

Renal Cell Type and State Estimation in Brightfield Histology Images: A Pilot Study on Diabetic Nephropathy

Jamie L. Fermin^a, Samuel Border^b, Ahmed Naglah^c, Benjamin Shickel^c, Patricio S. La Rosa^d, John E. Tomaszewski^e, Sanjay Jain^f, Tarek M. El-Achkar^g, Michael T. Eadon^g, and Pinaki Sarder^{c,h*}

^aDept. of Electrical and Computer Engineering, Univ. of Florida, Gainesville, FL

^bDept. of Biomedical Engineering, Univ. of Florida, Gainesville, FL

^cUniv. of Florida Intelligent Critical Care Center, Gainesville, FL

^dSeed Production Innovation, Crop Science Division, Bayer Company, St. Louis, MO

^eDept. of Pathology & Anatomical Sciences, Univ. at Buffalo, Buffalo, NY

^fDiv. of Nephrology, Dept. of Medicine, Washington Univ. School of Medicine, St. Louis, MO

^gDiv. of Nephrology, Dept. of Medicine, Indiana Univ. School of Medicine, Indianapolis, ID

^hDiv. of Nephrology, Hypertension, and Renal Transplantation – Quantitative Health Section, Dept. of Medicine, Univ. of Florida, Gainesville, FL

ABSTRACT

Multi-omics data, such as 10X Genomics Visium (spatial transcriptomics), measure gene expressions, molecular pathway activities, and can predict cell types/states but are often expensive and inaccessible in clinical settings. Thus, despite the emergence of multi-omics technologies, histopathological assessments under brightfield microscopy remain the diagnostic gold standard. In this work, we examine machine learning-based pipelines for predicting cell types/states from brightfield histology images using state-of-the-art (SOTA) deep learning (DL) models, aiming to enhance diagnostics and prognostics in clinical medicine.

Our proposed pipeline consists of two stages: (1) an Image-To-Text retrieval Network (ITTN) that leverages the CONtrastive learning from Captions for Histopathology (CONCH) model to assign histopathological text prompt from brightfield histology image, and (2) a Vision Language Model (VLM), which is built on the same CONCH model used in ITTN but incorporates a regression head to predict cell type/state proportions based on the paired image and text inputs. During training, we classify the image into one of four structural types (glomerulus, tubules, vessels, and interstitium) using the ITTN. These classification labels are then used to construct a new text prompt with a suitable histopathological description for each image in the test set. The new text prompt and raw image are used as paired inputs to the VLM to predict cell types/states. We also utilize SOTA models, such as CONCH (using only the vision encoder), ViT, and ResNet, which employ image-only inputs in separate regression pipelines.

We experimented and tested our proposed pipelines on a set of 10X Visium formalin-fixed paraffin-embedded whole slides images of diabetic nephropathy samples collected at Indiana University. Our experiments yielded a mean squared error of 0.0027 for the proposed pipeline, showing improvements of 20.59%, 27.03%, and 32.50% over CONCH (image only), ViT, and ResNet, respectively. The proposed pipeline aims to bridge the gap between traditional histopathology and molecular diagnostics, enhancing disease diagnosis and prognosis.

Keywords: Diabetic nephropathy, digital pathology, spatial transcriptomics, gene expression, regression, foundation model.

* Send correspondence to Pinaki Sarder, PhD; E-mail: pinaki.sarder@medicine.ufl.edu

1. INTRODUCTION

Despite the emergence of multi-omics technologies, histopathological assessments of tissue samples under bright-field microscopy remain a gold standard for the diagnosis of many diseases. Multi-omics data, such as 10X Genomics Visium data (spatial transcriptomics [ST]), is utilized in research and pre-clinical settings as they measure gene expressions, determine the activities of molecular pathways, and predict the populations of cell types/states. However, these technologies are expensive and often inaccessible in clinics, limiting usability.

In this study, we are using foundation models (FMs) to estimate renal cell types and states from histological whole slide images (WSIs) of renal tissue. FMs are large-scale machine learning models trained on extensive and diverse datasets, allowing them to agnostically learn generalizable features across various applications. Specifically, in histopathology, these models have been trained on a vast amount of data from different organs, allowing them to accurately recognize and differentiate key anatomical structures, such as glomeruli and tubules, in the context of the kidney.

We used human samples with formalin-fixed paraffin embedded (FFPE) in diabetic nephropathy (DN) to estimate these cell types/states. DN induces different pathological alterations in kidney tissues, such as the glomeruli. In the glomerulus, DN causes significant changes such as glomerular hypertrophy, mesangial expansion, and the formation of nodular glomerulosclerosis. These changes are clearly visible in the pixel space, which could provide a valuable context to learn the organizations of pixels.

By analyzing the spatial organization of these visible alterations in pixel space, we aim to determine whether this information can be indicative of the underlying molecular data. Essentially, we want to learn how these pixels are organized to estimate the cell type/state proportions accurately.

2. METHODS

In this study, we used a dataset consisting of nine paired brightfield WSI and 10X Visium¹ ST data from FFPE human kidney tissues, including five reference samples and four DN samples acquired from Indiana University.

The 10X Visium ST data contains “spot” (55 μm diameter regions of interest) of transcript count matrices, with a distance of 100 μm between spot centroids. These matrices can be converted into proportions of cell subtypes and states using a method called cell deconvolution.² Each spot captures near-whole transcriptome RNA (20K genes/sample) from about 10 or more cells. In the following, we discuss our methods to estimate cell type and states from brightfield histology using a state-of-the-art vision model followed by our proposed pipeline for such estimation.

Cell type proportion prediction using vision models: We fine-tuned three models: ResNet,³ Vision Transformer (ViT),⁴ and CONtrastive learning from Captions for Histopathology (CONCH).⁵ These models are FMs, pre-trained on large datasets. ResNet and ViT are pretrained on ImageNet,⁶ while CONCH is pretrained on the largest histopathology-specific vision-language dataset, consisting of 1.17 million image-caption pairs.

Each WSI was segmented into 512×512 sized tiles around each spot with overlap. Each tile, denoted as \mathbf{x}_i , is processed by an encoder model f_θ , with pretrained parameters θ , to extract a d -dimensional feature vector $\mathbf{z}_i = f_\theta(\mathbf{x}_i)$. These feature vectors are then fed into a multilayer perceptron (MLP) to predict the cell type proportions. The MLP processes the feature vectors and outputs a 74-dimensional vector \mathbf{y}_i corresponding to the proportions of the 74 different renal cell subtypes as listed in Table 1.

Cell type proportion prediction using the proposed pipeline: The proposed pipeline is detailed in Fig. 1. An image \mathbf{x}_i is passed to the Image-To-Text retrieval Network (ITTN) together with four prompts $\{t_1, t_2, t_3, t_4\}$. The text prompts that we used are “glomerulus”, “tubules”, “vessels”, and “interstitium”. The CONCH model extracts feature representations for both the image and text prompts. Specifically, the image encoder f_{θ_x} generates a normalized d -dimensional feature embedding $\mathbf{z}_{x_i} = \frac{f_{\theta_x}(\mathbf{x}_i)}{\|f_{\theta_x}(\mathbf{x}_i)\|}$. Similarly, the text encoder f_{θ_T} generates normalized d -dimensional feature embeddings for each prompt $\mathbf{z}_{t_j} = \frac{f_{\theta_T}(t_j)}{\|f_{\theta_T}(t_j)\|}$ for $j = 1, 2, 3, 4$.

We then calculate the cosine similarity score s between the normalized image embedding, and each normalized text embedding:

Table 1: Main Types and Subtypes of Renal Cells

Main Types	Sub Types
Proximal Tubule (PT)	PT, PT-S1, PT-S2, PT-S3, aPT, cycPT, dPT, dPT/DTL
Descending Thin Limb (DTL)	DTL, DTL1, DTL2, DTL3, dDTL3
Ascending Thin Limb (ATL)	ATL, dATL
Thick Ascending Limb (TAL)	TAL, aTAL1, aTAL2, M-TAL, dM-TAL, cTAL, dc-TAL, MD
Distal Convoluted Tubule (DCT)	DCT, DCT1, DCT2, dDCT, cycDCT
Connecting Tubule (CNT)	CNT, CNT-PC, dcCNT, cycCNT
Principal Cell (PC)	PC, C-PC, CCD-PC, OMCD-PC, M-PC, dOMCD-PC, dM-PC, IMCD, dIMCD
Intercalated Cell (IC)	IC, C-IC-A, CCD-IC-A, CNT-IC-A, dC-IC-A, OMCD-IC-A, M-IC-A, tPC-IC, IC-B
Papillary Epithelium (PapE)	PapE
Podocyte (POD)	POD, dPOD
Parietal Epithelial Cell (PEC)	PEC
Endothelial Cell (EC)	EC, EC-GC, EC-AEA, EC-DVR, EC-PTC, dEC-PTC, EC-AVR, dEC, cycEC, EC-LYM
Vascular Smooth Muscle/Pericyte (VSM/P)	VSM/P, MC, REN, VSMC, VSMC/P, dVSMC
Fibroblast (FIB)	FIB, MYOF, cycMYOF, M-FIB, dM-FIB, aFIB, dFIB
Immune Cell (IMM)	IMM, B, PL, T, NKT, MAST, MAC-M2, cycMNP, MDC, cDC, pDC, ncMON, N
Neural-like (NEU)	NEU, SC/NEU

$$s_{ij} = \cos(\mathbf{z}_{x_i}, \mathbf{z}_{t_j}) = \frac{\mathbf{z}_{x_i} \cdot \mathbf{z}_{t_j}}{\|\mathbf{z}_{x_i}\| \|\mathbf{z}_{t_j}\|} \quad (1)$$

The prompt t_j with the highest similarity score is selected as $t_{i,\hat{j}}$:

$$\hat{j} = \arg \max_{j \in \{1,2,3,4\}} s_{ij} \quad (2)$$

This selected prompt is then used to construct a new, detailed text prompt, t_{ij} , for further processing in the VLM. The VLM is built on the same pretrained CONCH model used in ITTN, but incorporates a regression head to predict cell type proportions. The input image \mathbf{x}_i and the detailed text prompt t_{ij} are both passed into the VLM. The image encoder f_{θ_x} processes the image tile to generate a feature embedding \mathbf{z}_{x_i} , while the text encoder f_{θ_T} processes the text prompt to produce a feature embedding $\mathbf{z}_{t_{ij}}$. These embeddings are concatenated and passed through an MLP in the VLM to predict the cell type/states proportion.

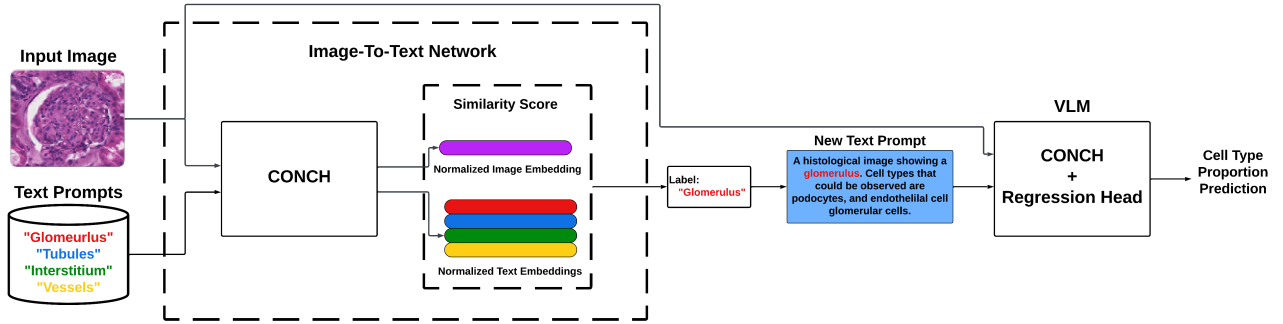


Fig. 1: Our proposed pipeline for predicting cell type/states proportions in kidney tissue images. An input histopathological image along with predefined text prompts are first processed by an ITTN to classify the kidney structure present. This classification is determined by calculating similarity scores between the normalized image embedding and normalized text embeddings for provided prompts. The text prompt with the highest similarity score is selected to construct a new detailed text prompt. This new text prompt, along with the input image, is then fed into a VLM, which uses this combined information to predict the proportions of various cell types/states.

3. EXPERIMENTS AND RESULTS:

All models were trained for 25 epochs with a batch size of 1024 using the Adam optimizer (learning rate 1×10^{-3}) and a learning rate scheduler with a decay factor of 0.1. Mean squared error (MSE) was used as a loss function.

Training was carried out on a single NVIDIA A100 GPU with 40 GB of RAM. Of the nine samples, seven were used for training and the remaining two were heldout for validation.

Model performance was evaluated using MSE, defined as:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M (\hat{y}_{ij} - y_{ij})^2 \quad (3)$$

where N is the number of samples, M is the number of cell subtypes, \hat{y}_{ij} is the predicted proportion of subtype j in sample i , and y_{ij} is the ground truth proportion.

From Table 2, we observe that the FM CONCH (Vision) achieved an MSE of 0.0034, which is lower than the MSEs of ViT (0.0037) and ResNet (0.0040). This experiment underscores the robustness of CONCH (Vision) in handling image-only data, as it outperforms models pretrained on ImageNet data. This superior performance can be attributed to the large domain gap between natural images and histopathological images. Models pretrained on histopathology-specific data, such as CONCH, are better equipped to capture the intricate patterns and textures unique to medical images, leading to more accurate predictions of cell type proportions. Table 3 shows the benefit of combining image and text data. The proposed pipeline (VLM) achieved the lowest MSE of 0.0027, underscoring the improved performance of VLM over vision-only models.

Model predictions for POD in DN samples reveal that VLM ($\mu = 0.35$, $\sigma = 0.15$) and CONCH (Vision) ($\mu = 0.22$, $\sigma = 0.10$) provide broader distributions with higher means compared to the ground truth, though VLM shows greater variability. ViT ($\mu = 0.20$, $\sigma = 0.09$) also indicates high variability but with a slightly lower mean, while ResNet ($\mu = 0.05$, $\sigma = 0.03$) consistently underestimates POD levels. For EC-GC, CONCH (Vision) ($\mu = 0.38$, $\sigma = 0.15$), and ViT ($\mu = 0.50$, $\sigma = 0.21$) predict higher variability with means closer to ground truth ($\mu = 0.58$, $\sigma = 0.14$) than VLM ($\mu = 0.29$, $\sigma = 0.18$), and ResNet ($\mu = 0.15$, $\sigma = 0.10$), which tend to underestimate these cell types.

Table 2: Cell Type Proportion Prediction Performance Comparing Different Vision Models

Metric	CONCH (Vision)	ViT	ResNet
MSE	0.0034	0.0037	0.0040

Table 3: Cell Type Proportion Prediction Performance Comparing Vision + Language vs. Only Vision foundation models

Metric	VLM	CONCH (Vision)
MSE	0.0027	0.0034

Fig. 2 illustrates the distribution of cell type proportions for podocytes (POD) and endothelial cell-glomerular cell (EC-GC) across all Visium spots intersecting with the glomeruli in both DN and reference samples. The ground truth distribution for podocytes (Fig. 2a) indicates significantly lower levels in DN samples ($\mu = 0.14$, $\sigma = 0.05$) compared to reference samples (Fig. 2b) ($\mu = 0.87$, $\sigma = 0.18$), reflecting reduced podocyte presence in diseased tissue. For EC-GC, DN samples show higher and more variable levels ($\mu = 0.58$, $\sigma = 0.14$), while reference samples exhibit minimal levels ($\mu = 0.03$, $\sigma = 0.06$), highlighting the pathological changes in DN. This observation supports earlier literature findings.⁷

In reference samples, VLM ($\mu = 0.68$, $\sigma = 0.25$) and CONCH (Vision) ($\mu = 0.66$, $\sigma = 0.27$) closely align with the ground truth for POD levels ($\mu = 0.87$, $\sigma = 0.18$), while ViT ($\mu = 0.56$, $\sigma = 0.31$) shows high variability and slight underestimation, and ResNet ($\mu = 0.20$, $\sigma = 0.08$) significantly underestimates. For EC-GC, all models slightly overestimate compared to ground truth ($\mu = 0.03$, $\sigma = 0.06$), with VLM ($\mu = 0.08$, $\sigma = 0.07$) and ViT ($\mu = 0.11$, $\sigma = 0.06$) showing more variability than CONCH (Vision) ($\mu = 0.06$, $\sigma = 0.03$) and ResNet ($\mu = 0.04$, $\sigma = 0.03$).

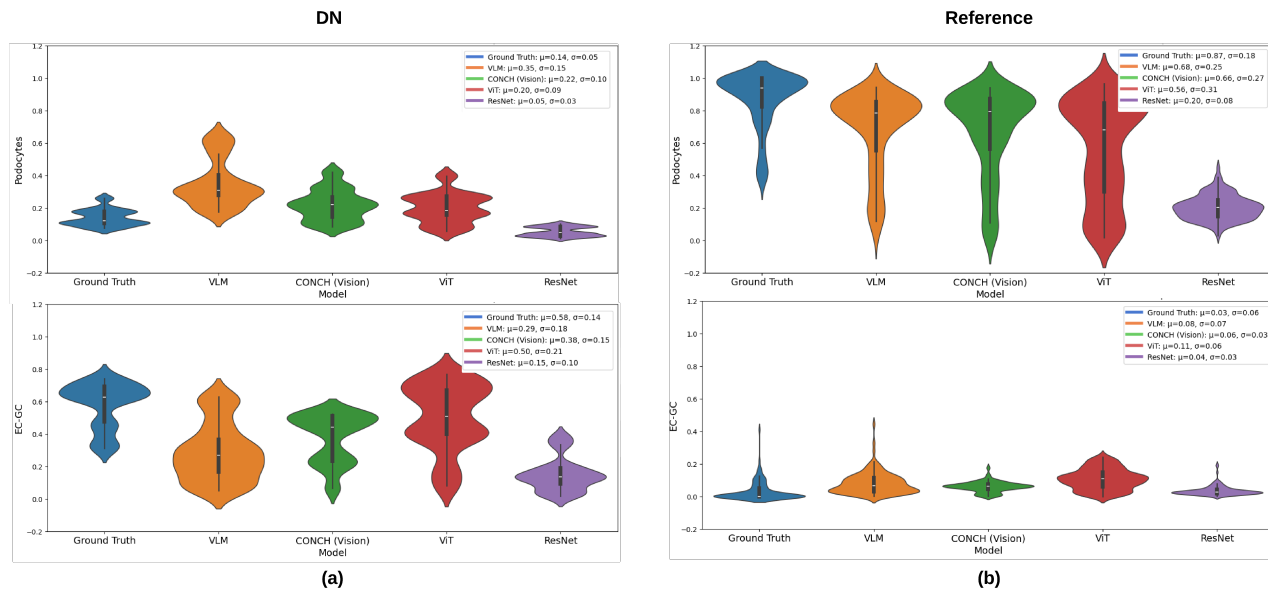


Fig. 2: We illustrate the cell type proportion distributions for podocytes (POD) and endothelial cell-glomerular cell (EC-GC) interactions from Visium spots that intersect with the glomeruli. **(2a)** depicts this for DN samples, while **(2b)** presents these distributions in reference samples. The figures compare the ground truth measurements with the predictions obtained from the four models, with μ and σ indicating the mean and standard deviation, respectively.

While VLM did not perform as well as the ViT and CONCH models in predicting POD and EC-GC in DN samples, it exhibited better sensitivity in identifying overall cell type proportions, as indicated by its lower overall MSE value. This increased sensitivity could be attributed to the text information incorporated by the VLM, which includes cell type information based on the expected cell types that can be found inside a particular kidney structure. However, these findings warrant further detailed examination in future studies.

4. CONCLUSIONS

In this study, we demonstrated the potential of using FMs tailored to histopathology to estimate cell type and state proportions. By examining DN FFPE human samples with clearly visible pathological alterations, we were able to successfully map these changes in the pixel space to their corresponding cell type and state proportions. Our results also show that leveraging multimodal image and text data, may offer superior results compared to using image data alone, opening up avenues to study the advantage of integrating multimodal features for enhanced predictive accuracy.

5. ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health [NIH/NIDDK R01 DK114485].

REFERENCES

- [1] Bergenstråhle, J., Larsson, L., and Lundeberg, J., “Seamless integration of image and molecular analysis for spatial transcriptomics workflows,” *BMC genomics* **21**, 1–7 (2020).
- [2] Melo Ferreira, R., Freije, B. J., and Eadon, M. T., “Deconvolution tactics and normalization in renal spatial transcriptomics,” *Frontiers in Physiology* **12**, 812947 (2022).
- [3] He, K., Zhang, X., Ren, S., and Sun, J., “Deep residual learning for image recognition,” in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).
- [4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929* (2020).
- [5] Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L. P., Gerber, G., et al., “A visual-language foundation model for computational pathology,” *Nature Medicine* **30**(3), 863–874 (2024).
- [6] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., “Imagenet: A large-scale hierarchical image database,” in [*2009 IEEE conference on computer vision and pattern recognition*], 248–255, Ieee (2009).
- [7] Albrecht, M., Sticht, C., Wagner, T., Hettler, S. A., De La Torre, C., Qiu, J., Gretz, N., Albrecht, T., Yard, B., Sleeman, J. P., et al., “The crosstalk between glomerular endothelial cells and podocytes controls their responses to metabolic stimuli in diabetic nephropathy,” *Scientific Reports* **13**(1), 17985 (2023).