# Deep Learning Framework for Classifying Whole-Slide Multiplex Immunofluorescence Images to Predict Immunotherapy Response in Melanoma Patients

Priyanka Rana, Tuba N Gide, Nurudeen A Adegoke, Yizhe Mao, Shlomo Berkovsky, Enrico Coiera, James S Wilmott, and Sidong Liu

***Abstract*—Immunotherapy has emerged as a prominent approach in melanoma treatment, however a substantial number of patients do not respond effectively. This highlights the critical need to accurately predict immunotherapy responses for designing personalised treatment strategies. Current practices rely predominantly on clinical data and the expertise of oncologists; however, a deeper understanding of molecular interactions through tissue-based biomarkers offers a promising avenue for advancement. Whole-slide multiplex immunofluorescence (mIF) images enable detailed analysis of cells/tissues in their microenvironment, deepening insights into disease mechanisms. However, numerous channels, an extensive image size, and spatially dispersed information of mIF images pose analytical challenges, requiring advanced techniques to effectively learn these intricate features for optimal performance. In this study, we introduce a novel deep-learning framework, Channel Optimisation with Multi-Instance Learning (COMIL), specifically designed to classify whole-slide mIF images for predicting immunotherapy response in melanoma patients. The study demonstrates that a feature extraction method that models inter-channel relationships and captures complex interdependencies among multiple channels of mIF images enhances classification performance. Additionally, incorporating this method within an MIL framework, optimised at both the slide and instance levels, further improves the classification performance of whole-slide mIF images. We evaluated COMIL on mIF images from the Melanoma Institute Australia, where it outperformed baseline methods with substantial improvements on both the internal (AUC increase by 28%) and external test sets (AUC increase by 20%), underscoring its potential for predicting immunotherapy response.**

***Index Terms*—Multiplex immunofluorescence images, multichannel images, melanoma, immunotherapy response.**

Received 15 December 2024; revised 25 July 2025 and 18 September 2025; accepted 14 October 2025. Date of publication 24 November 2025; date of current version 5 October 2026. The work of Tuba N Gide was supported by CINSW Early Career Fellowship under Grant 2020/ECF1244. This work was supported by the National Health and Medical Research (NHMRC), Centre for Research Excellence in Digital Health, and NHMRC Investigator under Grant GNT2008645. The data generation was supported in part by NHMRC Fellowship under Grant JSW APP1174325, in part by Melanoma Research Alliance young investigator Fellowship under Grant 700455, in part by Cancer Council NSW Project under Grant RG19-15 CINSW, and in part by Translational Program under Grant 2021/TPG2114. *(Corresponding author: Priyanka Rana.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Sydney Local Health District Human Research Ethics Committee under Ethics Board Protocol No. X17-0312 & 2019/ETH07604 and Reference No. 2020/ETH00426 & X20-0086, and performed in line with the Personalised Immunotherapy Platform (PIP-PREDICT).

Priyanka Rana, Shlomo Berkovsky, Enrico Coiera, and Sidong Liu are with the Centre for Health Informatics, Macquarie University, Sydney, NSW 2113, Australia (e-mail: priyanka.rana@mq.edu.au; shlomo.berkovsky@mq.edu.au; enrico.coiera@mq.edu.au; sidong.liu@mq.edu.au).

Tuba N Gide, Nurudeen A Adegoke, Yizhe Mao, and James S Wilmott are with the Melanoma Institute Australia, The University of Sydney, Sydney, NSW 2050, Australia, and also with the Charles Perkins Centre, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW 2050, Australia (e-mail: tuba.nurgide@melanoma.org.au; nurudeen.adegoke@melanoma.org.au; yizhe.mao@melanoma.org.au; james.wilmott@sydney.edu.au).

The code is available at https://github.com/priyarana/COMIL.

Digital Object Identifier 10.1109/JBHI.2025.3623576

## I. Introduction

In 2020, the global incidence of melanoma, a malignant tumour, exceeded 300,000 new cases in a single year, indicating a pressing concern as rates have been steadily increasing for several decades [1]. Immunotherapy, aiming to activate the immune system of the body to combat cancer, has emerged as a groundbreaking approach to the treatment of melanoma [2]. Immunotherapy drugs such as programmed cell death-1 (PD-1) and programmed cell death ligand-1 (PD-L1) inhibitors have shown promising outcomes in improving survival rates [3]. However, despite these advancements, a considerable proportion of patients do not respond to immunotherapy, which carries serious risks, including the potential for fatal reactions in some patients with melanoma undergoing treatment. This underscores the need to predict the immunotherapy response in melanoma patients, which is crucial to tailoring personalised treatment strategies that enhance therapeutic efficacy and minimise unnecessary side effects.

Current practice relies solely on clinical data or the expertise of oncologists which does not fully account for the complex biological variations present in individual patients. Predictive biomarkers (a biological molecule that is a sign of a

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

normal or abnormal process, or condition/disease) play a pivotal role in informing therapy decisions for individual patients. A comprehensive understanding of their molecular interactions, particularly through tissue-based biomarkers that assess multiple cell types, their spatial interactions, and distinct expression patterns of immunoregulatory molecules simultaneously, is essential for the accurate selection of patients who are likely to respond to immunotherapy [4].

Recent advancements in imaging technologies have enabled the visualisation of cellular and molecular interactions, helping enhance understanding of complex cellular dynamics. In particular, the emergence of microscopic multiplex immunofluorescence (mIF) images with abundant information across numerous channels has offered strong opportunities [5]. In a recent study, mIF images showed better diagnostic accuracy than other tissue imaging modalities when predicting responses to anti-PD-1 or anti-PD-L1 treatment [6], [7]. This is because mIF images enable simultaneous analysis of multiple biomarkers and their interactions within a single tissue section of the tumour microenvironment. This provides a deeper understanding of the underlying mechanisms, thereby allowing for more precise and personalised treatment decisions.

Whole-slide mIF images are characterised by their large size, scattered spatial information, and substantial number of image channels accommodating multiple biomarkers. This complexity often results in significant heterogeneity within an image. Therefore, their analysis requires advanced techniques to effectively learn both global and local images features for optimal performance. Recently, the classification of these images has gained increasing attention due to their promising applications in prognostic and diagnostic studies in the field of cancer research [6], [8]. These studies of mIF image classification mainly focus on selected regions (e.g., with higher tumour cell density) and channels (e.g. specific molecular biomarkers), and utilise handcrafted features based on marker intensity quantification. These features are modelled with subjective criteria defined by human experts. While handcrafted features can offer interpretability, they may be limited in capturing complex, non-linear patterns and relationships present in the data. This may result in underperformance in clinical settings or when evaluated on datasets from diverse populations or over extended periods.

Advanced deep learning methods can capture complex patterns within data, improving performance, reducing bias and enhancing the model's generalisability. However, the large size of whole-slide images often leads to challenges in processing and effective analysis, requiring specialised techniques for handling such extensive images. Therefore, Multiple Instance Learning (MIL) [9] which analyses images by splitting them into patches/instances of smaller dimensions is a preferred approach. With ground truth known only for the whole image, MIL is a weakly supervised method that considers image as a bag with all instances inheriting the image label. Accordingly, slide-level labels are assigned to every instance treating them as separate training examples; thus, increasing the number of labeled data points. However, this approach might not necessarily improve performance due to the potential introduction of noisy labels. This poses a significant challenge in developing effective machine learning classifiers for computational pathology. Thus, feature extraction and model optimisation become critical during training when detailed instance-level labels are unavailable. To accomplish this, various frameworks have been developed that can be categorised into three main types, classical MIL [9], Attention-based MIL (AttMIL) [10], and Clustering constrained Attention MIL (CLAM) [11].

The primary distinction between these frameworks lies in their approach in integrating instance-level features or predictions to generate slide-level features or predictions. The classical MIL [9] framework employs a max-pooling layer to derive the final slide-level prediction by selecting the instance with the highest predicted probability. However, this method may lead to suboptimal performance if the selected instance does not adequately represent the entire slide, particularly since slides can often have sparsely distributed or overlapping inter-class information. AttMIL [10] improves upon the classical MIL approach by incorporating an attention mechanism. This mechanism computes attention scores for each instance reflecting their relative importance and computes weighted-average of instance-level embeddings to obtain slide-level embedding. This approach enhances the model's ability to leverage information from multiple instances, rather than relying on a single instance with the highest probability. CLAM [11] builds on the AttMIL framework that employs attention mechanisms to identify subregions with high diagnostic value for precise whole-slide classification. Additionally, it utilises instance-level clustering on these identified representative regions to refine and constrain the feature space.

Another notable MIL framework based on the attention mechanism is Double–Tier Feature Distillation MIL (DTFD-MIL) [12] which utilises a two-tier method where the first tier applies attention mechanism and feature distillation strategies to randomly split instances in a bag into smaller bags (pseudo-bags). The second tier refines these feature representations at slide-level for classification. In contrast, Transformer-based Correlated MIL (TransMIL) [13] is a transformer-based approach that uses a self-attention mechanism to model both morphological and spatial correlations between instances, thereby enhancing the representation of complex patterns within whole-slide images.

A drawback inherent in existing MIL frameworks is that their optimisation approach is primarily focussed on slide-level training, with insufficient emphasis on instance-level features. This overlooks the variations in the importance of slide-level and local instance-level features during training, which may lead to suboptimal performance. Additionally, the current MIL frameworks are mainly designed for Hematoxylin and Eosin (H&E) stained or single channel immunohistochemistry slides, making them inadequate for the mIF image classification, due to the inherent complexity of the multiplex data with multiple channels. This is also highlighted by Lu et al. [6], showing that although mIF images can label multiple biomarkers of the tumour cells and immune elements, characterising the co-expression of these molecules in tissue microenvironment remains challenging. Existing frameworks are primarily designed to handle

only a single marker or standard three-channel images, thereby restricting their applicability to complex mIF datasets.

Accordingly, to fully realise the potential of mIF images within the MIL framework, it is essential to develop specialised channel optimisation techniques that can accommodate arbitrary number of channels and capture inherent inter-channel relationships. The current channel optimisation methods focus on convolutional channels within deep convolutional neural networks (CNNs) to enhance feature representation for standard RGB images. Squeeze-and-Excitation Networks (SE-Nets) [14] utilise global embedding information to capture channel relationships and modulate feature maps at the channel level using two fully connected (FC) layers. However, the parameter complexity in the FC layers makes it challenging to analyse interactions among channels across different layers limiting its ability to capture the contextual information [14]. Accordingly, Gated Channel Transformation (GCT) is proposed to enhance the integration and modelling of global information across convolutional channels [15].

While existing methods focus on optimising convolutional channels to enhance feature extraction in standard three-channel images, no methods currently exist for optimising image channels (enhancing the utilisation and extraction of informative features from individual channels) in whole-slide mIF images. As a result, the demand for advanced methodologies in whole-slide mIF image classification remains unmet. To address this gap, this study introduces Channel Optimisation integrated with Multi-Instance Learning (COMIL), an MIL framework specifically designed for whole-slide mIF image classification. COMIL addresses the challenge of extracting features from multiple mIF image channels and facilitates optimal training.

The proposed framework, COMIL innovates in two aspects compared to current MIL frameworks. Firstly, COMIL adapts the GCT [15] for feature extraction from mIF images, enabling it to accommodate an arbitrary number of biomarkers, thereby extending its applicability beyond conventional RGB images. Unlike the original deployment of GCT at the input stage before the convolutional layer, our approach deploys GCT after extracting pretrained features from each image channel. The GCT module transforms the pretrained features into global context embedding, capturing global information for each image channel. The obtained features are then normalised and subsequently gating adaptation is applied that employs a gating mechanism. Gating mechanisms have been effectively implemented to regulate the flow of information into, out of, and within the network [16]. Similarly, the utilised gating mechanism in COMIL dynamically modulates the contributions of all channels, enabling both competition and cooperation among them, thereby enhancing the network's ability to capture channel characteristics.

Secondly, during training, instead of relying on the conventional static weighting approach to integrate slide-level and instance-level losses, which predominantly emphasises slide-level losses, COMIL employs an adaptive loss (ADL) function for a dynamic and effective combination of both. ADL enables the model to prioritise the higher loss value during training, enabling more targeted model optimisation compared to static weighting methods. This approach improves the model's capacity to learn both slide-level global and fine-grained local patterns equitably.

We demonstrate the effectiveness of COMIL in predicting immunotherapy responses in melanoma patients and in identifying those unlikely to respond to treatment. Trained on an mIF image dataset from the Melanoma Institute Australia (MIA), consisting of seven image channels, with six channels representing protein biomarkers and one channel for autofluorescence, the proposed framework was evaluated on both internal and external test sets and significantly outperformed all compared methods for whole-slide mIF image classification.

## II. Methods

COMIL is specifically designed for mIF image classification and is built upon the MIL framework. The first step involves the segmentation of the tissue regions, followed by cropping instances or patches from within the segmented regions, each with dimensions of $224 \times 224 \times 7$ pixels. Each instance is then embedded into a 2,048-dimensional feature vector using a ResNet-101 backbone as the feature extractor. During training, these features undergo further refinement through GCT transformation, an attention mechanism, and instance-level clustering, optimising both image- and instance-level losses to provide final predictions.

### A. Dataset

A total of 211 labelled whole-slide mIF images of melanoma tissues are obtained from MIA for training. To evaluate the model under a realistic distribution shift, 42 additional images stained with a different biomarker sequence were acquired. Variations in staining order can alter signal intensity, marker expression, and spatial localisation, thereby challenging model generalisability. The image size is up to $24,000 \times 24,000$ pixels with a spatial resolution of 0.5 $\mu$m per pixel. Each image is presented with six protein biomarkers (CD8, CD16, CD68, DAPI, PD-L1 and SOX10) alongside an additional autofluorescence (AF) channel to identify various proteins within cells (Fig. 1(a)). All patients included in this dataset underwent immunotherapy, specifically anti-PD1+-CTLA-4, and classified as non-responsive (91 training and 19 test samples, positive class) or responsive (120 training and 23 test samples; negative class) to therapy based on progression-free survival status taken at 6 months following treatment. Biopsies were collected mainly within one year before treatment (89%), with a limited number taken more than a year before therapy. The details of image acquisition are available in [7], [17], and the input data patterns are described in the Appendix, Section A.

Informed consent was obtained from all participants and the study was conducted according to approved ethical guidelines. Ethical approval was granted by the Sydney Local Health District Human Research Ethics Committee. Ethics board protocol number: X17-0312 & 2019/ETH07604. Reference number 2020/ETH00426 & X20-0086. Clinical trial study name: Personalised Immunotherapy Platform (PIP-PREDICT). ClinicalTrials.gov ID: NCT06536257.

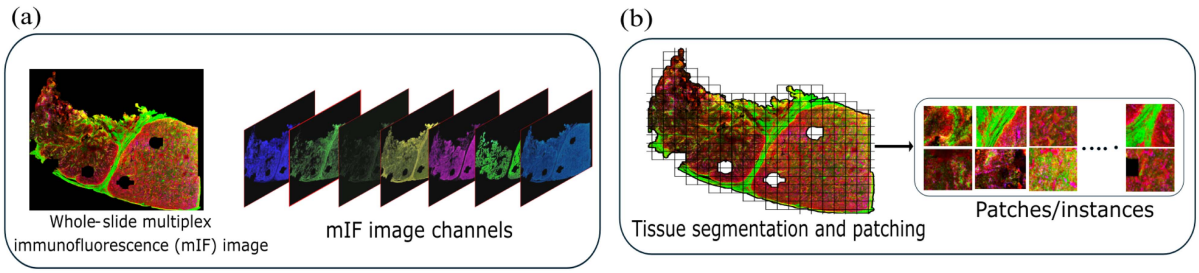



Fig. 1. (a) Each mIF image includes six protein biomarkers (CD8, CD16, CD68, DAPI, PD-L1 and SOX10) along with an additional autofluorescence channel. (b) The CD8, PD-L1 and DAPI biomarker images are used to generate a mask via Otsu's method, and the segmented tissue region is further divided into multiple $224 \times 224$ patches.

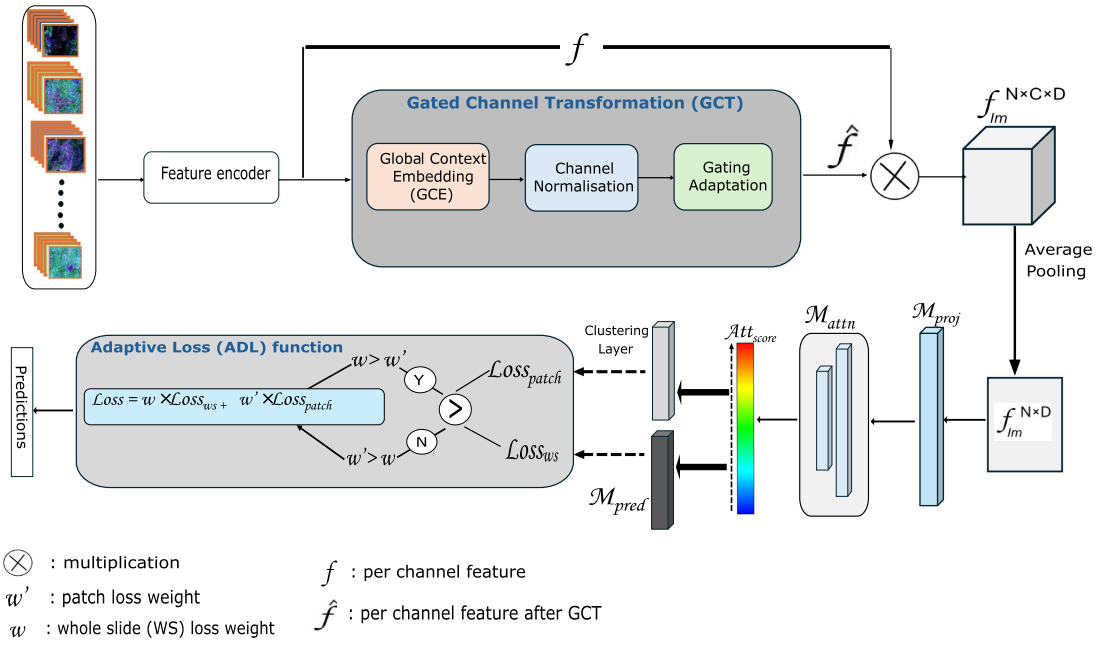



Fig. 2. Conceptual overview of the COMIL framework.

## B. Segmentation

As a first step, segmentation of the tissue from potentially noisy images is performed (Fig. 1(b)). It was observed that pixel distribution in CD8 and PD-L1 is comparatively more uniform and denser, resulting in a precise delineation of the object. Therefore, only CD8 and PD-L1 images, along with DAPI, were utilised to generate the segmentation mask through adaptive threshold with Otsu's method [18]. The grayscale images corresponding to the CD8, PD-L1, and DAPI channels were individually processed, with the Otsu method applied to each channel to determine an optimal threshold to separate signal-positive pixels from the background. The masks resulting from the CD8 and PD-L1 channels were used to approximate areas rich in immune cells, while the DAPI channel, which highlights nuclei, provided a structural reference for overall tissue architecture. The final tissue mask was obtained by overlapping the individual masks of CD8, PD-L1, and DAPI. This mask is used to segment the tissue regions in each channel, which are then combined to form the mIF image.

To construct MIL bags, the segmented tissue region is divided into multiple $224 \times 224$ patches, with each patch representing an instance. Features are then extracted from each instance and subsequently combined to represent the entire image.

## C. Feature Extraction

COMIL is designed to handle multiple channels of mIF images for feature extraction (Fig. 2). Within the MIL framework, each whole-slide mIF image is represented as a bag of $N$ instances. Instance-level features are extracted from each biomarker channel by combining it with DAPI and AF. Specifically, the input combinations now include CD8+DAPI+AF, CD16+DAPI+AF, CD68+DAPI+AF, PD-L1+DAPI+AF, and SOX10+DAPI+AF. This approach ensures that each biomarker is interpreted within a spatially and morphologically meaningful context, an essential aspect of mIF analysis. Furthermore, this design maintains compatibility with standard three-channel input requirements, enabling feature extraction using a ResNet-101 model pretrained on ImageNet. In addition to

the biomarker-combined inputs, we also include separate representations for DAPI and AF by replicating each channel twice to form valid three-channel inputs. This allows the model to capture biomarker-independent, morphology-rich features such as nuclear shape, cell density, and tissue architecture information that may otherwise be underrepresented when DAPI and AF are only used alongside strong biomarker signals.

These pretrained features from all channels are then aggregated (stacked) to represent the entire image, which is then processed through the GCT module. This unit captures inter-channel dependencies and refines the features by effectively modelling the relationships among channels. Following that, a spatial attention mechanism is applied to further enhance feature extraction, and subsequently instance-level clustering is performed to obtain the final predictions. In the following sections, we provide details of all the components of the proposed framework.

*1) Gated Channel Transformation:* The Gated Channel Transformation [15] mechanism is a channel-wise, contextual in formation modelling technique to enhance feature extraction process with deep CNNs. Notably, a single convolutional layer functions exclusively within a limited local context, often resulting in implicit and non-specific behaviours in the neuron outputs [19]. The GCT mechanism mitigates this constraint by incorporating global context, enabling more efficient and accurate contextual information processing.

GCT comprises three components: global context embedding, channel normalisation, and gating adaptation. The global context embedding aggregates and embeds the global contextual information in each channel by summing the squared activation across the spatial dimensions. The activation feature in a convolutional network is represented as $f \in \mathbb{R}^{C\times H\times W}$, where $H$ and $W$ are the spatial height and width, and $C$ is the number of channels. Accordingly, the global context embedding for $c$-th channel is represented as:

$$E_c = \alpha_c \left\{ \left[ \sum_{i=1}^{H} \sum_{j=1}^{W} \left( f_c^{i,j} \right)^2 \right] + \epsilon \right\}^{\frac{1}{2}} \tag{1}$$

where $\epsilon$ is a small constant introduced to ensure numerical stability. It prevents the problem of derivation at the zero point. Additionally, $\alpha$ ($[\alpha_1, \ldots, \alpha_C]$) is a trainable parameter termed embedding weight to adjust the significance of each channel. When $\alpha_c$ is low, channel $c$ will have minimal to no contribution to the channel normalisation process. Specifically, $\alpha$ enables the GCT module to learn and adapt to the unique characteristics of each channel.

The channel normalisation module employs normalisation technique to establish competitive relationships between neurons [20], thereby enhancing feature representation and promoting more robust and stable training. The gating adaptation module utilises a gating mechanism that dynamically updates the weight of each feature channel based on the learned, normalised global context. This mechanism facilitates both competition and cooperation among channels by modulating their contributions, enabling the network to prioritise the most informative channels through competition and fostering complementary relationships through cooperation. As a result, the network constructs richer and more discriminative feature representations, enhancing its overall performance.

GCT was originally designed to improve the feature representation of standard RGB images for visual recognition tasks [15]. It models contextual information by scaling the convolutional channels based on their significance and is typically deployed before the convolutional layer. Recognising the potential of GCT to capture relationships between channels, we extend its application to mIF images to improve feature extraction across multiple image channels beyond the conventional RGB format.

In our approach, GCT is integrated with the MIL framework and attention mechanism to improve feature extraction for mIF image instances, enabling the effective modelling of information across channels, thereby complementing local patch-level features for immunotherapy response prediction. We adapted all three components of GCT (global context embedding, channel normalisation, and gating adaptation) into the MIL framework. However, unlike the original approach, which deploys GCT at the input stage, COMIL applies it after extracting pretrained features from each image channel. This enables COMIL to adapt to the multichannel structure of mIF images, capture channel-specific representations, and refine them further, thereby facilitating their integration into the MIL framework. More specifically, by incorporating the designed GCT, COMIL is particularly well-suited for whole-slide mIF image classification, ensuring that both channel-specific and inter-channel information are captured and effectively used in downstream task. Furthermore, the tailored adaptation of GCT enables the extraction of features from multiple channels, distinguishing it from the conventional feature extraction process, which is typically limited to three-channel (RGB) images.

Specifically, following segmentation, the whole-slide mIF image is divided into multiple instances. Pre-trained ResNet-101 features are extracted from all instances of each image channel and embedded as $f_{c'} \in \mathbb{R}^{N\times 1\times 2048}$, for the $c'$-th channel, where $c' \in 1, 2, \ldots, C'$. Here $C'$ represents the total number of channels in the mIF image, and $N$ denotes the number of instances in the mIF image.

These features then pass through the GCT unit. First, the global context embedding for $c'$-th mIF image channel is computed as:

$$E_{c'} = \alpha_{c'} \left\{ \left[ \sum_{i=1}^{D} \left( f_{c'}^{i} \right)^2 \right] + \epsilon \right\}^{\frac{1}{2}} \tag{2}$$

where D is the dimension of the feature descriptor and equals 2048, $\epsilon$ is a small constant added for numerical stability, and $\alpha_{c'}$ is a training parameter that serves as the embedding weight. Next, the embedding is normalised using channel normalisation module as:

$$\hat{E}_{c'} = \frac{E_{c'}}{\|E\|_2} \tag{3}$$

where $\hat{E}_{c'}$ represents the normalised global context embedding of the $c^{\text{th}}$ channel and employs $l_2$ normalisation, defined as:

$$\|E\|_2 = \sqrt{\sum_{c'=1}^{C'} {E_{c'}}^2 + \epsilon} \tag{4}$$

where $\epsilon$ represents a small constant. The use of $l_2$ normalisation provides comparatively lower computational complexity, contributing to more stable and efficient training. Subsequently, gating adaptation is applied as:

$$\hat{f}_{c'} = f_{c'} \left[1 + \tanh\left(\gamma_{c'}\hat{E}_{c'} + \beta_{c'}\right)\right] \tag{5}$$

with gating weight $\gamma_{c'}$ and bias $\beta_{c'}$ managing the activations of the gating mechanism for the given image channel. A positive gating weight amplifies channel activations, guiding the model to focus more on the most informative channels by increasing the influence of highly activated channels while down-weighting others. This promotes competition and enhances the representation of the most discriminative features. In contrast, a negative gating weight suppresses strong activations and relatively boosts weaker ones, fostering cooperation by allowing less dominant channels to contribute more evenly. This promotes a more balanced integration of information across channels, generating gating weights that subsequently influence the attention scores and, in turn, the overall performance (see Appendix, Table V for impact of static positive or negative $\gamma$). Using a 1+tanh(x) gate activation prevents a vanishing gradient during training, which makes GCT capable of modelling identity mapping and makes the training process more stable.

Given $\hat{f}_{c'} \in \mathbb{R}^{N\times1\times2048}$ represents the feature matrix of the $c'$-th channel. Thus, $f_{Im}$ represents the complete mIF feature representation with dimensions $N \times C' \times 2048$, aggregating the feature matrices $\hat{f}_{c'}$ of the seven channels. Gating adaptation is followed by adaptive average pooling [21] across the channels to obtain features of dimension $f_{Im} \in \mathbb{R}^{N\times2048}$.

By combining normalisation methods with gating mechanisms, GCT establishes channel relationships that balance both competition and cooperation [20]. This dual approach improves feature representation by allowing channels to compete for relevance while also working together to integrate their outputs effectively. This results in more robust learning by stabilising the training process and handling complex data more efficiently.

Since COMIL is based on MIL with labels provided at the slide-level, we apply an attention mechanism to further enhance the feature representation. A detailed description of the attention mechanism is provided in the following section.

*2) Attention Mechanism:* Given that labels are assigned at the slide-level instead of the instance-level, the attention mechanism enables the model to determine which instances contribute more strongly to the overall label [10]. In this approach, the attention network computes attention scores that help the model focus on relevant features for classification.

For a whole-slide represented as a bag of N instances ($f_{Im} \in \mathbb{R}^{N\times2048}$), we denote the instance-level embedding corresponding to the $n$-th instance as $f_n$. Subsequent to the GCT module and adaptive average pooling, the model includes a projection module $M_{proj}$ followed by an attention module $M_{attn}$ (Fig. 2). The sequential, trainable fully connected layers in $M_{proj}$ transform the feature encoding into a more compact feature space. In particular, $M_{proj}$ compresses the feature vector of the $n^{\text{th}}$ instance ($f_n = R^{2048}$) using a single linear layer ($L_{proj} \in R^{1024\times2048}$) into a 1024-dimensional vector ($f'_n{=}L_{\text{proj}} \cdot f_n$). Following $M_{proj}$, $M_{attn}$ uses attention-based pooling [10] to derive a whole-slide feature representation by aggregating information from all *N* instances in a bag. More specifically, $M_{attn}$ is comprised of multiple stacked fully-connected layers, where $U_a \in R^{512\times1024}$ and $V_a \in R^{512\times1024}$ represent the first two layers of the attention network, while $W_a$ is the shared attention backbone for all classes. The tanh layer captures a rich and diverse range of feature representations, while the sigmoid layer serves as a gating mechanism that modulates this information, introducing non-linearity. Accordingly, the attention score of the $n^{\text{th}}$ instance for the $i^{\text{th}}$ class is given by:

$$\mathrm{a}_{i,n} = \frac{\exp\left(W_{a,i}\left(\tanh(V_a f'_n) \odot \mathrm{sigm}(U_a f'_n)\right)\right)}{\sum_{j=1}^{N}\exp\left(W_{a,i}\left(\tanh(V_a f'_j) \odot \mathrm{sigm}(U_a f'_j)\right)\right)} \tag{6}$$

The slide-level representation ($I_i \in R^{1024}$), which is aggregated based on the attention score distribution for the $i^{\text{th}}$ class is computed as:

$$I_i = \sum_{n=1}^{N} a_{i,n} f'_n \tag{7}$$

Subsequently, the classification layer $W_{c,i} \in R^{1\times1024}$ computes the slide-level score as:

$$S_{slide,i} = W_{c,i} \cdot I_i \tag{8}$$

*3) Instance-Level Clustering:* In order to improve the learning of class-specific features, we utilise instance-level clustering [11] which incorporates an additional binary clustering layer (Fig. 2) during training. Accordingly, we designate the top B instances with the highest in-class attention scores as having a positive cluster label, indicating their support for that class. Conversely, the lowest in-class attention scores are assigned a negative cluster label, signifying their counter evidence against the class. Since the instance-level labels are unavailable, the outputs of the attention network are used to generate pseudo labels for each whole-slide mIF image.

*4) Adaptive Loss Function:* The framework computes two loss values, one at image-level ($Loss_{Im}$) and another at instance-level ($Loss_{In}$). While $Loss_{Im}$ helps to optimise the global image features, $Loss_{In}$ allows the model to focus on local instance features. This enhances the model's robustness, by addressing heterogeneity present within the image. By incorporating both global and local features, the model becomes more adaptable to the complexities and variations in the data, improving its performance and generalisation capability.

To compute $Loss_{Im}$, $S_{slide}$ given by (8) is evaluated against the ground-truth image label using cross-entropy loss. Additionally, $Loss_{In}$ is also computed using cross-entropy loss by comparing the prediction scores for each sampled instance in the instance-level clustering to their respective pseudo-cluster labels. In standard practice, the two loss functions are combined

using a static weight value, which does not consider the potential variations in the importance of global and local features during the training process [11]. Furthermore, it places a greater emphasis on the image-level features while assigning limited importance to the instance loss features. This imbalance often results in suboptimal performance.

To address this limitation and improve the model's capacity to learn both slide-level global and fine-grained local features equitably, we propose adaptive loss function (ADL) to compare the two loss values during each iteration and apply more weight to the higher loss. More specifically, if $Loss_{Im}$ is greater than $Loss_{In}$, it suggests that the model is performing worse with respect to $Loss_{Im}$. Therefore, a higher weight is given to $Loss_{Im}$, causing the model to prioritise its improvement. Conversely, if $Loss_{In}$ is greater than $Loss_{Im}$, the model focuses on minimising $Loss_{In}$ by assigning it more weight. Accordingly, the model is optimised dynamically using both loss functions which are combined as:

$$Loss_{total} = w \times Loss_{Im} + w' \times Loss_{In}, \text{when } Loss_{Im} > Loss_{In}$$

$$Loss_{total} = w \times Loss_{In} + w' \times Loss_{Im}, \text{when } Loss_{In} > Loss_{Im}$$

where $w = \lambda$, $w' = 1 - \lambda$, and $\lambda \in [0.6, 1]$.

### D. Implementation Details

Feature extraction was performed on each $224 \times 224 \times 7$ pixel instance using a ResNet-101 model with a batch size of 64. Random combinations of image transformations were applied, such as horizontal/vertical/left/right flips, rotation at $90^\circ$, $180^\circ$, $270^\circ$.

Each image comprising a few hundred instances underwent further feature refinement with a batch size of one at the image level, along with an attention dropout of 0.25 applied after every layer to improve regularisation. For instance-level clustering, the number of instances with a top score was finetuned to 8 using hyperparameter B. The weight of combining two loss functions ($\lambda$) was finetuned to 0.7. Weighted sampling was used to balance the participation of the two classes during training. The model parameters were updated using the Adam optimiser with a weight decay of 0.00001 and an initial learning rate of 0.0002. Training was conducted for up to 40 epochs with early stopping. The embedding weight $\alpha$ was set to 1, and the gating weights and biases were initialised to zero, in accordance with the recommended implementation guidelines [15].

We evaluated the classification performance of COMIL on two test sets: (1) an internal test set comprising 15% of the 211 labelled whole-slide mIF images, randomly sampled from each class, and (2) an external test set consisting of 42 whole-slide mIF images. The remaining data, after withholding the internal test set, was split into 85% for training and 15% for validation. We employed a six-fold cross-validation scheme for hyperparameter tuning, using one fold for validation and the remaining five for training in each iteration. The optimal hyperparameters were selected on the basis of the average performance across the validation folds. The model was then retrained on the entire dataset, excluding both the internal and external test sets. To prevent information leakage, we ensured that all patches of the same whole-slide image and patients were assigned to the same fold. The methods were evaluated based on the area under the precision-recall curve (AUPRC), area under the curve (AUC) and F1-scores (threshold = 0.50).

For baseline models and comparison frameworks, each encoder was adapted to process the seven-channel mIF input by modifying the first convolutional layer, while keeping the rest of the architecture unchanged, as none of these models incorporate GCT. This follows our previous work [22], with further details provided in the Appendix, Section B.

## III. Results

A series of experiments were conducted to validate the proposed framework and evaluate its performance. To benchmark its effectiveness, we compared its performance with other state-of-the-art frameworks. In addition, ablation studies were performed on different variations of the proposed framework to evaluate the contribution of each component. We also specifically evaluated the effectiveness of the proposed components (GCT and ADL) for mIF image analysis and their integration with ResNet-101 features.

### A. Comparison With Other MIL Frameworks

We investigated other MIL frameworks to compare their performance with our proposed framework (Table I). The compared methods are 1) CLAM-SB [11], which applies a single-attention-branch method and instance clustering to focus on critical regions of interest within a whole-slide image; 2) CLAM-MB [11] extends this approach by incorporating multiple attention branches and instance clustering, enabling the model to capture a wider array of significant regions and features. 3) AttMIL [10] uses attention mechanism to selectively weigh and combine information from multiple instances within a bag, allowing the model to focus on the most relevant features for classification task; 4) DTFD [12] utilises a two-tier method: the first tier applies attention mechanisms and feature distillation strategies (MaxMinS, MaxS, AFS, MAS) to create pseudo-bags, and the second tier refines these features at the slide-level for classification; 5) TransMIL [13] is a transformer-based framework to capture the morphological and spatial correlations between instances, thereby enhancing the slide-level features. 6) PANTHER [23] employs a prototype-based approach grounded in a Gaussian mixture model to compress instances of whole-slide into a compact set of morphological prototypes, enabling efficient slide-level representation for downstream task. All frameworks were implemented using ResNet-101 as the feature extractor to ensure consistency.

PANTHER [23] achieves the second-best performance after COMIL, underscoring the effectiveness of its prototype-based approach grounded in a Gaussian mixture model. It is observed that CLAM-MB (multi-branch) outperforms CLAM-SB (single-branch), highlighting the superior generalisability of the multi-branch attention mechanism. However, within the

TABLE I
COMPARISON OF MIL FRAMEWORKS

| Model | AUPRC | AUC | F1-score | AUPRC | AUC | F1-score |
|---|---|---|---|---|---|---|
| | Internal Test Set | | | External Test Set | | |
| **COMIL (Our method)** | **0.81** | **0.82** | **0.67** | **0.64** | **0.67** | **0.64** |
| PANTHER [23] | 0.77 | 0.78 | 0.62 | 0.59 | 0.57 | 0.56 |
| AttMIL [10] | 0.58 | 0.61 | 0.60 | 0.52 | 0.54 | 0.50 |
| CLAM-SB [11] | 0.62 | 0.67 | 0.58 | 0.48 | 0.48 | 0.45 |
| DTFD [12] | 0.61 | 0.62 | 0.57 | 0.53 | 0.58 | 0.51 |
| CLAM-MB [11] | 0.68 | 0.66 | 0.60 | 0.60 | 0.59 | 0.55 |
| TransMIL [13] | 0.53 | 0.49 | 0.50 | 0.46 | 0.44 | 0.40 |

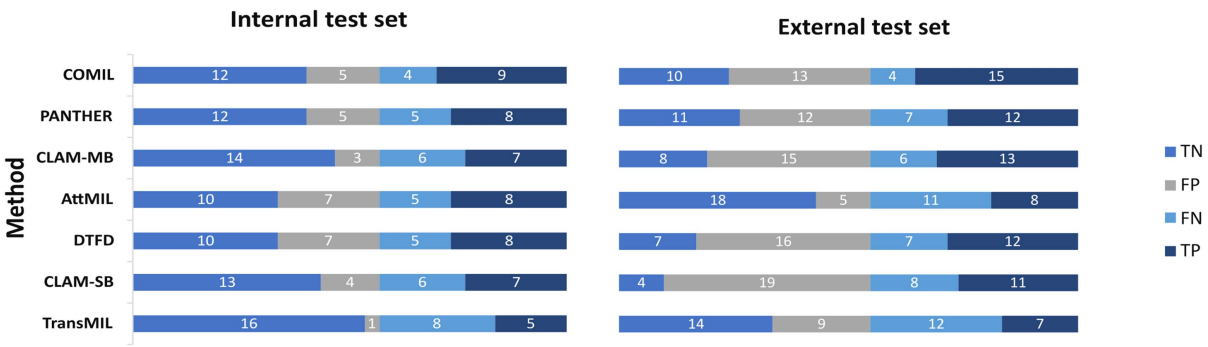



Fig. 3. Comparison of MIL frameworks using confusion matrices.

COMIL framework, the single-branch mechanism performs better, due to the enhanced feature representations provided by the GCT module. Moreover, methods that heavily rely on attention mechanisms, such as AttMIL and DTFD, demonstrate comparatively weaker performance. Notably, TransMIL, despite leveraging self-attention and focusing on inter-instance correlations, achieves the lowest performance overall. These results suggest that while attention-based methods can be effective for whole-slide images, they struggle to capture the intricate and multichannel features present in whole-slide mIF images without the support of a complimentary strategy. This underscores the need for advanced feature extraction methods to fully capture the intrinsic complexity of mIF images.

The results in Table I demonstrate that our proposed MIL framework significantly outperforms existing MIL frameworks across all three metrics on both test sets, validating its effectiveness in whole-slide mIF image classification. The high performance of the COMIL signifies the model's ability to distinguish between positive (non-responsive cases) and negative (responsive cases) classes across all possible thresholds. In the context of predicting therapy response, this ensures that the model can effectively identifies patients unlikely to respond to therapy, while minimising the risk of both false positives and false negatives.

We note that our model achieved F1-scores of 0.67 and 0.64 on the internal and external test sets, respectively, indicating a balanced trade-off between precision and recall (Fig. 3). This balance is particularly crucial in healthcare applications, where both false positives and false negatives can lead to detrimental outcomes. This strong performance highlights the capability of the proposed techniques (GCT and ADL) in our framework to complement the attention mechanism that effectively capture relevant patterns within multiple channels of mIF images. It also underscores the potential of our framework to advance whole-slide mIF image classification in computational pathology for predicting immunotherapy response.

### B. Ablation Studies

We perform ablation studies to assess the contribution of each component of the proposed framework (Table II). The results highlight the impact of these components on classification performance, evaluated using AUPRC, AUC and F1-score. The compared methods are as follows: (i) COMIL (GCT + Att-SB + AttCluster + CE + ADL), which is the proposed method incorporating GCT, attention-based single branch (Att-SB), instance-level clustering based on attention scores (AttCluster), Cross-Entropy (CE) loss function, and the proposed ADL function that dynamically computes the total loss by integrating losses from both instances and whole images during training; (ii) COMIL w/o GCT, excludes GCT from the proposed method and ResNet-101 features are extracted by replacing the first convolutional layer with a new layer to accommodate all seven input channels while retaining the other components; (iii) COMIL w/o ADL, where the method uses a static weight ($\lambda$) to combine loss functions, continuously giving more emphasis to the slide-level loss function. This method replaces the

TABLE II
ABLATION STUDY SHOWING PERFORMANCE ON INTERNAL AND EXTERNAL TEST SETS

| Method | Internal Test Set | | | External Test Set | | |
|---|---|---|---|---|---|---|
| | AUPRC | AUC | F1-score | AUPRC | AUC | F1-score |
| **COMIL (Our method)** | **0.81** | **0.82** | **0.67** | **0.64** | **0.67** | **0.64** |
| COMIL (w/o GCT) | 0.65 | 0.64 | 0.57 | 0.48 | 0.56 | 0.52 |
| COMIL (w/o ADL) | 0.66 | 0.71 | 0.62 | 0.51 | 0.55 | 0.54 |
| COMIL (with Att-MB replacing Att-SB) | 0.72 | 0.69 | 0.60 | 0.49 | 0.51 | 0.53 |
| COMIL (with Entropy based clustering replacing AttCluster) | 0.69 | 0.73 | 0.64 | 0.53 | 0.56 | 0.50 |
| COMIL (with Smooth-SVM loss function replacing CE) | 0.78 | 0.75 | 0.64 | 0.58 | 0.53 | 0.54 |
| COMIL (w/o Att. sigmoid) | 0.67 | 0.66 | 0.44 | 0.54 | 0.55 | 0.39 |
| COMIL (w/o Att. tanh) | 0.58 | 0.51 | 0.42 | 0.38 | 0.38 | 0.35 |
| COMIL (with randomised pixel positions) | 0.66 | 0.64 | 0.56 | 0.50 | 0.52 | 0.54 |

GCT: Gated Channel Transformation; Att-SB/Att-MB: Attention network (Single/Multiple Branch); ADL: Adaptive loss; AttCluster: instance-level clustering based on attention scores; CE: cross-entropy.

proposed ADL function while keeping all other components unchanged; (iv) COMIL (with Att-MB), in which Att-SB is replaced by attention-based multi-branch (Att-MB), while the other components remain the same; (v) COMIL (with Entropy based clustering), in which instance-level clustering based on entropy scores [24] (EntropyCluster) replaces AttCluster while keeping the other components unchanged; (vi) COMIL (with Smooth-SVM loss function), in which Smooth-SVM loss [25] function replaces CE previously used in CLAM [11], while the other components stay unchanged; (vii) COMIL (w/o Att. sigmoid), in which attention mechanism uses only tanh layer, while the other components remain the same; (viii) COMIL (w/o Att. tanh), in which attention mechanism uses only sigmoid layer, keeping all other components unchanged. We note that the model architecture, loss function, learning scheduler, optimiser and other initial hyperparameter settings were the same for all the experiments.

It is observed that each component of the proposed framework have a considerable impact on the overall classification performance. Specifically, eliminating GCT leads to a 20% and 25% drop in AUPRC, 22% and 16% in AUC and 15% and 18% in F1-score on the internal and external test sets, respectively. These results highlight the critical role of GCT in improving feature representation by capturing inter-channel relationships. Additionally, substituting ADL with a static $\lambda$ causes a 19% and 20% reduction in AUPRC, 13% and 18% in AUC and a 8% and 16% in the F1-score on the internal and external test sets, respectively, underscoring the importance of tuning the model for balanced optimisation at both the slide and instance levels. Both components significantly contribute to true positive and true negative predictions while minimising false positives and false negatives, thereby strengthening the model's robustness and reliability.

Furthermore, replacing Att-SB with Att-MB results in a 11% and 8% reduction in AUPRC, 16% and 23% in AUC and a 11% and 17% in F1-score on the internal and external test sets, respectively. This is because, when combined with GCT, Att-SB aggregates globally contextualized feature representations obtained from GCT into a single, coherent slide-level representation, thereby fully leveraging the synergy of GCT. In contrast, Att-MB partitions the representations across multiple branches, diluting the integration of global context, and limiting discriminative power. This introduces redundancy and weakens the signal, leading to the observed performance drop. In comparison, CLAM benefits from Att-MB over Att-SB (Table I) because its attention operates on raw features without GCT's global context, allowing it to capture diverse regions more effectively, which is advantageous in that framework.

Additionally, to further refine the process of selecting instances for training, we explored the use of entropy scores [24] for instance-level clustering (EntropyCluster). Entropy, recognised as an effective measure for extracting meaningful information [26], offers valuable insights into the model's uncertainty and confidence in selecting instances. By leveraging entropy, the model gains a broader view of the distribution of attention across instances, potentially enabling more informed decision-making. However, contrary to expectations, the model trained using entropy scores demonstrates relatively lower performance compared to the proposed method. This may be attributed to the fact that attention scores directly focus on key features, emphasising the most relevant parts of the image, while entropy scores capture uncertainty about the importance of the instance, which may not be representative of the true decision boundary of the task. This loss of specificity can hinder effective feature selection and degrade performance.

It is also observed that replacing the CE loss function with the Smooth-SVM loss function, as used in CLAM [11] leads to a decline in performance. The CE loss function outperforms Smooth-SVM because it aligns with the probabilistic nature of the model, providing smoother gradients and better handling of uncertain or overlapping class predictions as it applies higher penalties to misclassified examples. In contrast, smooth-SVM focuses on maximising margins between classes, which may not work as effectively in complex decision boundaries.

Furthermore, ablation results show that removing tanh or sigmoid from the gated attention mechanism significantly decreases performance. This is because tanh encodes content-based signals, while sigmoid gates feature contributions. Altering these activations destabilises attention computation, leading to a drop in model performance. Additionally, to evaluate

TABLE III
CHANNEL ABLATION STUDY

| Excluded biomarker | Internal Test Set | | | External Test Set | | |
|---|---|---|---|---|---|---|
| | AUPRC | AUC | F1-score | AUPRC | AUC | F1-score |
| CD8 | 0.58 | 0.62 | 0.57 | 0.51 | 0.52 | 0.50 |
| CD16 | 0.56 | 0.63 | 0.59 | 0.50 | 0.51 | 0.51 |
| CD68 | 0.73 | 0.73 | 0.62 | 0.59 | 0.61 | 0.60 |
| DAPI | 0.63 | 0.70 | 0.62 | 0.51 | 0.52 | 0.54 |
| PD-L1 | 0.65 | 0.57 | 0.56 | 0.53 | 0.55 | 0.54 |
| SOX10 | 0.63 | 0.68 | 0.57 | 0.52 | 0.52 | 0.55 |
| Autofluorescence | 0.64 | 0.59 | 0.61 | 0.45 | 0.48 | 0.46 |

Note: Each row shows model performance when the specified channel is excluded from the input. The reported metric values on the internal and external test sets reflect the impact of omitting that channel, highlighting its contribution to overall model performance.

whether COMIL effectively captures spatial and morphological patterns, we implemented a randomised-pixel baseline by shuffling pixel values. The substantial performance drop confirms reliance on meaningful structural information.

We also conducted a channel ablation study by systematically excluding each channel and evaluating the resulting performance impact (Table III). Our analysis revealed that CD8 and CD16 channels contribute most significantly to the model's predictive performance, consistent with findings from previous studies [17], [27]. Notably, the exclusion of AF has comparatively greater impact on the external test set, suggesting it provides contextual cues that support generalisation. These cues such as spatial patterns, structural artefacts, or morphological context may help the model maintain robust predictions under domain shift. Although not marking specific biological features, AF captures background signals that reflect tissue architecture and enhance feature representations. Overall, all channels contribute positively to the performance, indicating that the model effectively leverages complementary information across all biomarkers to achieve robust predictions.

### C. GCT-Driven Feature Extraction and Equitable Optimisation At Slide- and Instance-Levels

Our proposed framework, COMIL, introduces two key techniques in the MIL framework for whole-slide mIF image classification: GCT for improved feature extraction and ADL to optimise training equitably at both the slide- and instance-levels. GCT addresses the challenge of modelling the complex inter-channel relationships inherent in multichannel mIF images, enabling more effective feature extraction across biomarkers. ADL handles the limitation of imbalanced training supervision in MIL frameworks, where more emphasis on slide-level training can limit the model's ability to capture fine-grained local patterns. By adaptively weighting loss contributions, ADL enhances the model's ability to learn local features alongside global slide-level representations. Accordingly, COMIL achieved the best performance when evaluated against other MIL methods, demonstrating its superior capability in handling the complexities of whole-slide mIF images (Fig. 3, Table I). Notably, the two innovative components exhibited considerable improvements in performance in ablation studies, further highlighting their significance in optimising feature extraction and training within the MIL framework (Table II). Additionally, Fig. 4 qualitatively illustrates the contributions of the proposed components of COMIL, namely, GCT and ADL to the predictive performance.

The standard approach to feature extraction from mIF images typically involves replacing the first convolutional layer with a new layer that accommodates all seven input channels. However, as demonstrated by the results of the ablation study (Table II), this method (w/o GCT) does not adequately capture and integrate the complex relationships between multiple channels and spatial dimensions. Integrating GCT into the MIL framework enhances feature extraction by addressing these limitations. GCT incorporates a global context embedding module, which captures global information for each channel. Normalisation is then applied to appropriately scale the features, followed by gating adaptation, which fine-tunes the importance of each feature channel by modelling inter-channel relationships of competition and cooperation. GCT is deployed after extracting pretrained features from each mIF image channel, allowing capturing complex patterns and accommodating a varying number of channels, thus enhancing its adaptability.

As GCT is central to our feature extraction strategy, the version of COMIL without GCT serves as a natural baseline to assess overfitting. Accordingly, in addition to assessing the impact of the GCT by comparing COMIL with and without it using ResNet-101 features, we evaluated seven other encoders, ResNet-101/50/18 [28] and DenseNet-201/161/121 [29], UNI [30], and Virchow [31], on internal and external test sets. The consistent improvement observed with GCT across all architectures suggests that the performance gain is robust and not tied to a specific encoder, mitigating concerns of overfitting to a particular model architecture (Table IV).

Furthermore, we used logistic regression to evaluate summary statistics and morphological features (handcrafted features), including tissue area, total cell count, and mean biomarker intensities. As shown in Table IV, these features alone are insufficient, highlighting the need for deep learning features within an advanced framework to capture complex spatial and morphological patterns. We also evaluated KRONOS [32] features using COMIL with the ADL function but without GCT, since GCT is specifically designed for raw channel features to capture global contextualized inter-channel interactions,

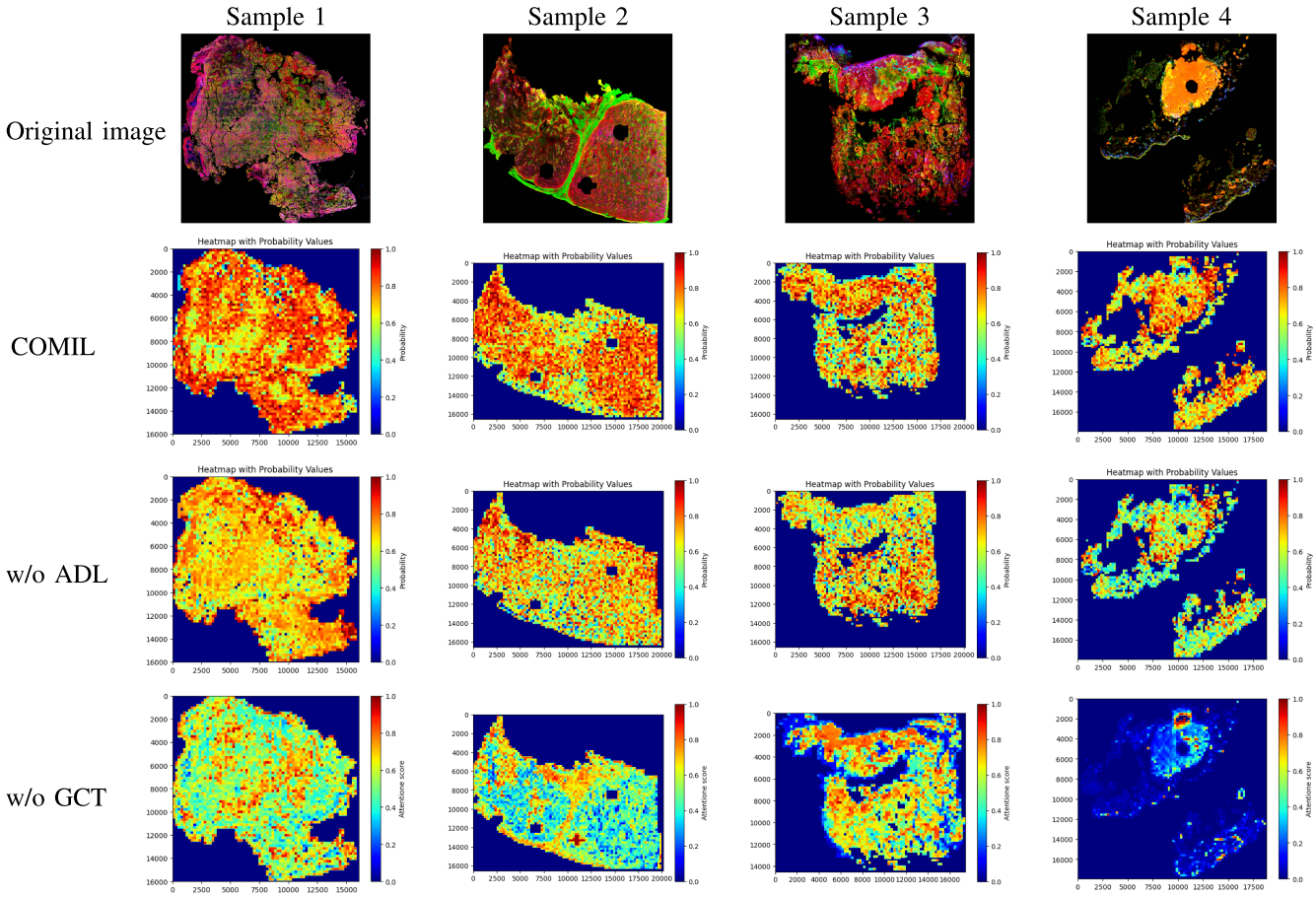



Fig. 4. Comparative heatmaps for four sample images across different methods. Samples 1 and 2 belong to the responder class, while samples 3 and 4 are from the non-responder class. The first row displays the original images; the second row shows the results of the proposed method (COMIL); the third row presents results without the global context transformation (w/o GCT); and the fourth row displays results without the adaptive loss function (w/o ADL).

TABLE IV
PERFORMANCE COMPARISON OF BASE AND PROPOSED (COMIL) MODELS WITH DIFFERENT FEATURE ENCODERS

| Method | Model | Internal Test set | | | External Test set | | |
|---|---|---|---|---|---|---|---|
| | | AUPRC | AUC | F1-score | AUPRC | AUC | F1-score |
| Handcrafted features KRONOS* | Base | 0.60 | 0.62 | 0.54 | 0.48 | 0.50 | 0.52 |
| | Proposed | 0.72 | 0.69 | 0.64 | 0.62 | 0.60 | 0.62 |
| VIRCHOW | Base | 0.64 | 0.59 | 0.53 | 0.45 | 0.51 | 0.44 |
| | Proposed | 0.74 | 0.72 | 0.64 | 0.58 | 0.54 | 0.56 |
| UNI | Base | 0.60 | 0.56 | 0.50 | 0.43 | 0.40 | 0.42 |
| | Proposed | 0.67 | 0.71 | 0.62 | 0.51 | 0.53 | 0.50 |
| DenseNet-121 | Base | 0.54 | 0.51 | 0.48 | 0.44 | 0.47 | 0.38 |
| | Proposed | 0.61 | 0.60 | 0.58 | 0.56 | 0.59 | 0.50 |
| DenseNet-161 | Base | 0.61 | 0.56 | 0.50 | 0.45 | 0.51 | 0.40 |
| | Proposed | 0.64 | 0.63 | 0.60 | 0.57 | 0.59 | 0.58 |
| DenseNet-201 | Base | 0.74 | 0.70 | 0.61 | 0.50 | 0.58 | 0.56 |
| | Proposed | 0.79 | 0.78 | 0.65 | 0.60 | 0.62 | 0.61 |
| ResNet-18 | Base | 0.53 | 0.54 | 0.43 | 0.43 | 0.42 | 0.36 |
| | Proposed | 0.61 | 0.63 | 0.49 | 0.52 | 0.47 | 0.40 |
| ResNet-50 | Base | 0.60 | 0.58 | 0.50 | 0.45 | 0.44 | 0.40 |
| | Proposed | 0.71 | 0.74 | 0.62 | 0.61 | 0.64 | 0.53 |
| ResNet-101 | Base | 0.65 | 0.64 | 0.57 | 0.48 | 0.56 | 0.42 |
| | Proposed | **0.81** | **0.82** | **0.67** | **0.64** | **0.67** | **0.64** |

Note*: KRONOS features were processed without GCT, since GCT targets raw channel features, whereas KRONOS provides high-level semantic mIF representations.

whereas KRONOS already provides high-level semantic representations. The results indicate that while KRONOS provides general-purpose semantic representations, the absence of explicit inter-channel modelling limits its discriminative power for our task.

COMIL demonstrates robustness against overfitting to pixel distributions and channel order. We systematically altered the positions of the biomarker channels. Specifically, we relocated key biomarkers, such as moving the DAPI channel from position 3 to positions 0 and 6 and applied maximal disruptions, including inverting the biomarker order and repositioning the AF channel. Across all these configurations, model performance remained consistent. This stability is attributed to the GCT module, which adaptively captures and integrates inter-channel dependencies in a position-invariant manner. By focusing on the relationships among biomarkers rather than their fixed input positions, GCT enables the model to maintain robust predictions despite permutations in channel arrangement.

Additionally, another limitation addressed by COMIL is the primary focus of current MIL frameworks on slide-level training and overlooking instance-level optimisation, which can lead to suboptimal performance. This is demonstrated by the ablation study (Table II, w/o ADL), where slide-level and instance-level loss function values are combined with fixed weights of 0.7 and 0.3, respectively, throughout training. COMIL introduces an ADL function that dynamically adjusts the emphasis between slide-level and instance-level loss functions based on their relative values. This enables more targeted and effective optimisation by directing gradients toward the most significant errors.

Furthermore, by leveraging attention scores, COMIL produces heatmaps that visually represent regions of interest. As shown in Fig. 4, the proposed method generates prominent attention scores to outline distinct, well-defined areas, highlighting the contribution of each region within the whole-slide image. In contrast, the heatmaps of rows 3 and 4, corresponding to methods without ADL and GCT, exhibit less pronounced attention. Specifically, the exclusion of ADL leads to reduced clarity in attention scores, while the absence of GCT results in less distinct attention areas. These findings corroborated by the results of the ablation study indicate that both ADL and GCT contribute significantly to the effective feature representation and performance of the classification model.

Notably, the responder samples (samples 1 and 2) exhibit more spatially coherent and dense regions in the heatmaps generated by COMIL compared to the non-responder samples (samples 3 and 4). This observation is consistent with established biological knowledge that responders typically demonstrate higher immune cell density, where overlapping membranes of adjacent immune cells contribute to increased signal intensities. The presence of high-intensity focused regions, combined with strong classification performance, highlights its effectiveness in addressing potential clinically significant tissue regions.

## IV. Discussion

Our study introduces COMIL, a deep learning-based method to predict immunotherapy responses in patients with melanoma. Trained on mIF images, COMIL outperforms previous methods [33] trained on clinical data to predict immunotherapy responses. This improvement highlights the significance of mIF imaging in capturing intricate molecular and cellular details that are often not discernible in clinical datasets alone. COMIL processes whole-slide mIF images, enabling simultaneous analysis of multiple biomarkers within a single tissue section by using the GCT module to capture interdependencies across channels, while leveraging MIL to capture spatially fine-grained information. This capability fully accounts for complex biological variations and improves the selection of patients who are likely to respond to immunotherapy [4].

We note that COMIL achieves strong performance on the internal test set, demonstrating effective learning within the training domain. While performance declined on the external test set, the framework maintained reasonable generalisability across datasets. This decrease is expected due to domain-specific variations such as differences in staining, acquisition, or cohort characteristics. To mitigate overfitting, we employed strict patient- and slide-level data partitioning, cross-validation for hyperparameter tuning, and regularisation techniques including dropout, early stopping and weight decay. These strategies helped preserve generalisation despite domain shifts, as reflected in the model's stable F1-score and overall robustness on external data.

Accordingly, COMIL advances whole-slide mIF image analysis by introducing a deep learning-based approach for whole-slide mIF images, in contrast to previous studies that relied on handcrafted features [6], [8] or deep learning studies dependent on manually annotated regions of interest[34], [35], [36]. Manual annotation requires expert input to define relevant regions, which limits scalability and and often overlooks the broader context available in whole-slide image. In contrast, COMIL processes the entire slide automatically, enabling scalable analysis while capturing both global tissue architecture and fine-grained biomarker patterns through automatically learned deep features. Given its superior performance, ResNet-101 was selected as the feature extractor for COMIL (Table IV). We attribute this to ImageNet pretraining on a diverse range of natural and synthetic satellite and space images. Notably, mIF images share structural characteristics with such imagery [8] which likely enabled ResNet-101 to generalise more effectively.

Furthermore, our study demonstrates that COMIL effectively addresses the significant challenge of analysing whole-slide mIF images, which are characterised by a large number of channels and extensive spatial dimensions (Fig. 1(a)). The GCT module accommodates an arbitrary number of biomarkers and provides a distinct advantage in capturing diverse features across various channels, and ADL effectively optimises training at both the slide- and instance-levels. The results indicate that the proposed framework outperforms existing MIL frameworks (Table I), with the ablation study (Table II) highlighting that the integration of the two proposed techniques into the MIL framework further enhances performance. Accordingly, COMIL proves to be an effective framework for handling complex mIF images. Additionally, we note that our feature extraction process is not channel-specific, thereby enabling the method to be adaptable to any channel type.

Our model's consistent F1-score, recall, and precision suggest that it is reasonably effective, providing a good balance between identifying positive cases (recall) and ensuring that the positive predictions are correct (precision). In practice, this balance is crucial to optimise therapeutic strategies, potentially improving clinical outcomes. Misclassifying a patient who would benefit from treatment (false negative) could lead to disease progression, potentially reducing survival rates. On the other hand, incorrectly predicting that a patient will respond to treatment (false positive) can result in the unnecessary administration of toxic therapies, leading to side effects, financial costs, and a decline in the patient's quality of life. Therefore, a predictive model for an immunotherapy response that accurately identifies non-responders and responders with balanced recall and precision aids in the personalisation of cancer care, allowing for more effective and individualised treatment plans. In addition, it improves resource allocation within healthcare systems, ensuring that expensive and intensive treatments are reserved for patients who are more likely to benefit from them.

Using COMIL, we are able to generate heatmaps that can potentially be utilised as a visualisation tool to identify regions of interest, offering valuable insights and assisting in clinical practice. Furthermore, since these heatmaps are based on attention scores and highlight key regions that contribute to the model's decision-making process, they can also assist clinicians in correlating morphological features with immunotherapy responses and resistance.

A limitation of our proposed framework is that it is based on MIL, which is a weakly supervised method. The main challenge of MIL is label ambiguity, as the model is trained on instances with image-level labels, which are often noisy. In future work, we aim to address this by developing advanced pseudo-labeling techniques to improve instance-level clustering. Furthermore, we will further improve the predictions by strengthening regularisation, particularly through data augmentation strategies specifically designed for mIF images. This would generate a more diverse set of samples for robust training, ultimately improving classification performance.

Moreover, the proposed feature extraction technique using GCT is versatile and can be extended to tasks beyond classification. One such potential application is mIF image generation, where the proposed GCT module can be used to develop methods for generating synthetic mIF images. Additionally, stain imputation is a promising use case. By utilising COMIL's ability to capture channel information effectively and integrating it with methods for imputing missing stains in mIF images, the reliance on repeated staining is minimised, thereby reducing the consumption of costly reagents.

As cancer is a multifaceted disease, the response to treatment is influenced by patient profiles, genetics, and the tumor microenvironment. Accordingly, we aim to develop multimodal training methods to capture these complexities and further enhance performance. Thus, while current performance provides a foundation, further improvements will likely be necessary to ensure the model's complete reliability in supporting clinical decision-making in cancer care.

## V. Conclusion

In this study, we introduced the Channel Optimisation with Multi-Instance Learning (COMIL) framework to predict immunotherapy responses in melanoma patients using whole-slide multiplex immunofluorescence (mIF) images. COMIL addresses the challenge of extracting features from multiple mIF channels by modelling inter-channel relationships and capturing complex dependencies, resulting in enhanced classification performance. Its integration with a multi-instance learning framework, coupled with optimised training at both the slide and instance levels, further improves its ability to capture the spatially dispersed information in whole-slide images. Our evaluation on mIF images from the Melanoma Institute Australia demonstrates that COMIL outperforms existing baseline methods, highlighting its potential as a robust tool for developing personalised treatment strategies in melanoma immunotherapy.

## Author Contributions

**Priyanka Rana**: Conceptualisation, Software, Methodology, Formal analysis, Investigation, Validation, Visualisation, Writing – original draft, Writing – review & editing. **Tuba N Gide**: Data curation, Resources, Writing - review & editing, Project administration. **Nurudeen A Adegoke**: Data curation, Resources, Writing - review & editing. **Yizhe Mao**: Data curation, Resources, Writing - review & editing. **Shlomo Berkovsky**: Resources, Writing - review & editing. **Enrico Coiera**: Resources, Writing - review & editing, Funding acquisition. **James S Wilmott**: Data curation, Resources, Writing - review & editing, Funding acquisition, Supervision. **Sidong Liu**: Data curation, Resources, Writing - review & editing, Supervision.

## Competing Interests

The authors declare that they have no competing interests.

## Appendix

### A. Input Data Patterns

A comparative analysis was conducted on tissue area ($mm^2$), total cell counts, and average cell intensities of all the biomarkers between responders and non-responders. Overall, no significant differences were observed in tissue area, total cell number, CD68, DAPI, or AF intensities. SOX10 showed a borderline difference. In contrast, the intensities of CD16, CD8, and PD-L1 were higher in responders. The p-value from the two-tailed Mann–Whitney U test indicates the observed differences are statistically significant.

### B. Implementation Details of Baseline and Comparison Frameworks That Do Not Utilise GCT

For the ImageNet-based models (ResNet and DenseNet), the first convolutional layer was modified to accept seven input channels by retaining pretrained weights for the original three RGB channels and initialising the additional four channels with the average of these weights. Similarly, in the UNI model, the

TABLE V
PERFORMANCE OF MODEL VARIANTS

| Model | Internal Test Set | | | External Test Set | | |
|---|---|---|---|---|---|---|
| | AUPRC | AUC | F1-score | AUPRC | AUC | F1-score |
| Positive $\gamma$ | 0.62 | 0.63 | 0.38 | 0.52 | 0.48 | 0.50 |
| Negative $\gamma$ | 0.64 | 0.74 | 0.35 | 0.39 | 0.38 | 0.21 |
| Trainable $\gamma$ (COMIL) | 0.81 | 0.82 | 0.67 | 0.64 | 0.67 | 0.64 |

patch embedding layer was adjusted to accommodate seven channels, with the extra channels initialised by averaging the pretrained RGB weights. For the Virchow model, the convolutional projection layer was replaced with a seven-channel version that preserved the original output dimensions and stride, copying the pretrained weights for the first three channels and initialising the remaining channels by averaging the pretrained RGB weights.

## ACKNOWLEDGMENT

The authors would like to thank Melanoma Institute Australia (MIA) patients and their families for their contributions to our study. Ongoing support from colleagues at the MIA and the Charles Perkins Centre, University of Sydney, is also greatly acknowledged. Support from the ClearBridge Foundation as well as from colleagues at MIA and Royal Prince Alfred Hospital is also gratefully acknowledged. This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government. We thank Satya Vedantam and Richard Miller for their assistance in facilitating access.

## REFERENCES

[1] P. P. Centeno, V. Pavet, and R. Marais, "The journey from melanocytes to melanoma," *Nature Rev. Cancer*, vol. 23, no. 6, pp. 372–390, 2023.
[2] J. Zhang, A. M. Joshua, Y. Li, C. H. O'Meara, M. J. Morris, and L. M. Khachigian, "Targeted therapy, immunotherapy, and small molecules and peptidomimetics as emerging immunoregulatory agents for melanoma," *Cancer Lett.*, vol. 586, 2024, Art. no. 216633.
[3] A. Abaza et al., "Programmed cell death protein 1 (PD-1) and programmed cell death ligand 1 (PD-L1) immunotherapy: A promising breakthrough in cancer therapeutics," *Cureus*, vol. 15, no. 9, 2023, Art. no. e44582.
[4] X. Lin, C. Zong, Z. Zhang, W. Fang, and P. Xu, "Progresses in biomarkers for cancer immunotherapy," *MedComm*, vol. 4, no. 5, 2023, Art. no. e387.
[5] D. Schapiro et al., "MCMICRO: A scalable, modular image-processing pipeline for multiplexed tissue imaging," *Nature Methods*, vol. 19, no. 3, pp. 311–315, 2022.
[6] S. Lu et al., "Comparison of biomarker modalities for predicting response to PD-1/PD-L1 checkpoint blockade: A systematic review and meta-analysis," *JAMA Oncol.*, vol. 5, no. 8, pp. 1195–1204, 2019.
[7] Z. Yaseen et al., "Validation of an accurate automated multiplex immunofluorescence method for immuno-profiling melanoma," *Front. Mol. Biosciences*, vol. 9, 2022, Art. no. 810858.
[8] S. Berry et al., "Analysis of multispectral imaging with the astropath platform informs efficacy of PD-1 blockade," *Science*, vol. 372, no. 6547, 2021, Art. no. eaba2609.
[9] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1-2, pp. 31–71, 1997.
[10] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.
[11] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, 2021.
[12] H. Zhang et al., "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18780–18790.
[13] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, and X. Ji, "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 2136–2147.
[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[15] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11791–11800.
[16] S. Hochreiter and J. Schmidhuber., "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[17] T. Gide et al., "Distinct immune cell populations define response to anti-PD-1 monotherapy and anti-PD-1/anti-CTLA-4 combined therapy," *Cancer Cell*, vol. 35, no. 2, pp. 238–255, 2019.
[18] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. TSMC-9, no. 1, pp. 62–66, Jan. 1979.
[19] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, pp. 169–191, 2003.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
[21] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. Int. Conf. Learn. Representations*, 2013.
[22] P. Rana, A. Sowmya, E. Meijering, and Y. Song, "Imbalanced classification for protein subcellular localization with multilabel oversampling," *Bioinformatics*, vol. 39, no. 1, 2023, Art. no. btac841.
[23] A. H. Song, R. J. Chen, T. Ding, D. F. Williamson, G. Jaume, and F. Mahmood, "Morphological prototyping for unsupervised slide representation learning in computational pathology," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 11566–11578.
[24] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.
[25] L. Berrada, A. Zisserman, and M. P. Kumar, "Smooth loss functions for deep top-k classification," in *Proc. Int. Conf. Learn. Representations*, 2018.
[26] B. Liu, C. Li, Z. Li, D. Wang, X. Ren, and Z. Zhang, "An entropy-based metric for assessing the purity of single cell populations," *Nature Commun.*, vol. 11, no. 1, 2020, Art. no. 3155.
[27] H. Lee et al., "Intratumoral CD16 macrophages are associated with clinical outcomes of patients with metastatic melanoma treated with combination anti-PD-1 and anti-CTLA-4 therapy," *Clin. Cancer Res.*, vol. 29, no. 13, pp. 2513–2524, 2023.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.
[30] R. J. Chen et al., "Towards a general-purpose foundation model for computational pathology," *Nature Med.*, vol. 30, no. 3, pp. 850–862, 2024.
[31] E. Vorontsov et al., "Virchow: A million-slide digital pathology foundation model," 2023, *arXiv:2309.07778*.
[32] M. Shaban et al., "A foundation model for spatial proteomics," 2025, *arXiv:2506.03373*.
[33] I. Pires da Silva et al., "Clinical models to define response and survival with anti-PD-1 antibodies alone or combined with ipilimumab in metastatic melanoma," *J. Clin. Oncol.*, vol. 40, no. 10, pp. 1068–1080, 2022.
[34] M. Sorin et al., "Single-cell spatial landscape of immunotherapy response reveals mechanisms of CXCL13 enhanced antitumor immunity," *J. Immunotherapy Cancer*, vol. 11, no. 2, 2023, Art. no. e005545.
[35] K. V. Hoebel et al., "Deep-learning model trained on multiplex immunofluorescence-stained tissue samples predicts the survival of patients with non-small cell lung cancer better than PD-l1 TPS alone," *Cancer Res.*, vol. 84, no. 6_Supplement, pp. 6189–6189, 2024.
[36] Z. Wu et al., "Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens," *Nature Biomed. Eng.*, vol. 6, no. 12, pp. 1435–1448, 2022.