GENERALIZED REPRESENTATION LEARNING FOR MULTIMODAL HISTOLOGY IMAGING DATA THROUGH VISION-LANGUAGE MODELING

Jacob S. Leiby & Dokyoon Kim
Department of Biostatistics, Epidemiology & Informatics
University of Pennsylvania
Philadelphia, PA, USA
{jleiby,dokyoon.kim}@pennmedicine.upenn.edu

Alexandro E. Trevino & Aaron T. Mayer
Enable Medicine
Menlo Park, CA, USA
{alex,aaron}@enablemedicine.com

Zhenqin Wu
School of Computing and Data Science
The University of Hong Kong
Pokfulam, Hong Kong
zqwu@cs.hku.hk

Zhi Huang
Department of Pathology and Laboratory Medicine
University of Pennsylvania
Philadelphia, PA, USA
zhi.huang@pennmedicine.upenn.edu

1 INTRODUCTION

Hematoxylin and eosin (H&E) staining provides detailed tissue morphology but lacks molecular specificity. Multiplexed spatial proteomics (SP) addresses this gap by measuring dozens of markers in situ, capturing molecular patterns, cellular composition, and cell–cell interactions at single-cell resolution (Giesen et al. (2014), Black et al. (2021)). Integrating H&E's morphological detail with SP's molecular specificity could enable richer tissue characterization to uncover disease mechanisms and improve biomarker discovery, as demonstrated in integrative spatial transcriptomics analysis (Shulman et al. (2025)). Textual information, including protein descriptions, cell-type annotations, and clinical metadata, can augment these analyses by grounding learned embeddings in established domain knowledge and enabling semantic queries (Schaefer et al. (2024)).

Recent self-supervised and vision-language models have generated robust morphological features from H&E to predict therapeutic biomarkers (Chen et al. (2024), Vorontsov et al. (2024)) and enabled cross-modal retrieval to empower search capabilities for clinical and educational use (Huang et al. (2023), Lu et al. (2024)). In SP, graph-based methods are employed to model cell–cell interactions and cellular neighborhoods (Wu et al. (2022)). Other work focuses on representation modeling directly from multiplexed SP images, employing convolutional or vision transformer architectures to embed multichannel fluorescence signals for retrieval tasks (Yu et al. (2023)). Early attempts to bridge SP and H&E have focused on predicting specific proteins from morphology alone, or integrating a limited set of markers for narrow tasks (Wu et al. (2024), Wu et al. (2023)). Although recent efforts aim at more general purpose SP embeddings through self-supervised learning (Wenckstern et al. (2025)), most studies rarely unify SP, H&E, and textual metadata in a single representation space, nor address heterogeneity in SP panels. To address these gaps, we propose a vision-language framework that combines a generalizable SP encoder capable of handling diverse marker sets with

a joint trimodal alignment strategy leveraging complementary features from SP, H&E, and text. By embedding explicit biological and clinical descriptors into natural language, our approach enriches learned representations with domain knowledge, while enabling intuitive multimodal retrieval. This single integrated framework captures molecular signals, morphological context, and biomedical semantics on multiple scales, thereby facilitating advanced discovery and translational applications.

2 METHODOLOGY

2.1 DATA CURATION

We curated CODEX spatial proteomics (Black et al. (2021)), aligned H&E imaging (either from the same or adjacent tissue slice), cell-type annotations, and clinical metadata from 64 studies consisting of a total of 3,925 tissue regions (Enable Medicine (2024)). Due to the high dimensionality of histology data, we cropped the regions into patches, resulting in over one million samples. We next developed an automated captioning pipeline to incorporate protein information, marker expression, cell-type annotations, and clinical metadata into natural language descriptions for each sample.

2.2 ENCODER ARCHITECTURES FOR MULTIMODAL ALIGNMENT

To create a unified representation space that jointly captures molecular, morphological, and textual information, we employ separate unimodal encoders for SP, H&E, and text. Each encoder produces a modality-specific embedding, which is then mapped to a shared 2048-dimensional space using feed-forward projection layers.

To handle heterogeneous marker panels in spatial proteomics and capture high-dimensional protein expression patterns, we propose a multistage transformer model trained from scratch. First, each marker channel is encoded into a vector representation using a vision transformer to capture spatial dependencies. Next, the channel embeddings are fused with an associated marker specific embedding initialized using a protein language model via summation (ESM Team (2024) ,Wenckstern et al. (2025)). Finally, the fused representations are input into a transformer encoder to model the inter-marker interactions.

We fine-tune a pretrained histology foundation model, PLIP, for the H&E modality (Huang et al. (2023)). Fine-tuning preserves the robust morphological features learned during pretraining while adapting the encoder to the specific nuances of our dataset.

We leverage PubMedBERT, a domain specific language model, for textual inputs (Gu et al. (2022)). To balance stability and adaptability, we only unfreeze the final two transformer layers for fine-tuning. This encourages the model to retain general biomedical linguistic knowledge while allowing the later layers to learn task-specific context.

After the unimodal encoders produce their respective embeddings, we align them in the joint-embedding space. The training objective consists of pair-wise contrastive losses between the modalities with an optional context-aware weighting scheme. Specifically, we follow the standard CLIP-style loss formulation between all possible pairs of modality and take the average (Radford et al. (2021)). Additionally, because some samples in a mini-batch may come from the same region, and thus may not be truly independent, we introduce an optional weighting scheme to down-weight the negative pairs originating from the same region.

2.3 EVALUATION AND PRELIMINARY RESULTS

Thus far we have conducted a pilot study using 648 regions from two cancer studies (head and neck cancer, colorectal cancer) consisting of approximately 80,000 total patches. We split the data at the region level into train and evaluation sets. We compare H&E performance to the baseline PLIP model, and CODEX performance to a baseline representation that consists of the mean expression value for each channel.

To evaluate our framework, we perform tasks at the patch and patient levels. At the patch level, we test the framework's ability to capture local cellular composition by retrieving patches with similar cell-type distributions, using the learned embedding space to measure similarity and validate against

ground-truth annotations. Our model improves (decreases) the average mean squared error of the top five retrieved cell composition vectors for H&E from 0.39 to 0.37. However, for CODEX we observe no improvement over the baseline. We also demonstrate a zero-shot classification capability by prompting the text encoder with dominant cell-type annotations and identify the associated image patches based on embedding similarities. For example, the model's average precision for tumor cells is 0.94 for H&E and 0.79 for CODEX compared to a random average precision of 0.32.

For patient-level evaluation, we use an attention-based multiple-instance learning (MIL) strategy in which patch-level embeddings are aggregated within each patient, and a classifier is trained to predict clinical phenotypes. Predicting human papillomavirus (HPV) status, a prognostic biomarker in head and neck cancer, our results show that the area under the precision recall curve improves from 0.90 to 0.92 for H&E and from 0.89 to 0.94 for CODEX. We further extend this to a retrieval task by computing patient-level similarity scores to identify clinically similar patients where we see improvement in CODEX based retrieval, however not for H&E.

3 Conclusion

Our preliminary results suggest that a trimodal framework for integrating molecular, morphological, and textual signals in histopathology has significant potential. Future research will expand to the entire dataset, allowing for a more comprehensive evaluation of our approach. Additionally, we will explore self-supervised learning for SP and refine alignment strategies. Post alignment, exploring multimodal generative large language models may unlock more advanced cross-modal capabilities, enabling deeper knowledge discovery and facilitating new clinical applications.

Meaningfulness Statement

A meaningful representation of life seamlessly links molecular functions at the protein level to macroscale structures of tissue and organ systems, integrating higher-order biological or clinical narratives. By aligning multiplexed spatial proteomics, H&E histology, and clinical phenotype data into a single embedding space, our work bridges the microscopic morphology of cells and tissues with macroscopic patient-level context. This context-aware representation learning encourages sharing of multi-scale biological information innate in the data to enhance the global representation of life.

References

Sarah Black, Darci Phillips, John W. Hickey, Julia Kennedy-Darling, Vishal G. Venkataraaman, Nikolay Samusik, Yury Goltsev, Christian M. Schürch, and Garry P. Nolan. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nature Protocols*, 16(8):3802–3835, August 2021. ISSN 1750-2799. doi: 10.1038/s41596-021-00556-8. URL https://www.nature.com/articles/s41596-021-00556-8. Publisher: Nature Publishing Group.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL https://www.nature.com/articles/s41591-024-02857-3. Publisher: Nature Publishing Group.

Enable Medicine. Enable Medicine Platform, 2024. https://www.enablemedicine.com/ [Accessed: 2024].

ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale Website, December 4 2024. URL https://evolutionaryscale.ai/blog/esm-cambrian.

Charlotte Giesen, Hao A O Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J Schüffler, Daniel Grolimund, Joachim M Buhmann, Simone Brandt, Zsuzsanna Varga, Peter J Wild, Detlef Günther, and Bernd Bodenmiller. Highly multiplexed imaging of

tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4):417–422, March 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2869. URL http://dx.doi.org/10.1038/nmeth.2869.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3 (1):1–23, January 2022. ISSN 2691-1957, 2637-8051. doi: 10.1145/3458754. URL http://arxiv.org/abs/2007.15779. arXiv:2007.15779 [cs].

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical Twitter. *Nature Medicine*, 29(9):2307–2316, September 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02504-3. URL https://www.nature.com/articles/s41591-023-02504-3. Publisher: Nature Publishing Group.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, March 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02856-4. URL https://www.nature.com/articles/s41591-024-02856-4. Publisher: Nature Publishing Group.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Moritz Schaefer, Peter Peneder, Daniel Malzl, Anna Hakobyan, Varun S. Sharma, Thomas Krausgruber, Jörg Menche, Eleni Tomazou, and Christoph Bock. Joint embedding of transcriptomes and text enables interactive single-cell RNA-seq data exploration via natural language. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024. URL https://openreview.net/forum?id=yWiZaE4k3K.

Eldad D. Shulman, Emma M. Campagnolo, Roshan Lodha, Amos Stemmer, Thomas Cantore, Beibei Ru, Tian-Gen Chang, Sumona Biswas, Saugato Rahman Dhruba, Andrew Wang, Rohit Paul, Sarath Kalisetty, Tom Hu, Maclean Nasrallah, Sheila Rajagopal, Stephen-John Sammut, Stanley Lipkowitz, Peng Jiang, Carlos Caldas, Simon Knott, Danh-Tai Hoang, Kenneth Aldape, and Eytan Ruppin. Ai-driven spatial transcriptomics unlocks large-scale breast cancer biomarker discovery from histopathology. *bioRxiv*, 2025. doi: 10.1101/2024.10.16.618609. URL https://www.biorxiv.org/content/early/2025/02/22/2024.10.16.618609.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, October 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-03141-0. URL https://www.nature.com/articles/s41591-024-03141-0. Publisher: Nature Publishing Group.

Johann Wenckstern, Eeshaan Jain, Kiril Vasilev, Matteo Pariset, Andreas Wicki, Gabriele Gut, and Charlotte Bunne. AI-powered virtual tissues from spatial proteomics for clinical diagnostics and biomedical discovery, January 2025. URL http://arxiv.org/abs/2501.06039. arXiv:2501.06039 [q-bio].

Eric Wu, Alexandro E Trevino, Zhenqin Wu, Kyle Swanson, Honesty J Kim, H Blaize D'Angio, Ryan Preska, Aaron E Chiou, Gregory W Charville, Piero Dalerba, Umamaheswar Duvvuri, Alexander D Colevas, Jelena Levi, Nikita Bedi, Serena Chang, John Sunwoo, Ann Marie Egloff, Ravindra Uppaluri, Aaron T Mayer, and James Zou. 7-UP: Generating in silico CODEX

from a small set of immunofluorescence markers. *PNAS Nexus*, 2(6):pgad171, June 2023. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgad171. URL https://doi.org/10.1093/pnasnexus/pgad171.

Eric Wu, Matthew Bieniosek, Zhenqin Wu, Nitya Thakkar, Gregory W. Charville, Ahmad Makky, Christian Schürch, Jeroen R. Huyghe, Ulrike Peters, Christopher I. Li, Li Li, Hannah Giba, Vivek Behera, Arjun Raman, Alexandro E. Trevino, Aaron T. Mayer, and James Zou. ROSIE: AI generation of multiplex immunofluorescence staining from histopathology images, November 2024. URL https://www.biorxiv.org/content/10.1101/2024.11.10.622859v1. Pages: 2024.11.10.622859 Section: New Results.

Zhenqin Wu, Alexandro E. Trevino, Eric Wu, Kyle Swanson, Honesty J. Kim, H. Blaize D'Angio, Ryan Preska, Gregory W. Charville, Piero D. Dalerba, Ann Marie Egloff, Ravindra Uppaluri, Umamaheswar Duvvuri, Aaron T. Mayer, and James Zou. Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nature Biomedical Engineering*, 6(12):1435–1448, November 2022. ISSN 2157-846X. doi: 10.1038/s41551-022-00951-w. URL https://www.nature.com/articles/s41551-022-00951-w.

Jennifer Yu, Zhenqin Wu, Aaron T. Mayer, Alexandro Trevino, and James Zou. A Multi-Granularity Approach to Similarity Search in Multiplexed Immunofluorescence Images, November 2023. URL https://www.biorxiv.org/content/10.1101/2023.11.26.568745v1. Pages: 2023.11.26.568745 Section: New Results.