# Deep Learning-based Prediction of Breast Cancer Tumor and Immune Phenotypes from Histopathology

Tiago Gonçalves[1,2,3], Dagoberto Pulido-Arias[1], Julian Willett[1], Katharina V. Hoebel[1], Mason Cleveland[1], Syed Rakin Ahmed[1], Elizabeth Gerstner[1], Jayashree Kalpathy-Cramer[1], Jaime S. Cardoso[2,3], Christopher P. Bridge[1,*], Albert E. Kim[1,*]

[1]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School
[2]Faculty of Engineering, University of Porto
[3]Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
[*]These authors contributed equally

## Introduction

- **Context** Efforts to develop a widely available gene expression panel quantifying facets of the tumor microenvironment (TME) have been limited due to the need for RNA extraction from fresh tissue instead of the more widely available paraffin-embedded tissue. Furthermore, pathology workflows mainly evaluate tumor cell characteristics (e.g., grading, subtyping) within H&E and are not able to reproducibly assess the TME [1].
- **Motivation** As spatial profiling efforts suggest that the physical colocalization of non-malignant cells to tumor cells modulate intracellular communication, tumor phenotypes, and clinical outcomes, there is likely biologically relevant information from H&E that is not used in clinical care.
- **Proposal** Here, we present an exploratory application of multiple instance learning (MIL) algorithms to characterize tumor and immune phenotypes from the H&E whole slide image (WSI) of primary breast tumors. We hypothesized that deep learning (DL) could predict these facets of the TME through recognizing biologically relevant features (e.g., tissue architecture, tumor-immune interface) within an individual H&E image (see Figure 1).

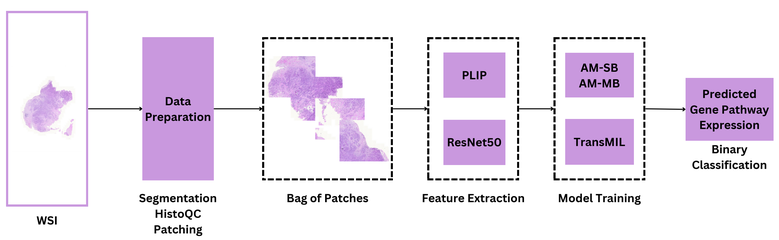


Figure 1. Overview of the proposed pipeline.

## Methodology

- **WSI Data Preparation:** We used the TCGA-BRCA dataset and performed quality assessment with HistoQC, discarding the low quality WSIs. We split each WSI into $256 \times 256$ patches, and extracted features using two different models: ResNet50, pre-trained on a natural image dataset (ImageNet); and PLIP [2], pre-trained on a pathology dataset (OpenPath).
- **Quantification of Gene Pathway Expression:** We employed single-sample GSEA (ssGSEA) to calculate gene set enrichment scores for 10 different pathways (see Table 1). ssGSEA transforms a single sample's gene expression profile to a gene set enrichment profile.
- **Models and Training Details:** The ssGSEA scores underwent binarization, with negative scores receiving the label $0$ and the rest label $1$. A random data split allocated $70\%$ to training, $15\%$ to validation, and $15\%$ to testing, ensuring exclusive assignment of a given patient's data to one split. We trained the models using CLAM-based models (AM-SB, AM-MB) [3] and TransMIL [4].

## Results and Discussion

- **Predictive Performance** Consistent with our hypothesis, our models predict TME phenotypes from H&E WSIs to reasonable accuracy (see Table 1), with models trained with PLIP-based features generally outperforming those trained with ResNet50-based features.

Table 1. **AUROC results on the test set, obtained for all model architectures, using ResNet50- and PLIP-derived features.** The best results per type of feature-extraction strategy are highlighted in bold. While the AM-SB and AM-MB model architectures outperform TransMIL prediction of all gene expression tasks, the increase in performance was marginal, suggesting that model architecture may not play a critical role in the classification of gene expression.

| Task | Architecture & Features | | |
|---|---|---|---|
| | AM-SB | AM-MB | TransMIL |
| | ResNet50 / PLIP | ResNet50 / PLIP | ResNet50 / PLIP |
| B-cell proliferation | 0.6974 / **0.7755** | **0.7322** / 0.7735 | 0.6960 / 0.7673 |
| T-cell mediated cytotoxicity | 0.7149 / 0.7703 | **0.7564** / **0.7770** | 0.7223 / 0.7196 |
| Angiogenesis | 0.7042 / **0.7435** | **0.7053** / 0.7213 | 0.7019 / 0.7214 |
| Epithelial-mesenchymal transition | **0.8110** / 0.7848 | 0.8023 / **0.8082** | 0.7545 / 0.7934 |
| Fatty acid metabolism | **0.6323** / **0.6030** | 0.6294 / 0.5920 | 0.5258 / 0.5634 |
| Glycolysis | **0.7996** / 0.8118 | 0.7954 / **0.8330** | 0.7834 / 0.8045 |
| Oxidative phosphorylation | 0.6894 / 0.7145 | **0.6926** / **0.7332** | 0.6699 / 0.6826 |
| Immunosuppression | 0.7996 / 0.8458 | **0.8133** / **0.8542** | 0.7572 / 0.8113 |
| Antigen processing and presentation | 0.7450 / 0.7806 | **0.7599** / **0.7924** | 0.7503 / 0.7342 |
| Cell cycling | 0.7768 / **0.7939** | **0.7809** / 0.7852 | 0.7121 / 0.7229 |

- **Preliminary Study on Explainability** We generated an attention map on an H&E WSI that was predicted to have a high level of T-cell mediated cytotoxicity (see Figure 2).

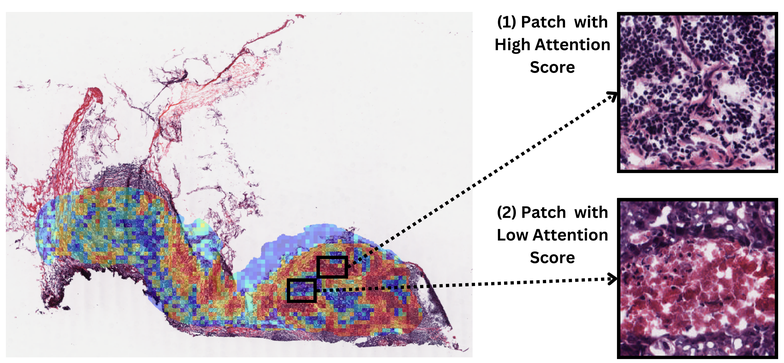


Figure 2. **Attention map obtained using the AM-SB architecture for a WSI predicted to have a high-degree of T-cell mediated cytotoxicity**. Red zones and blue zones indicate high and low attention scores, respectively. On the right, we provide two exemplar patches: 1) the high-attention patch illustrates abundant tumor-infiltrating lymphocytes without tumor cells, which are suggestive of high immune activity, and 2) the low-attention region demonstrated areas of tumor necrosis and minimal lymphocytes, consistent with low immune activity.

## Conclusions and Future Work

- Our work suggests that tumor and immune phenotypes can be inferred with reasonable accuracy using state-of-the-art DL methods. Importantly, we note that the choice of feature extraction strategy may impact performance. A model pre-trained on H&E-based features had improved performance for most tasks, compared to a model pre-trained on ImageNet-based features.
- As prediction of tumor and immune phenotypes can be inferred from H&E, these computational H&E biomarkers would have major implications for cancer patients and precision oncology as many hospitals only possess the capability for H&E and lack the resources for DNA/RNA sequencing.
- We will experiment with other feature extraction frameworks and feature fusion approaches (e.g., combining two different feature sets in a single model). Additionally, we intend to explore ordinal and multi-task classification and measures to boost model interpretability. Finally, we will trial multi-modal architectures, integrating relevant clinical metadata during training to assess the impact on performance.

## Acknowledgments

The results published or shown here are in whole or part based upon data generated by the TCGA Research Network (`https://www.cancer.gov/tcga`). This work was funded by the William G. Kaelin, Jr., M.D., Physician-Scientist Award of the Damon Runyon Cancer Research Foundation, American Association for Cancer Research Breast Cancer Research Fellowship, and Conquer Cancer/American Society of Clinical Oncology Young Investigator Award, and by the Portuguese Foundation for Science and Technology (FCT) through the Ph.D. grant "2020.06434.BD".

## References

[1] S. K. Longo, M. G. Guo, A. L. Ji, and P. A. Khavari, "Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics," *Nature Reviews Genetics*, vol. 22, no. 10, pp. 627–644, 2021.

[2] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical Twitter," *Nature Medicine*, pp. 1–10, 2023.

[3] 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 Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[4] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, *et al.*, "TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2136–2147, 2021.



tfgoncalves@mgh.harvard.edu, cbridge@mgh.harvard.edu, akim46@mgh.harvard.edu