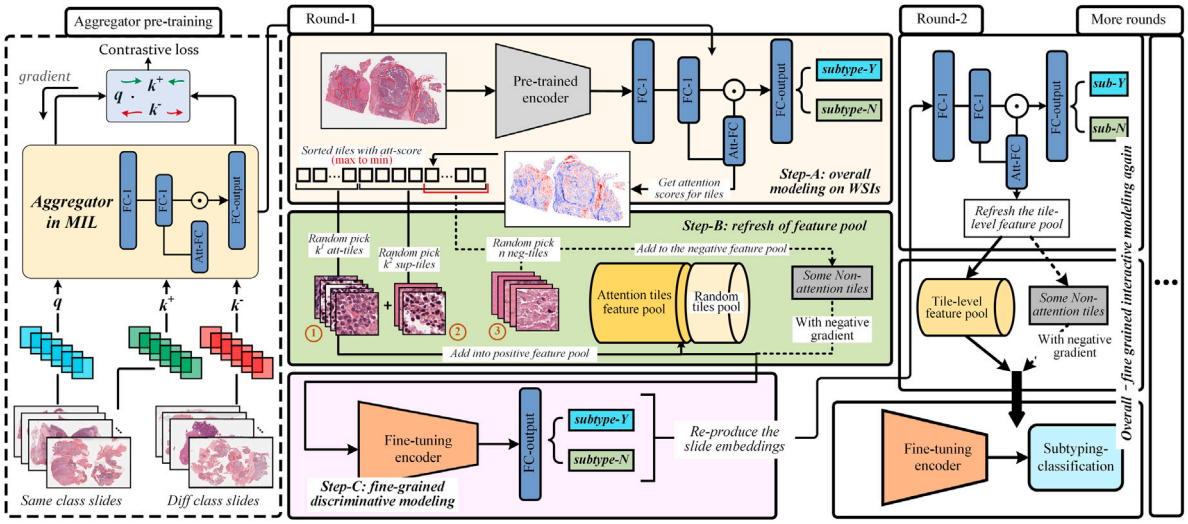

Fig. 2. Overview of the Inter-MIL framework. The framework is divided into three parts, from left to right, highlighting its various components and functions. Left, The aggregator is pretrained using contrastive learning, where embeddings of slides of the same and different subtypes are fed into the aggregator in pairs. The training objective is to minimise the distance between embeddings from the same subtype while maximising the distance between embeddings from different subtypes. Middle, The self-interaction MIL algorithm consists of three steps within each round: (1) train AttPool network with pretrained tile embeddings to obtain the attention value for each tile, (2) constructing a tile-level feature pool with high-attention tiles and supplementary tiles (defined in Eq. (4)), optionally including low-attention tiles (defined in Eq. (5)), and (3) fine-tuning the CNN encoder (using ResNet) with the tile-level feature pool. Right, The tile embeddings are reproduced for the next round of AttPool training, and the subsequent rounds of self-interaction MIL continue until convergence is achieved.

tiles $x_{i=1}^{k}$ with high attention scores to fine-tune $f_{cnn}^{t-1}(\cdot)$ to $f_{cnn}^{t}(\cdot)$, which encodes the fine-grained features. The training of $f_{cnn}^{t}(\cdot)$ aims to optimise prediction loss at the tile level. After this step, we regenerate the tile embeddings E^t using the fine-tuned feature encoder $f_{cnn}^{t}(\cdot)$ and continue with the next training round for the aggregator:

$$E^t = \left\{E_i^t = f_{cnn}^t(x_i)\right\}_{i=1}^{L}. \tag{3}$$

As shown by Fig. 2, in the middle column, the optimisation of Module-1 (beige block) and Module-2 (pink block) constitutes a training loop (round) of Inter-MIL, while the right column represents subsequent training rounds, which repeats the switching and interaction in overall and fine-grained feature optimisation until convergence. Constructing tile-level training materials (shown in green blocks) is a critical step in the Inter-MIL framework.

We now elaborate on the strategy for selecting representative training tiles. Our target is to construct a tile-level feature pool S^{pos}. Given a WSI, we rank tiles based on their attention scores in a monotonically decreasing order, i.e. $\{x_{d_1}, \ldots, x_{d_L} | att(E_{d_1}) \geq att(E_{d_2}) \geq \cdots \geq att(E_{d_L})\}$, where $att(E_i)$ refers to Eq. (1). We define a set of **attention tiles** S^{top} as a set of randomly sampled k^1 tiles out of the top K highest attention tiles, i.e. $S^{top} = \{x_i\}_{i=1}^{k^1} \subseteq \{x_{d_i}\}_{i=1}^{K}$. Similarly, we define a set of **supplementary tiles** S^{sup} as a set of randomly sampled k^2 tiles out of the remaining $L - K$ tiles, i.e. $S^{sup} = \{x_i\}_{i=1}^{k^2} \subseteq \{x_{d_j}\}_{j=K+1}^{L}$. K is determined by the total number of tiles L separately for each WSI. S^{sup} increases the diversity of tile features that may not be captured by S^{top}. We construct the set S^{pos} of tile-level feature pool as follows:

$$S^{pos} = S^{top} \cup S^{sup}. \tag{4}$$

The main goal of training on the representative tile feature set S^{pos} is to optimise the representation of tile-level feature encoders for fine-grained histological features. However, to attenuate the influence of noisy tile-level features. We also want the encoder to learn to distinguish and discard non-relevant tiles that could end up being allocated high attention scores. Therefore, we construct a set of **negative tiles** S^{neg} as:

$$S^{neg} = \{x_i\}_{i=1}^{n} \subseteq \{x_{d_i}\}_{i=(L-N)+1}^{L}, \tag{5}$$

which consists of n randomly sampled tiles from the N lowest attention tiles from each WSI. We adopt a different optimisation strategy for negative tiles S^{neg} than for attention tiles S^{top} and supplementary tiles S^{sup}. More details on the optimisation strategy for S^{neg} are provided in the next subsection. The pseudocode for constructing the tile-level feature training repository is provided in Algorithm 1.

3.3. Optimisation of tile-level feature encoder

As with standard MIL training, the aggregator is optimised based on the slide-level annotation y_s of slide s and the feed forward process shown in Eq. (1). The aggregator is optimised according to loss of weighted cross-entropy: $\mathcal{L}(y_s, y_{cls})$. Additionally, the encoder $f_{cnn}(\cdot)$ is trained based on the tile-level feature pool S^{pos} (and S^{neg}) described in the previous section. The optimisation process is as follows:

$$\theta_{f_{cnn}^t} = \theta_{f_{cnn}^{t-1}} + \gamma \cdot \nabla \mathcal{L}_{x_i \in S^{pos}} \left(y_{x_i}, f_{cnn}^{t-1}\left(x_i\right)\right), \tag{6}$$

where $\mathcal{L}_{x_i}$ refers to the loss of tile-level inputs, θ denotes the weights of the encoder, and γ is the learning rate. y_{x_i} is the annotation of tile x_i inherited from the slide it belongs to (i.e., y_s).

In contrast to modelling representative tile-level information in S^{pos}, we aim to reduce noise interference in the model. Treating low-attention tiles in S^{neg} as noise, we train the model to discard these tile-level features using an adversarial optimisation approach (Ganin et al., 2016), This approach trains the encoder to classify the noisy samples as poorly as possible by back propagating the negative gradients. Thus, we extend the optimisation of the encoder $f_{cnn}(\cdot)$ in (6) to incorporate S^{neg} as follows:

$$\begin{aligned}\theta_{f_{res}^t} = \theta_{f_{res}^{t-1}} &+ \gamma \cdot \nabla \mathcal{L}_{x_i \in S} \left(y_{x_i}, f_{res}^{t-1}(x_i)\right) \\ &- \gamma^{neg} \cdot \nabla \mathcal{L}_{x_j \in S^{neg}} \left(y_{x_j}, f_{res}^{t-1}(x_j)\right),\end{aligned} \tag{7}$$

where γ^{neg} denotes the learning rate of the negative training.

The overall optimisation and prediction pseudocode of InterMIL is given by Algorithm 2.

3.4. Contrastive pre-training of instance-bag aggregator

We designed an additional module to improve the convergence of the instance-bag aggregator (slide-level classifier) and bestow the aggregator with a better initialisation that increases the training stability.

Table 1
Setting of hyperparameters.

Parameter	Value	Notation
K	$0.05 \times L$	Selection range of high-attention tiles, for fine-tuning tile-level encoder.
N	$0.2 \times L$	Selection range of low-attention tiles, for denoising with adversarial optimisation.
k^1	50	The number of attention tiles for fine-tuning tile-level encoder.
k^2	$0.4 \times k^1$	The number of supplementary tiles for fine-tuning tile-level encoder.
n	$0.2 \times k^1$	The number of negative tiles for denoising on the tile-level encoder.
ep_{mlp}	10	In each round of interaction, the training epoch of aggregator.
ep_{mlp}^{init}	Dynamic	In the initial round of interaction, $ep_{mlp} \leftarrow ep_{mlp}^{init}$, the end of ep_{mlp}^{init} depends on $\mathcal{L}_{init}$.
ep_{cnn}	2	In each round of interaction, the training epoch of tile-level encoder.
N_{pt}	6000	The number of sampled contrastive pairs, $N_{pt} = N_{k_+} + N_{k_-}$, and $N_{k_+} = N_{k_-} = N_q$.
d_{pt}	8000	The size of the bag of tiles sampled from each slide.
ep_{pt}	$\upsilon_1 = 30,\ \upsilon_2 = 50$	The number of epochs for aggregator pre-training, for tasks of *OV-EMT* and *COLU-KRAS*, pick υ_1, for tasks of *LU-EGFR* and *BR-HER2*, pick υ_2.

Inspired by self-supervised contrastive learning algorithms (Chen et al., 2020; He et al., 2020), we adapt the contrastive learning based on sample data enhancement to the scenario of WSI classification. Thus we can pre-train the MIL model for the recognition ability of tile-level bags. Unlike unsupervised contrastive learning, which determines whether augmented samples are of the same origin, we currently apply supervised annotations to evaluate whether randomly sampled tile bags belong to the same subtype. Given $\mathrm{E^q} = \{E_i\}_{i=1}^{N_q}$ as a query bag of tile embeddings randomly picked from slides of any subtype, where N_q is the sampling quantity, $\mathrm{E^k} = \{E_i\}_{i=1}^{N_k}$ is the key embedding bag from slides of the same or different subtype with $\mathrm{E^q}$, where $N_k = N_q$. The learning target of the aggregator pre-training is to reduce the distance between tile embedding bags from the same subtype, while increasing the distance between tile embedding bags from different subtypes. This results in the following loss function:

$$\mathcal{L}_{pt} = -log \frac{exp(\left|\mathrm{E^q} \cdot \mathrm{E^k} - y_c(q \cdot k)\right| / \tau)}{\sum_{i \in C} exp(\left|\mathrm{E^q} \cdot \mathrm{E^k} - y_c(i)\right| / \tau)}, \tag{8}$$

where $y_c(q \cdot k) \in \{0, 1\}$ indicates if $\mathrm{E^q}$ and $\mathrm{E^k}$ are from the same subtype and $C = 2$, τ is a sensitivity parameter and we set $\tau = 1$.

The leftmost part of Fig. 2 illustrates a schematic of the optional module for aggregator pre-training, and Algorithm 3 presents its pseudocode.

4. Experiments and results

4.1. Deep learning setup

Deep learning model implementation. The source code and detailed environment configuration files of this study are available at github.com/superhy/LCSB-MIL. The deep learning model is implemented using the PyTorch-1.6 framework. Training and testing of our model were performed on four NVIDIA RTX 2080Ti GPUs.

The following settings are used in all tasks unless specified otherwise. For the two modules of MIL workflow, we use ResNet-18 pretrained on ImageNet (Deng et al., 2009) as the vision encoder (encoder) to generate the initial feature embedding of tiles, and we apply the Gated-AttPool (Lu et al., 2021a) as the instances-bag aggregator (aggregator) and the subtyping classifier, which contains three fully connected layers for classification and one attention-based pooling layer. Furthermore, we choose the weighted cross-entropy (WCE) loss to tackle the class imbalance. The Adam optimiser with a learning rate of 0.0001 is applied, and we set the batch size of 8 WSIs for training the aggregator and 128 tiles for the encoder. In the training stage, each interactive round consists of 5 epochs for the aggregator and 2 epochs of vision encoder training. We set a delayed stop mechanism in the aggregator training of the first round to improve the initial stability of the proposed Inter-MIL framework. Moreover, we set the convergence point at which the overall training of Inter-MIL stops, and we describe these parameters in detail in the following paragraphs.

Hyper-parameters in Inter-MIL. In the experiments, we set the hyperparameters for Inter-MIL as listed in Table 1. We adopt a delayed stop strategy for training the aggregator in the initial round to ensure that valuable high-attention tiles are selected. Specifically, e_{mlp}^{init} will continue to train until the loss value drops to $\mathcal{L}_{init}$. The number of epochs for aggregator pre-training uses υ_1 and υ_2 based on the size of the training cohorts. For tasks *OV-EMT* and *COLU-KRAS*, due to the smaller training set, we opt for υ_1, while for tasks *LU-EGFR* and *BR-HER2*, with the larger training sets available, we chose υ_2 to enable longer pre-training. The hyperparameter setting for the number of training epochs is based on the convergence state of the training loss, and the hyperparameter setting for the contrastive pre-training of the aggregator is chosen to be as large as possible within the limits of computational resources and acceptable training time. More explanation of other hyperparameters' setting principles can be found in Supplementary Information.

Experimental setup. We compare our proposed Inter-MIL method with several state-of-the-art MIL algorithms, including CNN-MIL (Campanella et al., 2019), AttPool (Ilse et al., 2018), Gated-AttPool (Ilse et al., 2018; Lu et al., 2021a), CLAM (Lu et al., 2021b), and FocAtt-MIL (Kalra et al., 2021). To provide a direct comparison with our method, we select Gated-AttPool with a fixed pretrained CNN encoder as the baseline.

We conducted 10 runs for the *OV-EMT* task and 5 runs for the other tasks in the training/testing evaluation procedure. For different datasets and tasks, we randomly selected a fixed proportion (70% or 50%, as explained in Section 3.1) of the data as the training set each run, with the remaining 30% or 50% as the testing set. This splitting process was repeated 5 or 10 times to create 5 or 10 sets of training/testing combinations. Due to the small amount of training data in *OV-EMT* and *COLU-KRAS* tasks, we do not split the data further into a validation set.

In external validation on *FOCUS-KRAS* cohort, we carried out 2 modes of evaluation: (1) Re-train mode: The FOCUS dataset is split into training and testing sets in a 70%/30% ratio, we repeat this and generate 5 pairs of training/testing sets. Models are then retrained and tested for 5 runs; (2) External-test mode: We utilised the models which were trained from 5 runs on the *COLU-KRAS* task to test on the entire FOCUS dataset.

In general, the summary information about the experimental data preparation is listed in Table 2.

In all experiments, we use the last epoch's model for testing. We use the output of the aggregator as the final prediction of the Inter-MIL. The interactive training rounds in Inter-MIL (including adInter-MIL) have two options of termination conditions, and meeting either one will conclude the rounds: (1) A total of 5 rounds is reached; (2) The training loss of the aggregator falls below the termination threshold $\mathcal{L}_{final}$. We empirically determine the value of $\mathcal{L}_{final}$ based on the performance of the baseline model and set $\mathcal{L}_{init}$ as $\mathcal{L}_{final} + 0.15$.

We record the performance in terms of (1) the average and standard deviation (std) of the area under the receiver operating characteristic curve (AUC), and (2) the balanced accuracy (BACC) with macro averaging across multiple testing runs. Note that the standard deviation

Table 2
Summary of data preparation information.

Task	Source	Split (%)	Runs
OV-EMT	TCGA	49/21 (70%/30%)	10
COLU-KRAS		79/33 (70%/30%)	5
LU-EGFR		130/131 (50%/50%)	5
BR-HER2		208/207 (50%/50%)	5
FOCUS (re-train)	S:CORT	140/60 (70%/30%)	5
FOCUS (ext-test)		–/200 (–/100%)	5

Task	Training data	Testing data
OV-EMT	TCGA-OV	TCGA-OV
COLU-KRAS	TCGA-COAD, LUAD	TCGA-COAD, LUAD
LU-EGFR	TCGA-LUAD	TCGA-LUAD
BR-HER2	TCGA-BRCA	TCGA-BRCA
FOCUS (re-train)	FOCUS	FOCUS
FOCUS (ext-test)	TCGA-COAD, LUAD	FOCUS

captures the impact of random data split on the training process, which is helpful to reflect the stability of model training and performance variance in scenarios with small numbers of training slides.

4.2. Evaluation on slide-level subtyping

In this section, we analyse the performance of the Inter-MIL method on the molecular subtyping task and compare it with baseline methods.

For Inter-MIL, we define the following variants in our experimental results report: 1. Inter-MIL-b, which is a simplified version without random tiles-level features; 2. adInter-MIL, which is based on the standard Inter-MIL but adds adversarial training (Ganin et al., 2016) on noisy tiles. In addition to these variants, we also design a pre-training module for aggregation classifiers in MIL. This is a hot-swappable module suitable for most MIL methods, so we use the prefix "PT-" to indicate that this module is applied.

As shown in Tables 3 and 4, the Inter-MIL variants, compared to the baseline Gated-AttPool method, achieved at least a 6% improvement and demonstrate the best performance. As illustrated in Fig. 3-a, Inter-MIL and its variants outperform the baseline method on all four tasks. Investigating the optimisation process, as shown in Fig. 3-c, the loss of Inter-MIL in the early iterations is on par with the baseline method Gated-AttPool, but then the training rapidly converges after the interactive optimisation of tile-level encoder.

As shown in Table 5, Inter-MIL's performance on the external validation cohort is evident. "FOCUS (re-trained)" refers to results after training on 140 WSIs from 70 patients, using the optimal learning rate and stopping point from the baseline and test on the 60 WSIs from 30 patients. In contrast, "FOCUS (external-test)" indicates the testing results from the full 200 FOCUS WSIs, using a model trained on the TCGA-*COLU-KRAS* cohort.

The external validation results shown in Table 5 highlight the robustness of the Inter-MIL approach. Given the significant preanalytical variability between the TCGA and the FOCUS cohort, it is remarkable that the performance in FOCUS as an external validation cohort only decreases by about 3%. In comparison, other baseline methods show a significant drop in their KRAS mutation prediction on FOCUS samples, with performance declining by as much as 10%.

To provide some additional context to this external validation result we perform re-training on the FOCUS cohort. Based on the known characteristics of the FOCUS cohort the obtained improvement in AUC is in line with our expectation and it demonstrated that the proposed inter-MIL approach makes effective use of the added data.

In addition, pretraining the aggregation classifier accelerates the optimisation of both Gated-AttPool and Inter-MIL series. As evidenced by Figs. 3-b, -d, and 9-b (in supplementary material), after the aggregation classifier was contrastively pretrained, the convergence of the training is expedited by at least 10 epochs. This acceleration was even more pronounced in tasks such as *LU-EGFR* and *BR-HER2*, which have more training data, where the convergence was more than 40 epochs ahead. With the pre-training of aggregator for 30 epochs on tasks *OV-EMT* and *COLU-KRAS*, and for 40 epochs on the other two tasks with bigger datasets, the performance was further improved on tasks *OV-EMT* and *COLU-KRAS*, but not significantly on the tasks *LU-EGFR* and *BR-HER2*, as shown in Tables 3, 4, and Fig. 3-e. In general, across the four subtyping tasks, Inter-MIL consistently demonstrated superior performance in terms of AUC and BACC, especially when the size of the training set is small.

4.3. Interpreting the model at various scales

In this section, we highlight the better interpretability generated by Inter-MIL, which provides insight into the optimisation process of features at various scales, from fine-grained tile-level features to the global biologically relevant features. Fig. 3-f showcases a test sample from the *OV-EMT* task, which demonstrates the interpretation workflow from bottom to top: (1) Inter-MIL iteratively optimises the encoder for tile instances. With different models, we use the Gradient-weighted Class Activation Mapping tool (Grad-CAM) (Selvaraju et al., 2017) to generate the gradient activation maps for individual tiles. Comparing the attention regions on example tiles between the baseline method and Inter-MIL, we observe that the high attention regions move toward cell nuclei areas in Inter-MIL as the tile level-encoder was iteratively optimised. (2) and (3) Macroscopically, we observe that the attention map also shifts from non-cancerous regions to tumour regions at the regional/global level during the optimisation. (4) We analyse the attention distribution statistics on the exemplar slide and find that tiles with the EMT prediction score $\geqslant 0.5$ receive higher attention in EMT-high cases while the EMT prediction score <0.5 co-occurs with higher attention in EMT-low cases. (5) By visualising the feature space of representative tiles in all test slides, we find that the tile-level features learned by Inter-MIL are significantly more differentiable than the baseline for tiles of different subtypes (i.e., EMT-Low vs. EMT-high). Therefore, the evolution of features and attention maps, the consistency of model attention and predictions, and an analysis of the latent feature space are now illustrated on a set of concrete examples.

4.3.1. Evolution of tile-level attention

Inter-MIL iteratively optimises the **tile-level** encoder to model fine-grained features, learning increasingly detailed histological features to enable MIL's aggregator to better assess the representativeness of each tile for subtyping. Fig. 4 illustrates the evolution of the attention distribution on tiles within a slide as the adInter-MIL improves the feature representation. After multiple rounds of interactive training, the model's attention shifts from the tiles of background tissue to the tiles on or near tumour regions (Fig. 4-a and b). Furthermore, after one interactive training, the attention distribution of tiles in different regions has changed significantly, while round-3 is further fine-tuned based on round-2. For instance, Fig. 4-a shows a WSI which contains a blood clot. After the third round of interactive training, the model no longer pays attention to this area which does not have any diagnostic relevance. Fig. 4-b presents an example where the attention shifts away from stromal tissue, instead, the model focuses more on the tumour areas, even showing the fissures between thin-strip-like tumour areas.

Examining the second part of Fig. 4-a and b, we compare the allocation of attention to tumour and non-tumour tiles before and after self-interactive training. Initially, attention towards non-tumour tiles is comparable to or even exceeds that of tumour tiles. However, after a few rounds of self-interactive training, attention towards non-tumour tiles diminishes to zero while tiles from tumour regions garner progressively more focus. Further analysis of tile-level heatmaps, using gradient-based activation maps, reveals a gradual concentrated attention on nuclear-containing regions within tumour tiles, moving away from background areas.

Table 3
Results on multiple molecular subtyping tasks with ROC-AUC ± std (%) over 10 runs for task *OV-EMT* and 5 runs for tasks *COLU-KRAS*, *LU-EGFR*, and *BR-HRER2*.

Methods	*OV-EMT*	*COLU-KRAS*	*LU-EGFR*	*BR-HER2*
CNN-MIL (Campanella et al., 2019)	59.86 ± 1.11	50.56 ± 0.15	–	–
AttPool (Ilse et al., 2018)	61.38 ± 1.35	61.41 ± 0.78	–	–
Gated-AttPool (Ilse et al., 2018; Lu et al., 2021a)	62.46 ± 1.13	59.94 ± 0.06	64.15 ± 0.39	54.84 ± 0.14
CLAM (Lu et al., 2021b)	62.62 ± 1.10	62.18 ± 0.80	65.17 ± 0.25	54.17 ± 0.11
FocAtt-MIL (Kalra et al., 2021)	56.89 ± 1.86	63.76 ± 1.05	64.19 ± 0.12	53.53 ± 0.15
Inter-MIL (ours)	71.91 ± 0.50	64.78 ± 0.18	69.81 ± 0.12	63.08 ± 0.04
Inter-MIL-b (ours)	68.53 ± 0.21	64.01 ± 0.3	67.99 ± 0.02	62.00 ± 0.07
adInter-MIL (ours)	**74.55 ± 0.43**	**66.34 ± 0.31**	**70.65 ± 0.08**	**63.21 ± 0.03**
PT - Gated-AttPool (ours)	*64.18 ± 1.01*	*62.34 ± 0.70*	*65.21 ± 0.25*	*57.21 ± 0.22*
PT - Inter-MIL (ours)	*74.41 ± 0.26*	*70.38 ± 0.28*	*70.15 ± 0.24*	*62.94 ± 0.03*
PT - adInter-MIL (ours)	*77.00 ± 0.40*	*71.38 ± 0.11*	*71.33 ± 0.08*	*64.02 ± 0.11*

Table 4
Results on multiple molecular subtyping tasks with BACC ± std (%) over 10 runs for task *OV-EMT* and 5 runs for tasks *COLU-KRAS*, *LU-EGFR*, and *BR-HRER2*.

Methods	*OV-EMT*	*COLU-KRAS*	*LU-EGFR*	*BR-HER2*
Gated-AttPool (Ilse et al., 2018; Lu et al., 2021a)	70.45 ± 1.87	59.41 ± 0.08	60.34 ± 0.20	53.29 ± 0.02
CLAM (Lu et al., 2021b)	69.00 ± 1.08	62.97 ± 0.27	61.11 ± 0.20	55.63 ± 0.11
FocAtt-MIL (Kalra et al., 2021)	61.71 ± 0.72	61.74 ± 0.10	57.32 ± 0.02	53.56 ± 0.02
Inter-MIL (ours)	84.86 ± 0.45	64.50 ± 0.45	66.04 ± 0.33	56.69 ± 0.07
adInter-MIL (ours)	**85.45 ± 0.48**	**65.85 ± 0.24**	**69.56 ± 0.03**	**63.21 ± 0.03**
PT - Gated-AttPool (ours)	*71.45 ± 0.89*	*62.49 ± 0.36*	*61.19 ± 0.21*	*57.40 ± 0.14*
PT - Inter-MIL (ours)	*85.45 ± 0.38*	*67.86 ± 0.42*	*69.81 ± 0.17*	*58.52 ± 0.04*
PT - adInter-MIL (ours)	*85.77 ± 0.61*	*69.78 ± 0.16*	*70.87 ± 0.15*	*63.67 ± 0.14*

Table 5
Results on KRAS mutation status prediction with ROC-AUC ± std (%) and BACC ± std (%) over 5 runs for external validation on FOCUS cohort, compared with the internal validation results on *COLU-KRAS* task.

Methods	FOCUS (re-trained)		FOCUS (external-test)		*COLU-KRAS* (for comparison)	
	AUC	BACC	AUC	BACC	AUC	BACC
AttPool (Ilse et al., 2018)	70.79 ± 0.39	60.79 ± 0.45	51.55 ± 0.06	50.32 ± 0.02	61.41 ± 0.78	58.83 ± 0.52
Gated-AttPool (Ilse et al., 2018; Lu et al., 2021a)	69.64 ± 0.23	62.06 ± 0.11	53.08 ± 0.12	52.49 ± 0.09	59.94 ± 0.06	59.41 ± 0.08
CLAM (Lu et al., 2021b)	71.49 ± 0.25	63.12 ± 0.44	51.27 ± 0.05	49.94 ± 0.00	62.18 ± 0.80	62.97 ± 0.27
Inter-MIL (ours)	**78.01 ± 0.14**	69.59 ± 0.04	**62.41 ± 0.12**	58.43 ± 0.30	64.78 ± 0.18	64.50 ± 0.45
adInter-MIL (ours)	75.27 ± 0.15	**70.07 ± 0.01**	61.89 ± 0.04	**58.61 ± 0.36**	66.34 ± 0.31	65.85 ± 0.24

Additional comparisons of attention maps of different methods, and more examples of the model's attention-shifting evolution process can be found in Figures 10, 11, 12, and 13. Figures 14 and 15 show the top attention tiles respectively given by adInter-MIL and baseline methods. These figures can be found in the Supplementary Material. In conclusion, the fine-grained attention maps are improved and become concentrated around informative features such as nuclei after Inter-MIL interaction training, with the tile-level encoder being optimised.

4.3.2. Consistency across attention and different classes

In each interactive training round, Inter-MIL optimises the tile-level encoder with high-attention tiles as training material, thereby enabling even other tiles to attain classification scores. In this section, we investigate whether, following the interactive optimisation of Inter-MIL, the tiles from high-attention areas across different molecular subtypes simultaneously receive higher predictive scores aligned with their specific subtype. For instance, in EMT-high slides, do high-attention tiles tend towards EMT-high scores? Conversely, in EMT-low slides, do high-attention tiles align with EMT-low scores? We demonstrate this by presenting positive and negative examples from tasks *OV-EMT* and *COLU-KRAS* in Fig. 5. For each example, the left shows the original slide, the middle visualisation shows the attention value for each tile, and then the right shows the fine-grained classification score on each tile, in which the attention value is taken from the aggregator after Inter-MIL training. We can see that for the EMT-low case, the classification outcome corresponding to the high attention area is closer to 0, while in the EMT-high case, the classification score corresponding to the high attention area is close to 1. A similar observation holds for the KRAS-no and KRAS-yes examples.

Additionally, the most right part of Fig. 5 presents the attention distribution of the tiles on the example slides. The upper figure illustrates that tile distribution across attention ranges approximates a normal distribution. In the lower figure, red bars denote the proportion of tiles classified as EMT-high and KRAS-yes (with scores > 0.5), whereas green bars indicate the proportion of tiles classified as EMT-low and KRAS-no (with scores < 0.5). These results validate that Inter-MIL ensures a more consistent alignment between the aggregator's attention distribution and the fine-grained classification scores at tile-level. Thereby indicating that Inter-MIL can facilitate communication and alignment of slide-level and tile-level features.

4.3.3. More discriminative features

In the preceding sections, we showcased how Inter-MIL facilitates the alignment of visual features across tile-level and slide-level. In this section, we demonstrate the proposed Inter-MIL leads to more discriminative feature space.

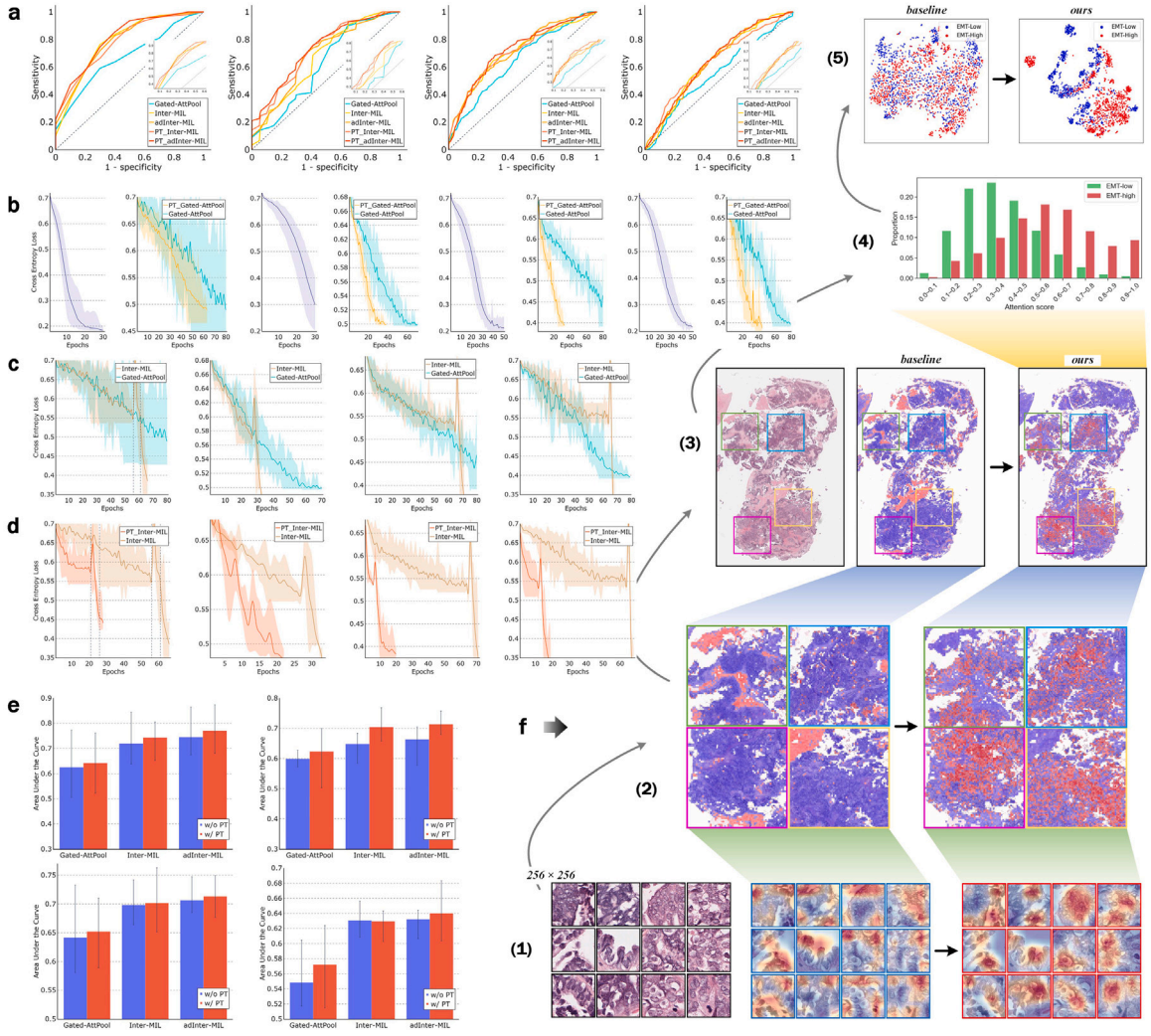

Fig. 3. Highlighted results and discussion. a, AUC-ROC curves of different models, for tasks: *OV-EMT*, *COLU-KRAS*, *LU-EGFR*, and *BR-HER2*, from left to right. Likewise below. b, log of loss for MIL aggregator pretraining and log of loss comparison for Gate-AttPool models with/without aggregator pretraining. c, log of loss comparison for Gate-AttPool models and Inter-MIL models. d, log of loss comparison for Inter-MIL models with/without aggregator pretraining. e, AUC performance comparison for Gate-AttPool, Inter-MIL, and adInter-MIL models with/without aggregator pretraining. From left to right in a∼d and left-top to right-bottom in e, the charts illustrate the results on *OV-EMT*, *COLU-KRAS*, *LU-EGFR*, and *BR-HER2* tasks. f, Model interpretation comparison at various scales of the baseline model (on the left) and adInter-MIL model (ours, on the right), which uses the case of *OV-EMT* task as an instance. From bottom to top, f-(1), Fine-grained scale. Gradient activation heatmaps of the instanced tile images. f-(2) and f-(3), Macroscopic scale. Attention heatmaps on representative regions, and their corresponding location on the slide. f-(4), Attention score statistics in the slide-level. Attention score distribution of tiles with different prediction results on EMT-low/high. f-(5), Feature space visualisation at the test cohort level. The feature space t-SNE (Van der Maaten and Hinton, 2008) mapping of high informative tiles from all test slides.

Fig. 6-a displays the feature distribution of the 100 highest and lowest attention tiles of all slides in the *OV-EMT* test set. It is mapped to the 2-dimensional coordinates with help of the t-SNE (Van der Maaten and Hinton, 2008) dimensionality reduction method. Green and yellow dots refer to the lowest-attention tiles from EMT-low/high slides, respectively, while blue and red dots refer to the highest-attention tiles from EMT-low/high slides. The distribution of tile features with high attention is clearly different from that of low attention tiles, both in the baseline GatedAttPool-MIL model (denoted as model 'X' in the figure) and adInter-MIL model (denoted as model 'Y' in the figure). However, there is a noticeable difference between GatedAttPool-MIL and adInter-MIL as the distributions of high attention tiles of EMT-low and EMT-high are clearly distinguishable in adInter-MIL model, but not in the GatedAttPool-MIL model. High-attention tile examples from EMT-low/high, and low-attention tiles are shown on the right side. It is noticeable that low-attention and high-attention tiles present distinct visual representations, yet differentiating high-attention tiles of EMT-low and EMT-high subtypes demands finer visual features.

Fig. 6-b illustrates the difference in feature distribution between the GatedAttPool-MIL model and the adInter-MIL model for high-attention and low-attention tiles, respectively. We note that the features of the high-attention tiles benefit more from the proposed Inter-MIL and show an improved separation of the relevant classes in feature space, while the low-attention tile features are less affected. This suggests that the improvement of classification is more significant on highly informative features rather than low attention regions that are potentially noisy.

Fig. 6-c further presents examples of querying high-attention tiles in the feature space. One example is EMT-high, and another is EMT-low. Dark green dots indicate high attention tiles from queried slides. We can observe that, whether it is EMT-low or EMT-high, the query points in the optimised feature space are closer to the corresponding subtype cluster and farther away from the others. In contrast, querying different subtypes with the baseline model is not as convenient.

More results regarding the learned tile-level features can be found in Figures 17 and 18 (see Supplementary Material). Among them, Figure 17 provides the feature distributions when sampling varying numbers of top attention tiles in all four molecular subtyping tasks. The results

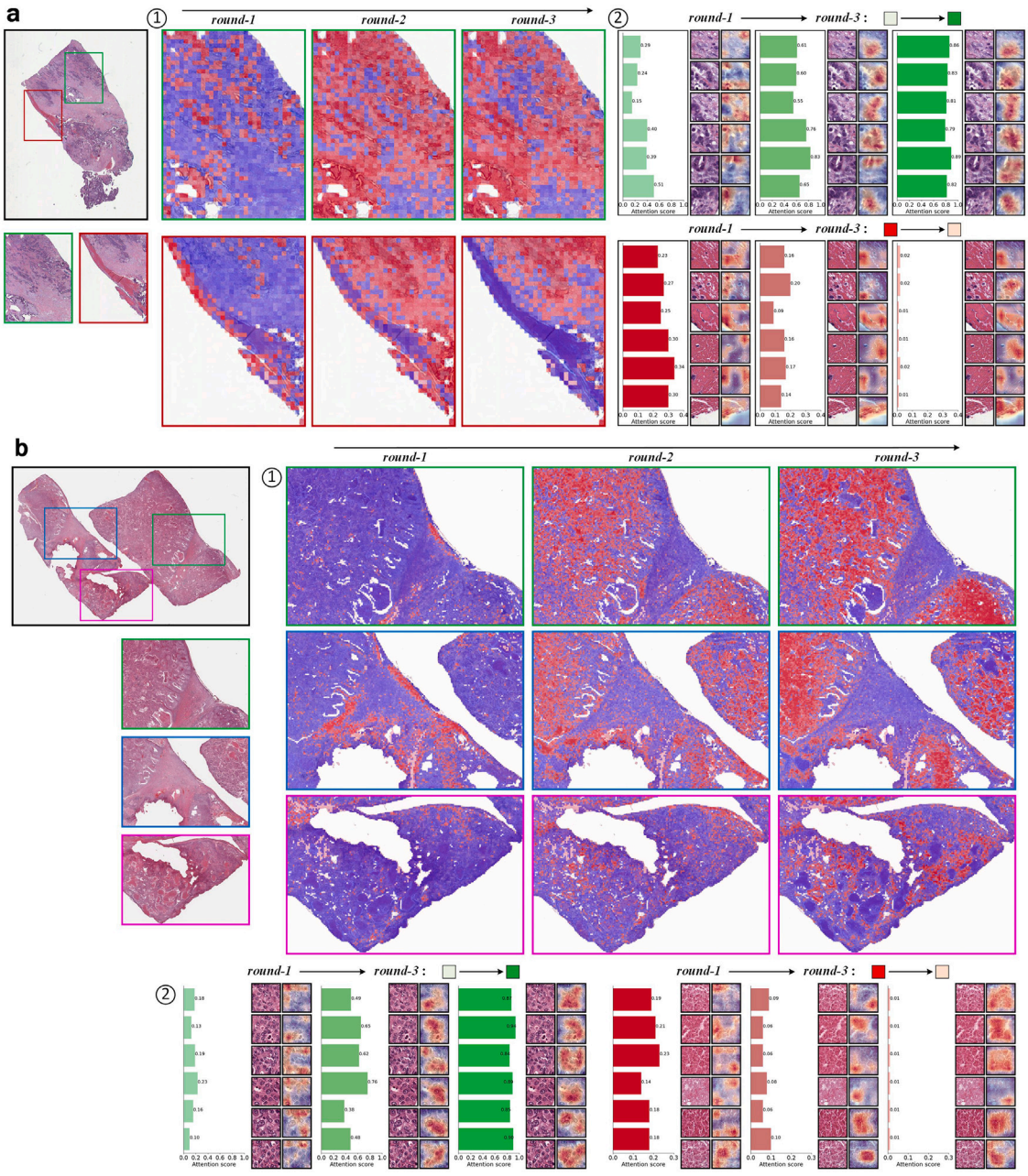

Fig. 4. Attention evolution of tile-level features after each self-interaction round. a, Example test case from the *OV-EMT* task. b, Example test case from the *COLU-KRAS* task. The images on the left show the location of the example regions in the original WSIs. ① shows the evolution of the attention heatmap in different slide regions. Colour transition from blue to red indicates a rise in attention and vice versa. ② examples of tiles highly informative to the morphological classification task (green histograms) and of low relevance to the task (red histograms). The histogram, tile image and its attention heatmap demonstrate the attention evolution of these regions. Here in both presented scenarios, attention scores increase over time for tiles representing densely nuclear regions, and decrease for tiles containing connective tissue.

reveal that after self-interactive training, the feature spaces of different subtypes become more separated. Furthermore, Figure 18 shows that in the feature spaces of Inter-MIL and adInter-MIL, high-attention tiles from a test slide primarily neighbour high-attention tiles from slides of the same molecular subtype. Conversely, in GatedAttPool-MIL, many neighbours of high-attention tiles originate from slides of different molecular subtypes.

4.4. Impact of key hyperparameters

As explained in Section 3, we need to determine how many representative tiles are selected for tile-level encoder fine-tuning in each round of interactive training. The key hyperparameter is k^1. Fig. 7 provides the comparison results for Inter-MIL on different hyperparameter settings of k^1, on tasks *OV-EMT* and *COLU-KRAS*, showing that there may be some fluctuations in performance on different hyperparameter settings. As we compared multiple values of the selected tile numbers: too many tiles lead the encoder to learn excessive noise, while too few tiles result in insufficient learning of fine-grained features. However, the Inter-MIL model still outperforms the baseline under different settings of k^1.

5. Discussion

In this paper, we introduce Inter-MIL to tackle the challenges of complex molecular trait analysis, especially when specific histological

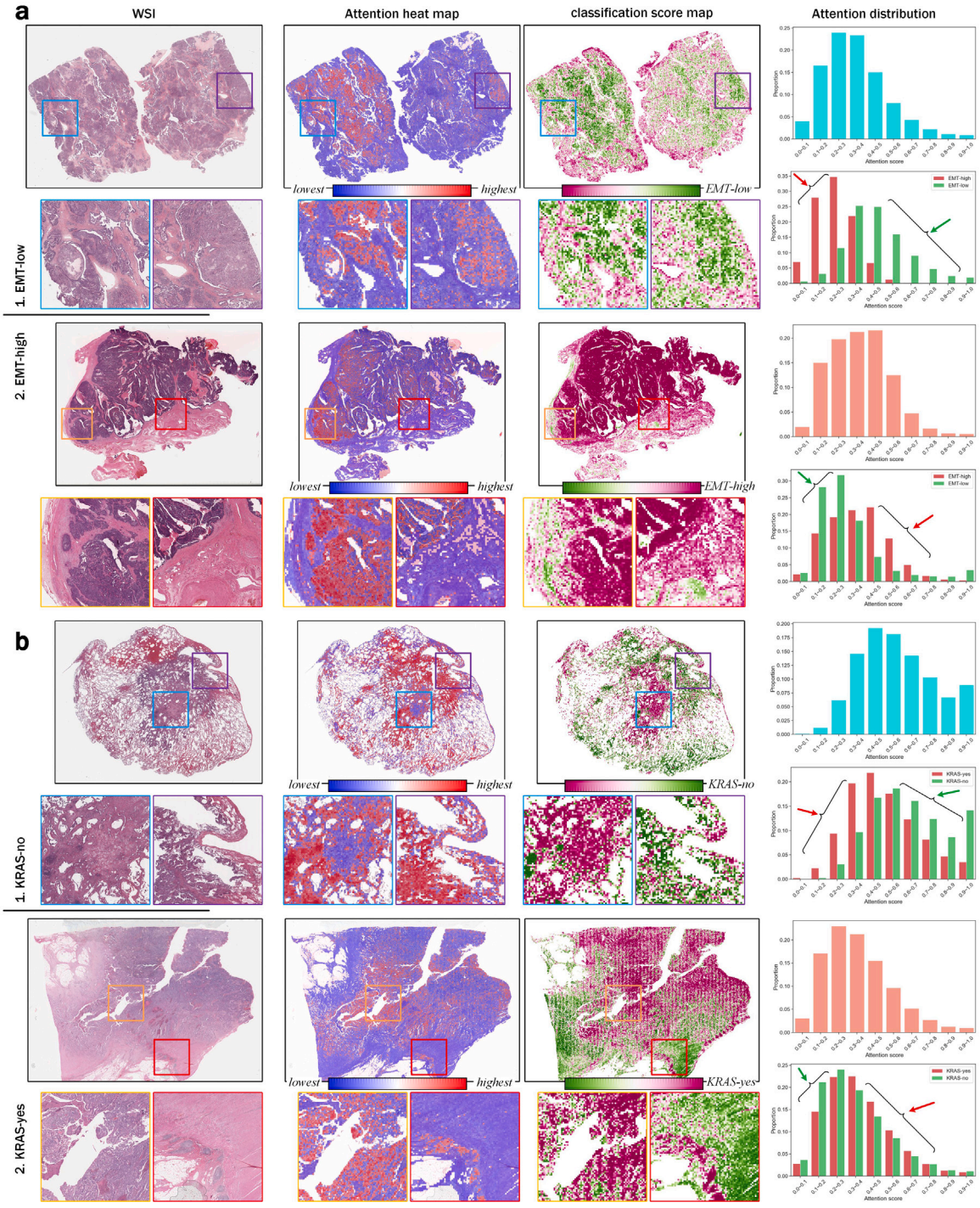

Fig. 5. Distributions of slide-level attention and classification scores in two morpho-molecular subtyping classification tasks. a, Example cases from the *OV-EMT* classification task, top: EMT-low case, bottom: EMT-high case. b, Example cases from the *COLU-KRAS* task, top: KRAS-no case, bottom: KRAS-yes case. For both a and b, from left to right: 1. the original WSI and the selected regions of interest; 2. attention heatmap; 3. classification score map; 4. Tile-attention histograms. Top: the proportion of tiles in the different attention ranges, bottom: proportions of tiles with prediction results of EMT-low/high (KRAS-no/yes). Here, we observe that in cases of different subtypes, tiles with higher attention obtain prediction scores that correspond more closely to their subtypes.

biomarkers for subtypes are not clearly defined. Inter-MIL simulates the pathologists' practice of frequently adjusting microscope magnification to capture both fine-grained and overall tissue features (Jaarsma et al., 2015), thereby enabling a seamless transition from subtle, fine-grained histology features to macro morphology features. Inter-MIL introduces an iterative knowledge interaction in weakly supervised learning for WSIs. It leverages global slide-level features as representative training material for tile-level encoding, enhancing the discriminative feature space and simplifying slide-level classification. Our results demonstrate the effectiveness of this interactive optimisation, showing improved model performance.

A potential issue with the use of the TCGA dataset in our experiments is associated with the lack of KRAS labels, which may bias the model. This is because, even though the reasons for missing labels are unknown, they are often not missing at random. E.g., an unknown KRAS status may be correlated to the complexity of the case and the

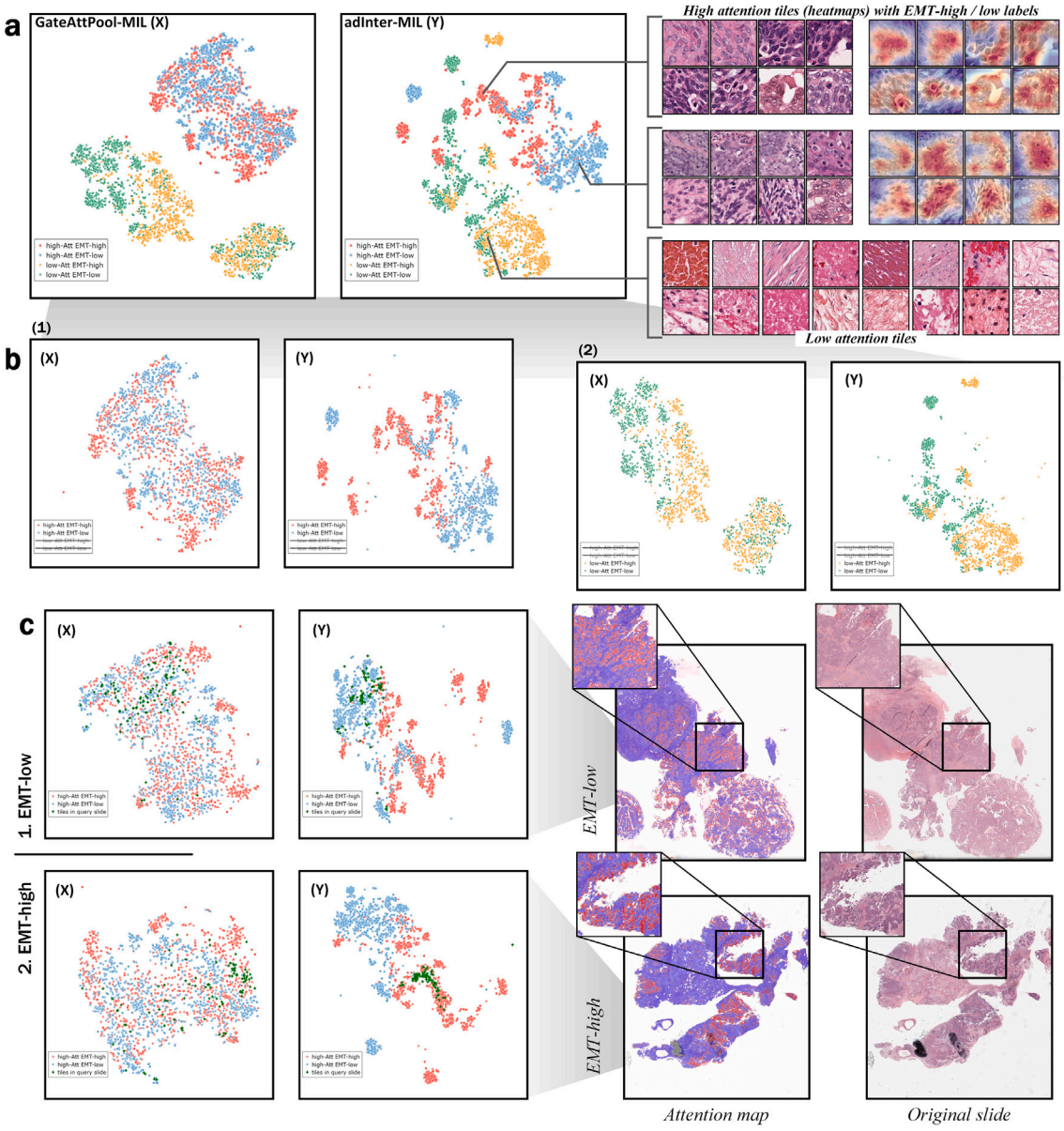

Fig. 6. A comparison of the feature spaces of all the representative tiles in the *OV-EMT* dataset for two of trained models. a, Left: the distribution of highly informative tiles and tiles with low task relevance in the learned cohort feature space, for the GatedAttPool-MIL (X) model and the adInter-MIL (Y) models; Right: example highly informative EMT-high/EMT-low tiles and examples of tiles without discriminative features. b, Comparison of the feature spaces of the (X) and (Y) models for highly informative tiles and tiles with low task relevance. c, The distributions of the highly informative tiles (green) taken from the two example cases: with EMT-low and EMT-high status respectively over the feature spaces of the two tested models. It can be seen that informative tiles form clearer, more separate clusters in the feature space of the adInter-MIL model. The tiles corresponding to the two example cases are located within the clusters corresponding to their correct label in the adInter-MIL feature space.

distribution of true KRAS status in these patients can be different from the overall distribution. Therefore, we carried out external validation to examine the generalisability of our method on unseen cohorts. Also, training and evaluating Inter-MIL's on our private cohort samples demonstrates its superior utilisation of new data.

Baseline methods typically rely on extensive training samples to compensate for the lack of optimisation of tile-level detail features (Zheng et al., 2022). However, even with extremely small training sets, Inter-MIL's self-interactive tile-level embedding optimisation achieves significant accuracy, demonstrating superior training efficiency compared to methods that necessitate larger datasets.

Inter-MIL also enhances model interpretability by providing reliable fine-grained features. By focusing on biologically informative regions and discarding noise, it enables more representative phenotype profiling. Re-optimisation of tile-level features yields slide-level embeddings enriched with fine-grained information, leading to more precise attention allocation and improved tile-level optimisation in subsequent rounds. This establishes a positive feedback loop, as evidenced by the clustering of fine-grained features in the latent space, Inter-MIL allows initially incorrect attention distributions to be rectified by using more optimal fine-grained feature representations learned from other slides. This self-correcting mechanism, absent in baseline models, optimises alignment for both coarse and fine-grained visual features, countering the impact of visual artefacts in small datasets.

Inter-MIL extends interpretability beyond attention, offering pseudo tile-level classification scores, which baseline methods cannot provide

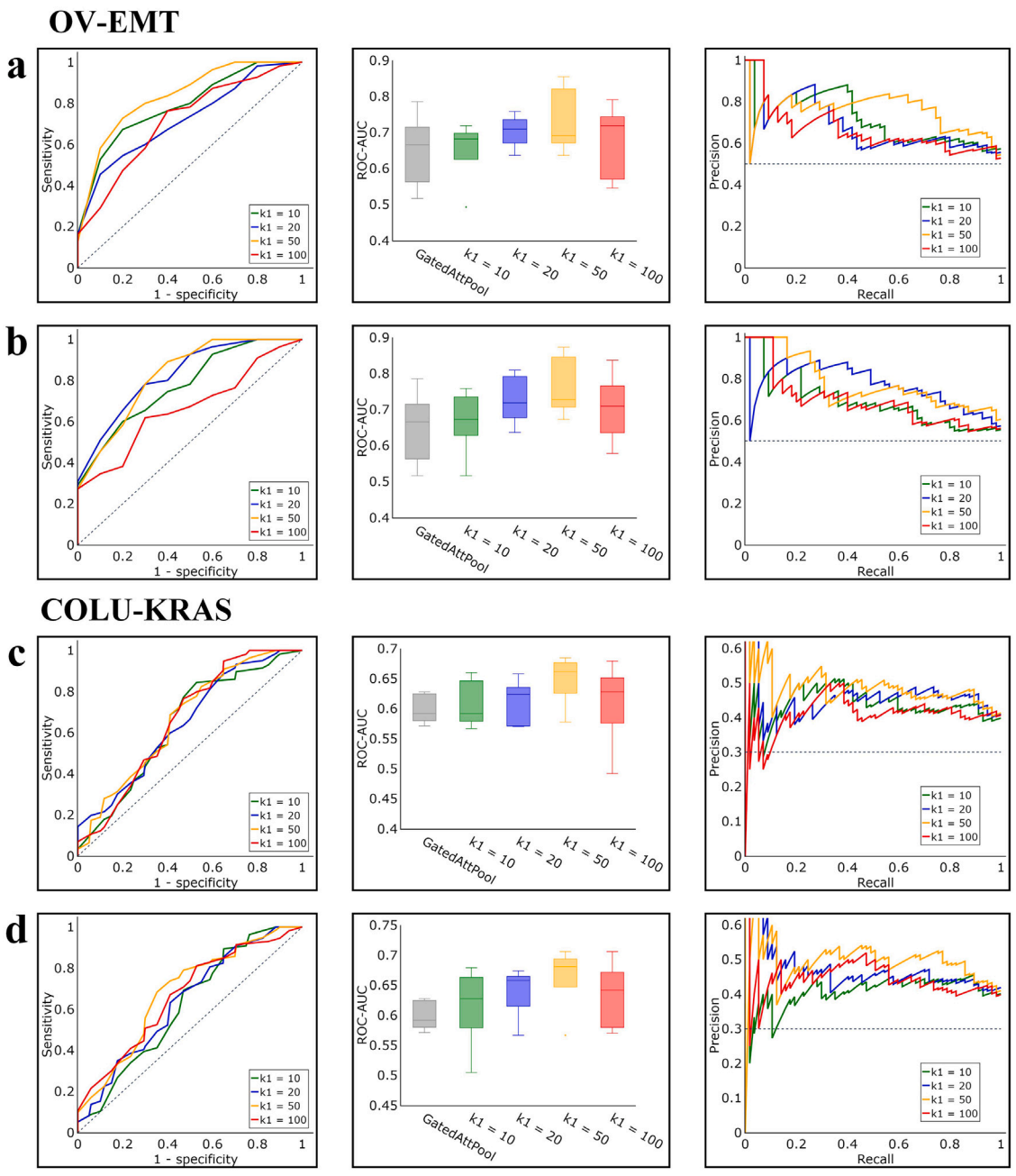

Fig. 7. Fluctuations in performance with various values of parameter k^1, which could be 10, 20, 50 (used), and 100. a and b refer to the methods of Inter-MIL and adInter-MIL, on the task of *OV-EMT*, while c and d show the performance of Inter-MIL and adInter-MIL, on the task of *COLU-KRAS*. For a, b, c, and d, the left shows the AUC-ROC curve under different values of k^1, the mid shows the result comparison of ROC-AUC, and the right shows the Precision–Recall Curve (PRC) under different values of k^1.

without detailed annotations. These scores clarify tile-level pseudo labels, moving beyond mere indicators of importance to actual class predictions, and also offer valuable reference information for researchers interested in the association between tile-level representations and subtypes.

Beyond the Inter-MIL framework, we introduce two auxiliary modules: (1) Adversarial optimisation for low-attention tiles, inspired by Ganin et al. (2016), enhancing model focus by reducing emphasis on non-critical regions, as evidenced by more concentrated attention. This module led to the development of adInter-MIL, a variant of Inter-MIL. The choice between Inter-MIL and adInter-MIL depends on data quality. For datasets with significant noise (e.g., contaminants, overexposure) or have not undergone thorough cleaning/normalisation of data, adInter-MIL effectively mitigates noise's impact on attention allocation. However, in rigorously quality-controlled datasets, adInter-MIL offers no substantial improvement over Inter-MIL; and (2) Contrastive pre-training for bag-of-tiles aggregators, inspired by He et al. (2020) and Chen et al. (2020), providing pre-trained parameters that enhance the aggregators' discriminative capabilities from the outset. While this accelerates aggregator convergence and boosts classification, it results in a sparse attention map due to selective tile sampling for contrasting training. Nonetheless, the potential of aggregator pre-training to enhance WSI analysis, especially when performing the pre-training on external fundamental tasks like tumour/benign classification (Tolkach et al., 2020) or training with unsupervised fashion like self-supervised learning (He et al., 2020), is promising.

The proposed Inter-MIL framework is scalable and adaptable, it can be adapted to various encoder and aggregator backbones, and the multi-task mode can be extended by setting multiple output heads in interactive training modules. Inter-MIL also has the potential to

accommodate a wider array of histopathological analyses including tumour classification (Dolezal et al., 2022), prognosis (Lu et al., 2020; Foersch et al., 2023), and therapy response prediction (Foersch et al., 2023), without necessitating pixel-level annotations. It also provides a biologically pertinent tile-level feature pool, offering more informative materials to enrich correlation analysis in multi-task and multi-modal studies (Lipkova et al., 2022a; Chen et al., 2022c).

One limitation of Inter-MIL is its slight sensitivity to hyperparameters, such as k^1, analysed in Section 4.4. Furthermore, Inter-MIL relies on the initial round of MIL to learn a preliminary attention distribution on the slides. If the initial round fails or concludes prematurely, it may result in suboptimal outcomes in subsequent self-interactive learning phases.

6. Conclusion

In summary, we introduce a novel weakly supervised MIL approach for predicting molecular subtypes from histological WSIs, utilising a self-interactive algorithm to bridge multi-scale histopathological features. This method facilitates the learning of highly discriminative features in latent space and enhances interpretability through improved visualisation outcomes.

Notably, Inter-MIL introduces a simple and efficient communication mechanism for features across different scales in scenarios with a small amount of data, an achievement not accomplished in other studies. Thus, Inter-MIL presents a viable solution to practical challenges such as datasets with a scant amount of cases and indeterminate biomarker locations. Moreover, Inter-MIL's design allows seamless integration with other models, enabling users to adopt any advanced deep learning architecture for encoders and aggregators or to leverage other pre-trained fundamental models.

Future efforts will aim to tackle existing technical hurdles, perform more robust uncertainty estimation, broaden Inter-MIL's applicability across varied tasks, and assess the tile-level feature pool's utility in diverse applications. Moreover, we intend to investigate Inter-MIL's integration with more DNN architectures like Graph Neural Networks (Lee et al., 2022) and Vision Transformers (Chen et al., 2022b; Azad et al., 2024b) to adeptly capture contextually rich spatial information on WSIs.

CRediT authorship contribution statement

Yang Hu: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Korsuk Sirinukunwattana:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Bin Li:** Writing – review & editing, Validation. **Kezia Gaitskell:** Validation, Investigation, Data curation. **Enric Domingo:** Validation, Resources, Data curation. **Willem Bonnaffé:** Writing – review & editing, Validation. **Ruby Wood:** Writing – review & editing, Validation. **Nasullah Khalid Alham:** Visualization, Software. **Stefano Malacrino:** Software. **Dan J Woodcock:** Writing – review & editing, Supervision. **Clare Verrill:** Supervision, Funding acquisition. **Ahmed Ahmed:** Resources, Investigation, Funding acquisition. **Jens Rittscher:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare that they have no competing financial interests.

Acknowledgements

KS, CV and JR – were supported by the PathLAKE Centre of Excellence for digital pathology and artificial intelligence which is funded by the Data to Early Diagnosis and Precision Medicine strand of the HM Government's Industrial Strategy Challenge Fund, managed and delivered by Innovate UK on behalf of UK Research and Innovation (UKRI) (Grant ref: 104689/application number 18181). CV - was supported by the NIHR Oxford Biomedical Research Center. Views expressed are those of the authors and not necessarily those of the PathLAKE Consortium members, the NHS, the UKRI, the NIHR, Innovate UK or the Department of Health.

JR is an adjunct professor of the Ludwig Oxford Branch.

We thank Professor Ian Mills and Oxford Prostate Cancer Biology Group for their guidance and suggestions on molecular biology for this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103437.

Data availability

The TCGA datasets and images analysed in this study are openly and publicly available at https://portal.gdc.cancer.gov/.

The FOCUS cohort was reviewed and approved as part of S:CORT consortium by the South Cambs Research Ethics committee (REC ref 15/EE/0241).

The S:CORT FOCUS cohort used in this study is available to all academic researchers on submission of a data request to the S:CORT data access committee. For commercial agencies, the data is in the process of being made available through Cancer Research Horizons acting on behalf of the funders and consortium members.

References

Abraham, J.P., Magee, D., Cremolini, C., Antoniotti, C., Halbert, D.D., Xiu, J., Stafford, P., Berry, D.A., Oberley, M.J., Shields, A.F., et al., 2021. Clinical validation of a machine-learning–derived signature predictive of outcomes from first-line oxaliplatin-based chemotherapy in advanced colorectal CancerAI analysis of molecular data to predict FOLFOX response. Clin. Cancer Res. 27 (4), 1174–1183.

Alsaafin, A., Safarpoor, A., Sikaroudi, M., Hipp, J.D., Tizhoosh, H., 2023. Learning to predict RNA sequence expressions from whole slide images with applications for search and classification. Commun. Biol. 6 (1), 304.

Aubreville, M., Stathonikos, N., Bertram, C.A., Klopfleisch, R., Ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T.A., Maier, A., et al., 2023. Mitosis domain generalization in histopathology images—the MIDOG challenge. Med. Image Anal. 84, 102699.

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D., 2024a. Advances in medical image analysis with vision Transformers: A comprehensive review. Med. Image Anal. 91, 103000. http://dx.doi.org/10.1016/j.media.2023.103000, URL https://www.sciencedirect.com/science/article/pii/S1361841523002608.

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D., 2024b. Advances in medical image analysis with vision Transformers: A comprehensive review. Med. Image Anal. 91, 103000.

Barker, J., Hoogi, A., Depeursinge, A., Rubin, D.L., 2016. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. Med. Image Anal. 30, 60–71.

Bilal, M., Raza, S.E.A., Azam, A., Graham, S., Ilyas, M., Cree, I.A., Snead, D., Minhas, F., Rajpoot, N.M., 2021. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. Lancet Digit. Health 3 (12), e763–e772.

Binder, A., Bockmayr, M., Hägele, M., Wienert, S., Heim, D., Hellweg, K., Ishii, M., Stenzinger, A., Hocke, A., Denkert, C., et al., 2021. Morphological and molecular breast cancer profiling through explainable machine learning. Nat. Mach. Intell. 3 (4), 355–366.

Brunt, E.M., 2010. Pathology of nonalcoholic fatty liver disease. Nat. Rev. Gastroenterol. Hepatol. 7 (4), 195–203.
Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Med. 25 (8), 1301–1309.
Cao, L., Wang, J., Zhang, Y., Rong, Z., Wang, M., Wang, L., Ji, J., Qian, Y., Zhang, L., Wu, H., et al., 2023. E2EFP-MIL: End-to-end and high-generalizability weakly supervised deep convolutional network for lung cancer classification from whole slide image. Med. Image Anal. 88, 102837.
Cen, X., Dong, W., Lv, W., Zhao, Y., Dubee, F., Mentis, A.-F.A., Jovic, D., Yang, H., Li, Y., 2024. Towards interpretable imaging genomics analysis: Methodological developments and applications. Inf. Fusion 102032.
Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022b. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: O'Conner, L. (Ed.), Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vol. 42600, IEEE, New Jersey, pp. 16144–16155.
Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (Eds.), International Conference on Machine Learning. Vol. 119, PMLR, Virtual, pp. 1597–1607.
Chen, H., Li, C., Wang, G., Li, X., Rahaman, M.M., Sun, H., Hu, W., Li, Y., Liu, W., Sun, C., et al., 2022a. GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection. Pattern Recognit. 130, 108827.
Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al., 2022c. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell 40 (8), 865–878.
Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A., 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Med. 24 (10), 1559–1567.
Couture, H.D., Williams, L.A., Geradts, J., Nyante, S.J., Butler, E.N., Marron, J., Perou, C.M., Troester, M.A., Niethammer, M., 2018. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer 4 (1), 30.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Flynn, P. (Ed.), 2009 IEEE Conference on Computer Vision and Pattern Recognition. Vol. 14067, IEEE, New Jersey, pp. 248–255.
Dolezal, J.M., Srisuwananukorn, A., Karpeyev, D., Ramesh, S., Kochanny, S., Cody, B., Mansfield, A.S., Rakshit, S., Bansal, R., Bois, M.C., et al., 2022. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. Nature Commun. 13 (1), 6572.
Foersch, S., Glasner, C., Woerl, A.-C., Eckstein, M., Wagner, D.-C., Schulz, S., Kellers, F., Fernandez, A., Tserea, K., Kloth, M., et al., 2023. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. Nature Med. 1–10.
Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17 (1), 1–35.
Gao, Z., Jia, C., Li, Y., Zhang, X., Hong, B., Wu, J., Gong, T., Wang, C., Meng, D., Zheng, Y., et al., 2022. Unsupervised representation learning for tissue segmentation in histopathological images: from global to local contrast. IEEE Trans. Med. Imaging 41 (12), 3611–3623.
Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., Vermeulen, L., Wang, X., 2019. DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis 8 (9), 44.
Gianni, L., Eiermann, W., Semiglazov, V., Manikhas, A., Lluch, A., Tjulandin, S., Zambetti, M., Vazquez, F., Byakhow, M., Lichinitser, M., et al., 2010. Neoadjuvant chemotherapy with trastuzumab followed by adjuvant trastuzumab versus neoadjuvant chemotherapy alone, in patients with HER2-positive locally advanced breast cancer (the NOAH trial): a randomised controlled superiority trial with a parallel HER2-negative cohort. Lancet 375 (9712), 377–384.
Godson, L., Alemi, N., Nsengimana, J., Cook, G.P., Clarke, E.L., Treanor, D., Bishop, D.T., Newton-Bishop, J., Gooya, A., Magee, D., 2024. Immune subtyping of melanoma whole slide images using multiple instance learning. Med. Image Anal. 103097.
Guo, R., Xie, K., Pagnucco, M., Song, Y., 2023. SAC-Net: Learning with weak and noisy labels in histopathology image segmentation. Med. Image Anal. 86, 102790.
Han, C., Lin, J., Mai, J., Wang, Y., Zhang, Q., Zhao, B., Chen, X., Pan, X., Shi, Z., Xu, Z., et al., 2022. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. Med. Image Anal. 80, 102487.
Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I., 2020. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In: O'Conner, L. (Ed.), Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vol. 42600, IEEE, New Jersey, pp. 3852–3861.
He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: O'Conner, L. (Ed.), Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vol. 42600, IEEE, New Jersey, pp. 9729–9738.
Hekselman, I., Yeger-Lotem, E., 2020. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. Nature Rev. Genet. 21 (3), 137–150.
Hong, R., Liu, W., DeLair, D., Razavian, N., Fenyö, D., 2021. Predicting endometrial cancer subtypes and molecular features from histopathology images using multi-resolution deep learning models. Cell Rep. Med. 2 (9), 100400.
Hoque, M.Z., Keskinarkaus, A., Nyberg, P., Seppänen, T., 2024. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. Inf. Fusion 101997.
Hu, Z., Artibani, M., Alsaadi, A., Wietek, N., Morotti, M., Shi, T., Zhong, Z., Gonzalez, L.S., El-Sahhar, S., KaramiNejadRanjbar, M., et al., 2020. The repertoire of serous ovarian cancer non-genetic heterogeneity revealed by single-cell sequencing of normal fallopian tube epithelial cells. Cancer Cell 37 (2), 226–242.
Hu, Z., Cunnea, P., Zhong, Z., Lu, H., Osagie, O.I., Campo, L., Artibani, M., Nixon, K., Ploski, J., Gonzalez, L.S., et al., 2021. The oxford classic links epithelial-to-mesenchymal transition to immunosuppression in poor prognosis ovarian cancers. Clin. Cancer Res. 27 (5), 1570–1579.
Huang, Z., Shao, W., Han, Z., Alkashash, A.M., De la Sancha, C., Parwani, A.V., Nitta, H., Hou, Y., Wang, T., Salama, P., et al., 2023. Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. NPJ Precis. Oncol. 7 (1), 14.
Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (Eds.), International Conference on Machine Learning. Vol. 80, PMLR, Stockholm, pp. 2127–2136.
Jaarsma, T., Jarodzka, H., Nap, M., van Merriënboer, J.J., Boshuizen, H., 2015. Expertise in clinical pathology: Combining the visual and cognitive perspective. Adv. Health Sci. Educ. 20 (4), 1089–1106.
Jain, M.S., Massoud, T.F., 2020. Predicting tumour mutational burden from histopathological images using multiscale deep learning. Nat. Mach. Intell. 2 (6), 356–362.
Jung, H.A., Lim, J., Choi, Y.-L., Lee, S.-H., Joung, J.-G., Jeon, Y.J., Choi, J.W., Shin, S., Cho, J.H., Kim, H.K., et al., 2022. Clinical, pathologic, and molecular prognostic factors in patients with early-stage EGFR-mutant NSCLC. Clin. Cancer Res. 28 (19), 4312–4321.
Kalra, S., Adnan, M., Hemati, S., Dehkharghanian, T., Rahnamayan, S., Tizhoosh, H., 2021. Pay attention with focus: A novel learning scheme for classification of whole slide images. In: de Bruijne, M. (Ed.), Medical Image Computing and Computer Assisted Intervention, MICCAI 2021. Vol. 12908, Springer, Strasbourg, pp. 350–359.
Kalra, S., Tizhoosh, H.R., Choi, C., Shah, S., Diamandis, P., Campbell, C.J., Pantanowitz, L., 2020. Yottixel–an image search engine for large archives of histopathology whole slide images. Med. Image Anal. 65, 101757.
Kapse, S., Das, S., Zhang, J., Gupta, R.R., Saltz, J., Samaras, D., Prasanna, P., 2024. Attention De-sparsification Matters: Inducing diversity in digital pathology representation learning. Med. Image Anal. 93, 103070.
Kassab, M., Jehanzaib, M., Başak, K., Demir, D., Keles, G.E., Turan, M., 2024. FFPE++: Improving the quality of formalin-fixed paraffin-embedded tissue imaging via contrastive unpaired image-to-image translation. Med. Image Anal. 91, 102992.
Kers, J., Bülow, R.D., Klinkhammer, B.M., Breimer, G.E., Fontana, F., Abiola, A.A., Hofstraat, R., Corthals, G.L., Peters-Sengers, H., Djudjaj, S., et al., 2022. Deep learning-based classification of kidney transplant pathology: a retrospective, multicentre, proof-of-concept study. Lancet Digit. Health 4 (1), e18–e26.
Krishnan, R., Rajpurkar, P., Topol, E.J., 2022. Self-supervised learning in medicine and healthcare. Nat. Biomed. Eng. 1–7.
Lee, Y., Park, J.H., Oh, S., Shin, K., Sun, J., Jung, M., Lee, C., Kim, H., Chung, J.-H., Moon, K.C., et al., 2022. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. Nat. Biomed. Eng. 1–15.
Li, T., Kung, H.-J., Mack, P.C., Gandara, D.R., 2013. Genotyping and genomic profiling of non–small-cell lung cancer: implications for current and future therapies. J. Clin. Oncol. 31 (8), 1039.
Li, B., Li, Y., Eliceiri, K.W., 2021a. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: O'Conner., L. (Ed.), Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vol. 46437, IEEE, New Jersey, pp. 14318–14328.
Li, W., Li, J., Polson, J., Wang, Z., Speier, W., Arnold, C., 2022. High resolution histopathology image generation and segmentation through adversarial training. Med. Image Anal. 75, 102251.
Li, J., Li, W., Sisk, A., Ye, H., Wallace, W.D., Speier, W., Arnold, C.W., 2021c. A multi-resolution model for histopathology image classification and localization with multiple instance learning. Comput. Biol. Med. 131, 104253.
Li, K., Qian, Z., Han, Y., Eric, I., Chang, C., Wei, B., Lai, M., Liao, J., Fan, Y., Xu, Y., 2023. Weakly supervised histopathology image segmentation with self-attention. Med. Image Anal. 86, 102791.
Li, H., Yang, F., Zhao, Y., Xing, X., Zhang, J., Gao, M., Huang, J., Wang, L., Yao, J., 2021b. DT-MIL: Deformable transformer for multi-instance learning on histopathological image. In: de Bruijne, M. (Ed.), International Conference on Medical Image Computing and Computer-Assisted Intervention. Vol. 12908, Springer, Cham, pp. 206–216.

Lievre, A., Bachet, J.-B., Le Corre, D., Boige, V., Landi, B., Emile, J.-F., Côté, J.-F., Tomasic, G., Penna, C., Ducreux, M., et al., 2006. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Res. 66 (8), 3992–3995.

Lipkova, J., Chen, R.J., Chen, B., Lu, M.Y., Barbieri, M., Shao, D., Vaidya, A.J., Chen, C., Zhuang, L., Williamson, D.F., et al., 2022a. Artificial intelligence for multimodal data integration in oncology. Cancer Cell 40 (10), 1095–1110.

Lipkova, J., Chen, T.Y., Lu, M.Y., Chen, R.J., Shady, M., Williams, M., Wang, J., Noor, Z., Mitchell, R.N., Turan, M., et al., 2022b. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. Nature Med. 28 (3), 575–582.

Lu, C., Bera, K., Wang, X., Prasanna, P., Xu, J., Janowczyk, A., Beig, N., Yang, M., Fu, P., Lewis, J., et al., 2020. A prognostic model for overall survival of patients with early-stage non-small cell lung cancer: a multicentre, retrospective study. Lancet Digit. Health 2 (11), e594–e606.

Lu, M.Y., Chen, T.Y., Williamson, D.F., Zhao, M., Shady, M., Lipkova, J., Mahmood, F., 2021a. AI-based pathology predicts origins for cancers of unknown primary. Nature 594 (7861), 106–110.

Lu, W., Toss, M., Dawood, M., Rakha, E., Rajpoot, N., Minhas, F., 2022. Slidegraph+: whole slide image level graphs to predict her2 status in breast cancer. Med. Image Anal. 80, 102486.

Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021b. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5 (6), 555–570.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Malla, S.B., Fisher, D.J., Domingo, E., Blake, A., Hassanieh, S., Redmond, K.L., Richman, S.D., Youdell, M., Walker, S.M., Logan, G.E., et al., 2021. In-depth clinical and biological exploration of DNA damage immune response as a biomarker for oxaliplatin use in colorectal cancer. Clin. Cancer Res. 27 (1), 288–300.

Niehues, J.M., Quirke, P., West, N.P., Grabsch, H.I., van Treeck, M., Schirris, Y., Veldhuizen, G.P., Hutchins, G.G., Richman, S.D., Foersch, S., et al., 2023. Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: A retrospective multi-centric study. Cell Rep. Med. 4 (4).

Niyas, S., Bygari, R., Naik, R., Viswanath, B., Ugwekar, D., Mathew, T., Kavya, J., Kini, J.R., Rajan, J., 2023. Automated molecular subtyping of breast carcinoma using deep learning techniques. IEEE J. Transl. Eng. Health Med. 11, 161–169.

Oner, M.U., Kye-Jet, J.M.S., Lee, H.K., Sung, W.-K., 2023. Distribution based MIL pooling filters: Experiments on a lymph node metastases dataset. Med. Image Anal. 87, 102813.

Pan, X., Cheng, J., Hou, F., Lan, R., Lu, C., Li, L., Feng, Z., Wang, H., Liang, C., Liu, Z., et al., 2023. SMILE: Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations. Med. Image Anal. 102867.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: O'Conner, L. (Ed.), Proceedings of the IEEE International Conference on Computer Vision. Vol. 35066, IEEE, New Jersey, pp. 618–626.

Seymour, M.T., Maughan, T.S., Ledermann, J.A., Topham, C., James, R., Gwyther, S.J., Smith, D.B., Shepherd, S., Maraveyas, A., Ferry, D.R., et al., 2007. Different strategies of sequential and combination chemotherapy for patients with poor prognosis advanced colorectal cancer (MRC FOCUS): a randomised controlled trial. Lancet 370 (9582), 143–152.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y., 2021. TransMIL: Transformer based correlated multiple instance learning for whole slide image classication. Adv. Neural Inf. Process. Syst. 34.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., Waterston, R.H., 2017. DNA sequencing at 40: past, present and future. Nature 550 (7676), 345–353.

Sirinukunwattana, K., Domingo, E., Richman, S.D., Redmond, K.L., Blake, A., Verrill, C., Leedham, S.J., Chatzipli, A., Hardy, C., Whalley, C.M., et al., 2021. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. Gut 70 (3), 544–554.

Springenberg, M., Frommholz, A., Wenzel, M., Weicken, E., Ma, J., Strodthoff, N., 2023. From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. Med. Image Anal. 87, 102809.

Tolkach, Y., Dohmgörgen, T., Toma, M., Kristiansen, G., 2020. High-accuracy prostate cancer pathology using deep learning. Nat. Mach. Intell. 2 (7), 411–418.

Tomczak, K., Czerwińska, P., Wiznerowicz, M., 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. 19 (1A), A68.

Tomita, N., Tafe, L.J., Suriawinata, A.A., Tsongalis, G.J., Nasir-Moin, M., Dragnev, K., Hassanpour, S., 2022. Predicting oncogene mutations of lung cancer using deep learning and histopathologic features on whole-slide images. Transl. Oncol. 24, 101494.

Tsai, P.-C., Lee, T.-H., Kuo, K.-C., Su, F.-Y., Lee, T.-L.M., Marostica, E., Ugai, T., Zhao, M., Lau, M.C., Väyrynen, J.P., et al., 2023. Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. Nature Commun. 14 (1), 2102.

Uegami, W., Bychkov, A., Ozasa, M., Uehara, K., Kataoka, K., Johkoh, T., Kondoh, Y., Sakanashi, H., Fukuoka, J., 2022. MIXTURE of human expertise and deep learning—Developing an explainable model for predicting pathological diagnosis and survival in patients with interstitial lung disease. Mod. Pathol. 1–9.

Xing, X., Zhu, M., Chen, Z., Yuan, Y., 2024. Comprehensive learning and adaptive teaching: Distilling multi-modal knowledge for pathological glioma grading. Med. Image Anal. 91, 102990.

Yan, R., Shen, Y., Zhang, X., Xu, P., Wang, J., Li, J., Ren, F., Ye, D., Zhou, S.K., 2023. Histopathological bladder cancer gene mutation prediction with hierarchical deep multiple-instance learning. Med. Image Anal. 87, 102824.

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., Gao, S., Yuan, X., Tian, G., Liang, Y., et al., 2022. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. Comput. Struct. Biotechnol. J. 20, 333–342.

Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y., 2022. DTFD-MIL: Double-Tier feature distillation multiple instance learning for histopathology whole slide image classification. In: O'Conner, L. (Ed.), Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vol. 52688, IEEE, New Jersey, pp. 18802–18812.

Zhao, B., Deng, W., Li, Z.H.H., Zhou, C., Gao, Z., Wang, G., Li, X., 2024. LESS: Label-efficient multi-scale learning for cytological whole slide image screening. Med. Image Anal. 103109.

Zhao, Y., Lin, Z., Sun, K., Zhang, Y., Huang, J., Wang, L., Yao, J., 2022. SETMIL: spatial encoding transformer-based multiple instance learning for pathological image analysis. In: Wang., L. (Ed.), International Conference on Medical Image Computing and Computer-Assisted Intervention. Vol. 13436, Springer, Singapore, pp. 66–76.

Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., et al., 2020. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In: O'Conner., L. (Ed.), Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vol. 52688, IEEE, New Jersey, pp. 4837–4846.

Zheng, X., Wang, R., Zhang, X., Sun, Y., Zhang, H., Zhao, Z., Zheng, Y., Luo, J., Zhang, J., Wu, H., et al., 2022. A deep learning model and human-machine fusion for prediction of EBV-associated gastric cancer from histopathology. Nature Commun. 13 (1), 2790.