Connecting Gene Expression and Tissue Morphology with Conditional Generative Models

Frederieke Lohmann [1 2] **Alberto Valdeolivas** [1] **Jelica Vasiljević** [1]

Abstract

Inferring tissue morphology from gene expression remains widely unexplored. We present a two-stage conditional generative framework that leverages for the first time spatial transcriptomics data from the Visium HD platform to demonstrate this inference is feasible. Starting from near-whole-transcriptome profiles, the model synthesizes histology-like images that are plausible, as validated by FID scores and expert review. Model interpretation further reveals biologically meaningful links between specific genes and morphological patterns.

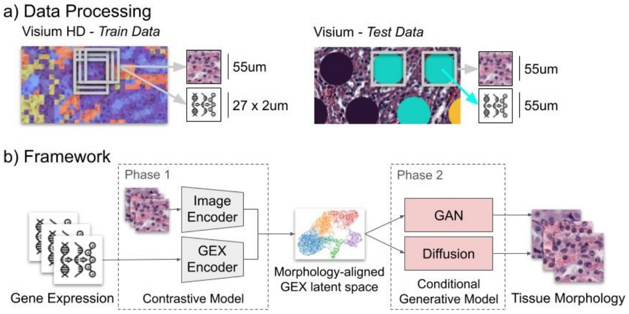


Figure 1. a) Data Processing strategy and b) two-phase modeling framework

Introduction

Spatial Transcriptomics (ST) is an innovative technology that allows the capture of gene expression (GEX) profiles within their natural tissue environment. Uncovering the relationship between GEX and morphology is crucial for tasks such as predicting the structural impact of gene perturbations and detecting early morphological changes linked to disease onset. Yet, it has received limited attention in existing literature. Recent studies have attempted to link gene expression with morphology using bulk RNA-seq data (Carrillo-Perez et al., 2023; 2024), a strategy that obscures intra-sample heterogeneity critical for spatial arrangement insights. Alternative approaches infer cellular morphology either from single-gene perturbations (Navidi et al., 2024) or from narrowly focused gene panels (Wu & Koelzer, 2024; Wu et al., 2023), limiting their scope. To fill this gap, our study leverages recent advancements in ST—specifically the 10x Visium HD platform—to model this relationship across the entire transcriptome (Hahn et al., 2025). Inspired by the effectiveness of generative models in text-to-image applications, we developed a two-stage conditional generative framework that synthesises histological images conditionally on GEX. Our model produces plausible histology-like images, confirmed by competitive FID scores and expert pathologist reviews. Interpretability analyses further point to associations between certain GEX patterns and distinct morphological traits, setting the stage for future work to identify genes whose disruption may contribute to disease or drug-induced toxicity. To our knowledge, this is the first approach to integrate high-resolution ST and conditional generative modelling to predict histological images. The code and data is available at https://github.com/lohmannf/genes2morphology.

Method

In this work, we characterise the mapping from GEX profiles to tissue morphology as a one-to-many mapping and model it using a two-phase conditional generative approach. The workflow is depicted in Figure 1b.

Phase 1: Contrastive Representation Learning. In the first step, we align tissue morphology and GEX by mapping them into a common representation space. The goal of this phase is dimensionality reduction of the GEX profiles to avoid overfitting in the second phase. While selecting highly variable genes (HVGs) or spatially variable genes (SVGs) is common practice, they have to be determined across all training and testing samples to be meaningful, limiting generalizability to new unseen samples. To mitigate this issue,

[1] Roche Pharma Research & Early Development, Data & Analytics, F. Hoffmann-La Roche Ltd, Roche Innovation Center Basel, Basel, Switzerland [2] ETH Zürich, Zürich, Switzerland. Correspondence to: Frederieke Lohmann <flohmann@ethz.ch>.

Proceedings of the Workshop on Generative AI for Biology at the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

we instead take a contrastive approach to dimensionality reduction. Inspired by prior work (Xie et al., 2023; Min et al., 2024; Lee et al., 2024), we train a bimodal contrastive model composed of an image encoder and a GEX encoder that integrates the two modalities into a shared latent space. The GEX encoder is a Multi-Layer Perceptron (MLP) with a projection head, while the image encoder combines the fixed pathology foundation model UNI (Chen et al., 2024) with a trainable projection head. The representations of matching data pairs are aligned using the CLIP loss (Radford et al., 2021).

Phase 2: Conditional Generation. In the second phase, the learned GEX representations are used to condition an image generative model. This phase is agnostic to the specific conditional generative architecture. We illustrate this flexibility by applying it to two major types of generative models: Generative Adversarial Networks (GANs) and diffusion models. For GANs, we adapt the StyleGAN-T architecture (Sauer et al., 2023) as StyleGAN-G, while for diffusion models, we use Stable Diffusion (SD) (Rombach et al., 2022). Both models are conditioned on the GEX embeddings obtained with the fixed GEX encoder pretrained in the first phase. The StyleGAN-T architecture is modified so that the discriminator incorporates a feature extractor trained on histological images instead of ImageNet (Dosovitskiy et al., 2020; Filiot et al., 2024), and it is trained from scratch. Similarly to StyleGAN-T, to ensure alignment between the GEX condition and the generated image, the spherical distance between the GEX embedding and the generated image embedding obtained with the pretrained image encoder is used as an additional guidance loss term. On the other hand, due to computational costs, the SD model was fine-tuned (U-Net part) on histological image data while conditioning on the GEX embeddings following the approach of Navidi et al. (2024). To mirror the usage of CLIP guidance in the GAN-based approach and improve prompt alignment, we employ classifier-free guidance (Ho & Salismans, 2022) with different strengths w at inference time. Training details for both models are described in Appendix A.

Data. We conducted experiments on six formalin-fixed paraffin-embedded (FFPE) healthy mouse kidney samples, profiled using two spatial transcriptomics platforms: Visium HD (one sample) and standard Visium (five samples). Visium HD offers much higher spatial resolution, capturing gene expression in $2\mu m$ bins, compared to the $55\mu m$ spots of standard Visium. To leverage the high resolution of Visium HD samples, we aggregated adjacent $2\mu m$ bins into $\sim 300,000$ "pseudo-spots" by summing gene expression profiles over 27 × 27 $2\mu m$ bins with a stride of 5 bins. We retained 18,248 genes common to both platforms and normalised and log-transformed gene expression profiles per sample. To obtain GEX-morphology data pairs, the corresponding tissue morphology to each (pseudo-)spot is extracted from the Macenko stain-normalised (Macenko et al., 2009) WSI as a $55\mu m \times 55\mu m$ patch at the (pseudo-)spot's spatial location. We resize patches to 128×128 pixels to ensure uniform absolute resolution. The preprocessing workflow is illustrated in Figure 1a. Given the need for a large training dataset, Visium HD data were used for model training, while standard Visium samples served as test data. Additional details on datasets and preprocessing are provided in Appendix B.

Results

We performed various experiments to quantitatively and qualitatively evaluate the plausibility of the generated images. We also analyse the interpretability of the gene expression encoder and its relation to the biological relevance of the generated morphologies.

Table 1. FID and KID on 36528 generated images

MODEL	FID ↓	KID ↓
STYLEGAN-G	**18.41**	**0.0127(0.0015)**
SD ($w = 1$)	84.28	0.0666(0.0035)
SD ($w = 3$)	78.99	0.0784(0.0043)
SD ($w = 7$)	84.69	0.0925(0.0045)

Image Plausibility. Figure 2 shows histological patches generated using multiple noise vectors (columns) and gene expression inputs (rows). The GEX profiles corresponding to real patches (shown in the leftmost column) are used as input. The generated images accurately capture the tissue morphology of their corresponding reference images and can be visually compared to real tissue patches shown in Appendix C. Notably, the generated images more closely resemble the training data in hue and overall quality (Appendix C, top row), as the model does not explicitly account for experimental variations such as tissue preparation. To quantitatively measure the quality of generated images, Table 1 reports Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018) for both the GAN- and diffusion-based models. Since predicting tissue morphology from GEX using ST platforms such as Visium has not been previously addressed, there are no directly comparable methods. Therefore, we compare our results to the most closely related approaches. In this setting, StyleGAN-G achieves substantially better FID scores than prior GAN-based methods (e.g., RNA-GAN, FID = 83.89 (Carrillo-Perez et al., 2023)), while the diffusion-based model performs comparably to related diffusion methods (MorphoDiff, FID ≥ 78 (Navidi et al., 2024)). We attribute the relatively lower performance of SD-models in this setting to the limited generalizability of pretrained diffusion models to histopathology data, as noted in prior

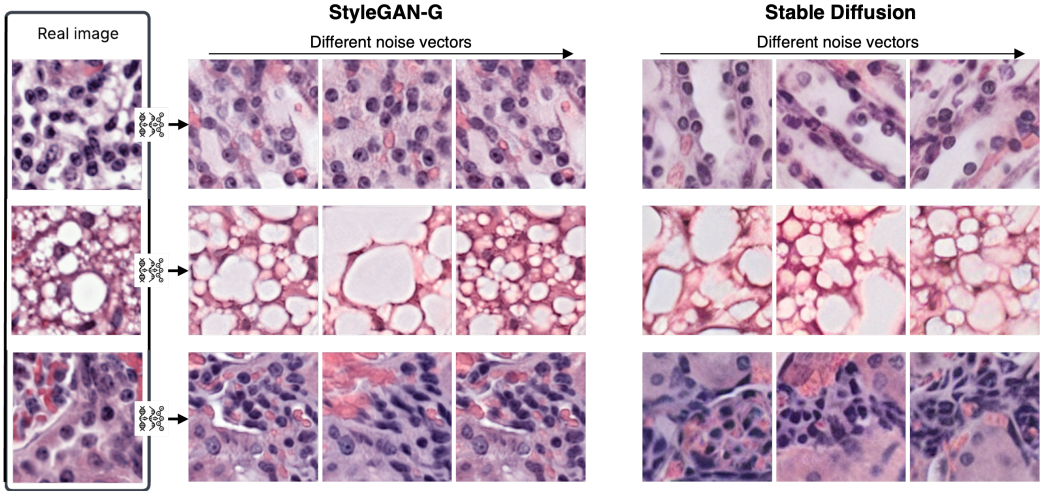


Figure 2. 128 × 128px images generated from different gene expression prompts with StyleGAN-G and Stable Diffusion ($w = 3$). Corresponding reference image shown in the leftmost column. Columns correspond to the same random noise vector.

Table 2. Average expert classification performance on 50 images

Accuracy ↑	Precision ↑	Recall ↑
0.486(0.095)	0.492(0.096)	0.488(0.111)

work (Müller-Franzes et al., 2023). Due to the substantial training requirements of dedicated diffusion models and the limited availability of data, we selected the StyleGAN-G architecture for subsequent experiments. However, as more ST datasets become available, diffusion models specifically tailored to histopathology may outperform the GAN-based approach. To further assess the morphological plausibility of the generated images, we conducted an expert validation study with a panel of 10 veterinary pathologists. We presented each expert with a morphologically diverse, balanced dataset of 50 real and generated images and asked to classify them. The average classification performance across the dataset is shown in Table 2. The expert panel achieved a near-random accuracy of 0.486 on the entire dataset, confirming that the images generated with StyleGAN-G are highly realistic. In Appendix D, we compare the distributions of real and synthetic images in the learned image embedding space of the contrastive model as an additional evaluation of generated image quality.

GEX Embedding Space. To assess the quality of the learned embedding space, we applied sample-wise Leiden clustering (Traag et al., 2019) on the GEX embedding space (GEX data projected using the pretrained gene expression encoder). For visualization, we combine all samples together and project them using UMAP (McInnes et al., 2018). The hyperparameters of the clustering are chosen to ensure the identification of key kidney regions within each sample, including the cortex (CX), outer stripe of the outer medulla (OSOM), brown adipose tissue (BAT), inner stripe of the outer medulla (ISOM), and inner medulla&papilla (IM&P) (Kumaran & Hanukoglu, 2024). Figure 3a+b presents the UMAP of the GEX embedding space. In Figure 3a, embeddings are color-coded by Leiden clustering results, while Figure 3b displays colors corresponding to sample origin. Figure 3d-h maps these clusters onto test set images, demonstrating successful identification of key kidney regions across all samples. The embedding space exhibits two important properties. GEX embeddings from different samples are well-integrated as demonstrated in Figure 3b. Additionally, clusters representing identical tissue regions from various samples consistently localize to similar areas within the latent space, as shown in Figure 3a. In contrast, in the UMAP space of the GEX data, a clear separation between samples can be observed (Figure 3c). To quantify the alignment in the embedding space, we performed joint clustering across all test samples both in the GEX embedding and GEX space and compared it to sample-wise clustering and sample index using the Adjusted Rand Index (ARI). A visualization of the joint clustering can be found in Appendix E. The results show a strong alignment with separately identified clusters (ARI = 0.79) and minimal correlation with the sample index (ARI = 0.005) in the embedding space. In contrast, the overlap between clustering and sample index in GEX space is significant (ARI = 0.588), indicating stronger batch effects.

Biological Relevance. Building on the observed alignment between GEX embedding space clusters and kidney regions (Figure 3), we aim to identify which genes primarily drive the differences between clusters and therefore relate to the underlying morphological distinctions. We employed the In-

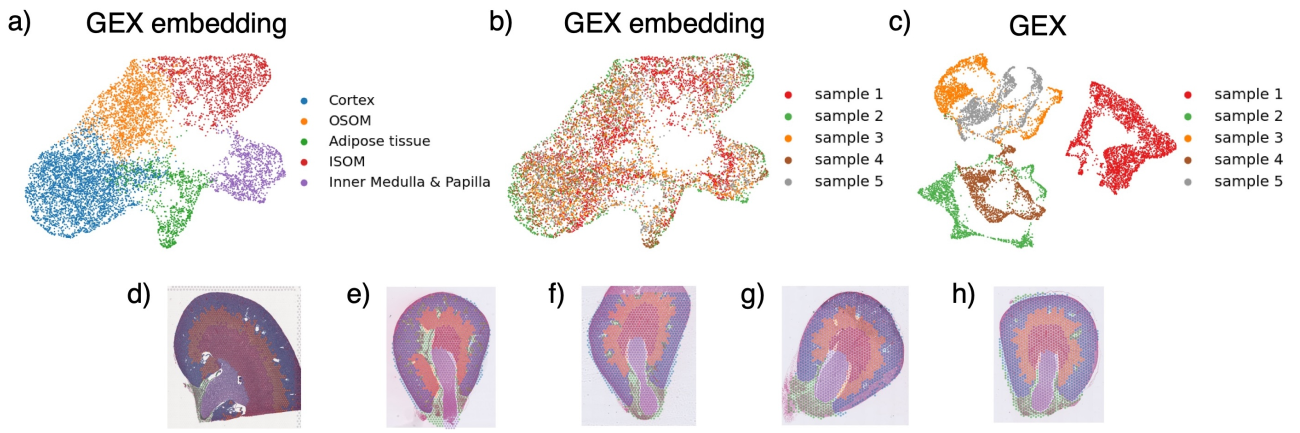


Figure 3. UMAP of the gene expression embedding space colored by (a) clusters found on each sample individually and (b) sample index. (c) UMAP of the gene expression space of samples 1-5 colored by sample index. Tissue cuts with spots colored by clusters found in gene expression embedding space for samples (d) 1, (e) 2, (f) 3, (g) 4, and (h) 5. Colors in (a) and (d)-(h) refer to the same clusters.

Table 3. Significantly enriched GO:CC terms in sample 4 with more than 4 leading genes.

Term	NES ↑	Query	Reference
Lipid Droplet	1.710422	BAT	IM&P
Basolateral Plasma Membrane	1.702832	CX	BAT
Basolateral Plasma Membrane	1.660759	CX	OSOM
Endoplasmatic Reticulum Lumen	2.098712	IM&P	OSOM
Intracellular Organelle Lumen	2.002516	IM&P	OSOM
Collagen-Containing Extracellular Matrix	1.936724	IM&P	OSOM

tegrated Gradients (IG) method (Sundararajan et al., 2017), as implemented in Heimberg et al. (2024), to identify genes that contribute most significantly to differences between GEX embedding clusters. This method assigns an attribution score to each input feature (in this case, each gene) by quantifying its importance along a continuous path between two points in the embedding space. Therefore, we applied IG to assess gene contributions between pairs of cluster-averaged expression profiles as identified in Figure 3d-h. This approach allowed us to determine which genes most significantly drive the differences between clusters in the gene expression embedding space. Our analysis involved the following steps: 1) We averaged the GEX embedding across all data points within each cluster. 2) For each pairwise comparison of cluster-averaged embeddings, we identified the 100 genes with the highest attribution scores. 3) We interpreted the attribution scores as indicators of a gene's influence on differences between clusters, analogous to differential gene expression analysis. We hypothesized that genes with high attribution scores would be associated with cellular components and structures that define kidney tissue architecture. To test this hypothesis, we performed Gene Set Enrichment Analysis (GSEA) Preranked (Subramanian et al., 2005) on these top 100 genes ranked by attribution score. We compared the most highly attributed genes to Gene Ontology Cellular Component (GO:CC) terms (Ashburner et al., 2000), as this ontology describes structural components rather than biological or molecular functions. We report enriched terms at a significance level of 2%. Table 3 illustrates the results for test sample 4, showing the normalized enrichment score (NES) of all significantly enriched GO:CC terms with at least 4 leading genes. This representation highlights the most relevant structural components associated with the highly attributed genes in this sample. The results for the remaining samples are provided in Appendix F.2. Inspection of the top enriched terms (see Appendix F.1) shows that they align with prominent anatomical features of their respective query regions, indicating that the model relies on structurally relevant genes to discriminate morphological patterns across different tissue regions.

Conclusion

We introduce, to the best of our knowledge, the first approach to synthesise histology images directly from high-resolution GEX data. Experiments on five kidney samples demonstrate that our model learns biologically meaningful gene expression representations, generates histologically accurate tissue morphologies across diverse renal structures, and produces images that achieve competitive FID scores and expert pathologists find indistinguishable from real tissue. Although our proof-of-concept work relies on only six sections of healthy mouse kidney—hence on a single tissue type and species—it nonetheless shows that histological images can be synthesized directly from GEX data, opening new avenues for systematic exploration of gene–morphology links. As larger and more diverse ST datasets emerge, this approach could evolve into a powerful tool for pinpointing molecular drivers of early disease and for in silico simulations of the structural consequences of gene perturbations.

Acknowledgements

The work of Frederieke Lohmann was conducted as part of an internship at F. Hoffmann-La Roche Ltd. Alberto Valdeolivas and Jelica Vasiljević are employees of F. Hoffmann-La Roche Ltd. This work was supported by the Roche Postdoctoral Fellowship to Jelica Vasiljević.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

10XGenomics. Adult Mouse Kidney (FFPE), 2021. URL https://www.10xgenomics.com/datasets/adult-mouse-kidney-ffpe-1-standard-1-3-0.

10XGenomics. Visium HD Spatial Gene Expression Library, Mouse Kidney (FFPE), 2024. URL https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-kidney.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. ISSN 1061-4036. doi: 10.1038/75556.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Carrillo-Perez, F., Pizurica, M., Ozawa, M. G., Vogel, H., West, R. B., Kong, C. S., Herrera, L. J., Shen, J., and Gevaert, O. Synthetic whole-slide image tile generation with gene expression profile-infused deep generative models. *Cell Reports Methods*, 3(8):100534, 2023. ISSN 2667-2375. doi: 10.1016/j.crmeth.2023.100534.

Carrillo-Perez, F., Pizurica, M., Zheng, Y., Nandi, T. N., Madduri, R., Shen, J., and Gevaert, O. Generation of synthetic whole-slide image tiles of tumours from RNA-sequencing data via cascaded diffusion models. *Nature Biomedical Engineering*, pp. 1–13, 2024. doi: 10.1038/s41551-024-01193-8.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., and Mahmood, F. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. ISSN 1078-8956. doi: 10.1038/s41591-024-02857-3.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*, 2020. doi: 10.48550/arxiv.2010.11929.

Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Camara, A., Kain, A. M., Saillard, C., and Schiratti, J.-B. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. *medRxiv*, pp. 2023.07.21.23292757, 2024. doi: 10.1101/2023.07.21.23292757.

Hahn, K., Amberg, B., Monné Rodriguez, J. M., Verslegers, M., Kang, B., Wils, H., Saravanan, C., Bangari, D. S., Long, S. Y., Youssef, S. A., et al. Points to consider from the estp pathology 2.0 working group: Overview on spatial omics technologies supporting drug discovery and development. *Toxicologic Pathology*, pp. 01926233241311258, 2025.

Heimberg, G., Kuo, T., DePianto, D. J., Salem, O., Heigl, T., Diamant, N., Scalia, G., Biancalani, T., Turley, S. J., Rock, J. R., Bravo, H. C., Kaminker, J., Heiden, J. A. V., and Regev, A. A cell atlas foundation model for scalable search of similar human cells. *Nature*, pp. 1–3, 2024. ISSN 0028-0836. doi: 10.1038/s41586-024-08411-y.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv*, 2017. doi: 10.48550/arxiv.1706.08500.

Ho, J. and Salismans, T. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022.

Kumaran, G. K. and Hanukoglu, I. Mapping the cytoskeletal architecture of renal tubules and surrounding peritubular capillaries in the kidney. *Cytoskeleton*, 81(4-5):227–237, 2024. doi: https://doi.org/10.1002/cm.21809.

Lee, Y., Liu, X., Hao, M., Liu, T., and Regev, A. PathOmCLIP: Connecting tumor histology with spatial gene expression via locally enhanced contrastive learning of Pathology and Single-cell foundation model. *bioRxiv*, pp. 2024.12.10.627865, 2024. doi: 10.1101/2024.12.10.627865.

Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., and Thomas, N. E. A method for normalizing histology slides for quantitative analysis. *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1107–1110, 2009. doi: 10.1109/isbi.2009.5193250.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2 2018.

Min, W., Shi, Z., Zhang, J., Wan, J., and Wang, C. Multimodal contrastive learning for spatial gene expression prediction using histology images. *Briefings in Bioinformatics*, 25(6):bbae551, 2024. ISSN 1467-5463. doi: 10.1093/bib/bbae551.

Müller-Franzes, G., Niehues, J. M., Khader, F., Arasteh, S. T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., Kather, J. N., and Truhn, D. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 7 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-39278-0.

Navidi, Z., Ma, J., Miglietta, E. A., Liu, L., Carpenter, A. E., Cimini, B. A., Haibe-Kains, B., and Wang, B. MorphoDiff: Cellular Morphology Painting with Diffusion Models. *bioRxiv*, pp. 2024.12.19.629451, 2024. doi: 10.1101/2024.12.19.629451.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. *arXiv*, 2021. doi: 10.48550/arxiv.2103.00020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Saharia, C., Chan, W., Saxena, S., Lit, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Sauer, A., Karras, T., Laine, S., Geiger, A., and Aila, T. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. *arXiv*, 2023. doi: 10.48550/arxiv.2301.09515.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506580102.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic Attribution for Deep Networks. *arXiv*, 2017. doi: 10.48550/arxiv.1703.01365.

Traag, V. A., Waltman, L., and van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 3 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-41695-z.

Wu, J. and Koelzer, V. H. SST-editing: in silico spatial transcriptomic editing at single-cell resolution. *Bioinformatics*, 40(3):btae077, 2024. ISSN 1367-4803. doi: 10.1093/bioinformatics/btae077.

Wu, J., Berg, I., and Koelzer, V. H. IST-editing: Infinite spatial transcriptomic editing in a generated gigapixel mouse pup. *bioRxiv*, pp. 2023.12.23.573175, 2023. doi: 10.1101/2023.12.23.573175.

Xie, R., Pang, K., Chung, S. W., Perciani, C. T., MacParland, S. A., Wang, B., and Bader, G. D. Spatially Resolved Gene Expression Prediction from H&E Histology Images via Bi-modal Contrastive Learning. *arXiv*, 2023. doi: 10.48550/arxiv.2306.01859.

A. Training Details

A.1. Contrastive Image-Gene Encoder

The contrastive image-gene encoder is pretrained on a subset of 10000 image-gene spot pairs drawn uniformly at random from the VisiumHD training sample. We train for a total of 4200 iterations at a learning rate of $5e-6$ on a NVIDIA A100-SXM4 80GB GPU. A batch size of $N = 128$ and a dropout rate of $p = 0.8$ is used. Validation of model convergence is performed on sample 1.

A.2. StyleGAN-G

StyleGAN-G is trained on all available data pairs from the VisiumHD sample at a learning rate of $2e-3$. Training is conducted in 2 progressive growing steps on 2 NVIDIA A100-SXM4 80GB GPUs. First, we train at 64×64 resolution for 16,000 iterations. For subsequent training at 128×128, we keep all layers up to 64×64 resolution fixed and train for 50,000 iterations. We use a total batch size of 128 and a per-GPU batch size of 8. The guidance strength was set to $\lambda = 0.2$. In preliminary experiments, training of StyleGAN-G did not converge when jointly optimizing a non-pretrained GEX encoder, instead of using the fixed GEX encoder from Phase 1, highlighting the importance of learning a principled GEX embedding.

A.3. Stable Diffusion

The Stable Diffusion Pipeline was fine-tuned for 55 epochs on a NVIDIA A100-SXM4 80GB GPU, keeping all components except the U-Net frozen. We use the HuggingFace implementation of SD and fine-tune at a batch size of 32 and a learning rate of $1e-5$. To enable classifier-free guidance, the GEX embedding prompt is set to the embedding of the vector of all zeroes with probability $p_{\text{uncond}} = 0.1$. We use the definition of classifier-free guidance reported in Saharia et al. (2022), where $w = 1$ corresponds to using no guidance.

B. Data Details

The Visium HD processed training sample and the standard Visium processed sample 1 were sourced from 10XGenomics (2024; 2021) and represent the coronal section of a kidney. The remaining standard Visium processed samples 2-5 originate from a proprietary dataset and represent the transverse section of a kidney. Samples 1-5 yield 3124, 1490, 1670, 1421, and 1427 data pairs respectively. In Visium HD data, the location of the center $2\mu m$ bin is used as the spatial location of the pseudo-spot. The stain normalisation is applied to each WSI individually before extracting image patches.

C. Real Image Data

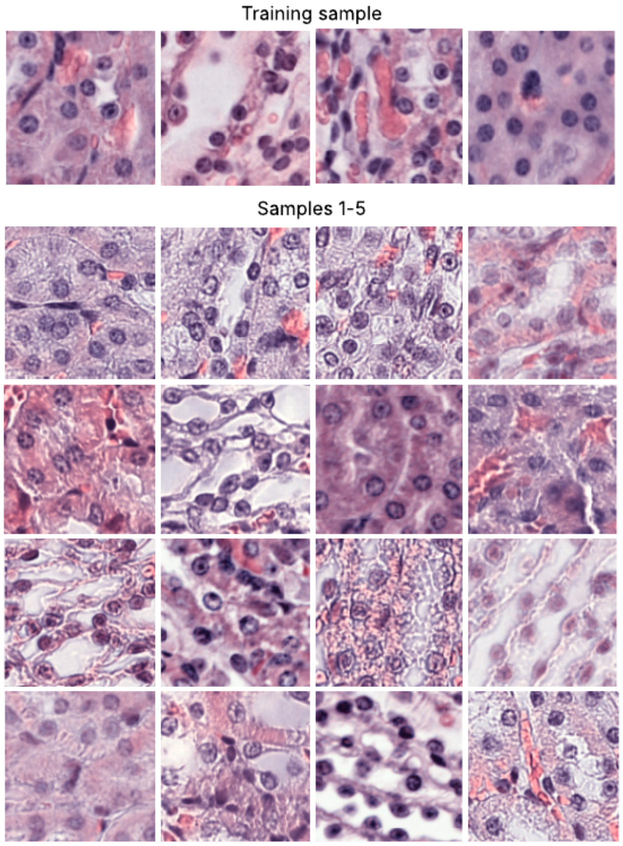


Figure C.1. Representative example patches from the training data (top row) and samples 1-5 (bottom)

D. Image Embedding Space

Figure D.1 displays a UMAP of the image embedding space generated using our CLIP model on both synthetic and real images (see also Table 1). It demonstrates that synthetic and real images are well mixed and thus further strengthens the claim that the generated images accurately reflect the true data distribution.

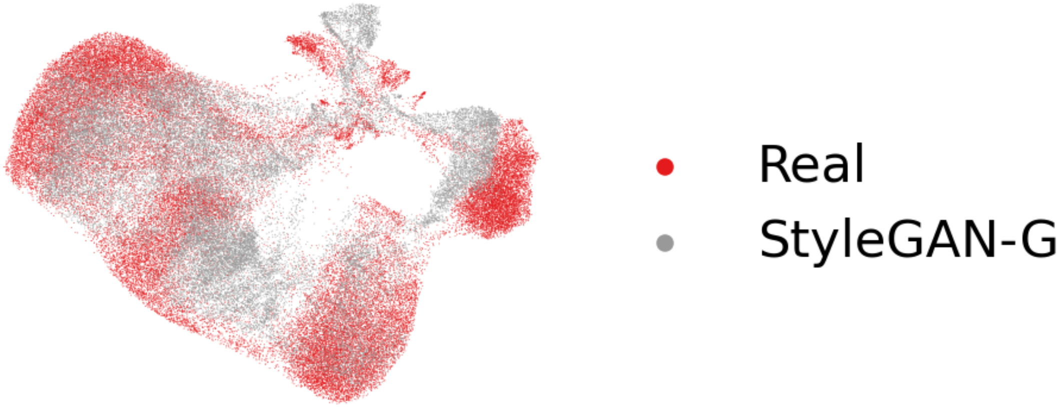
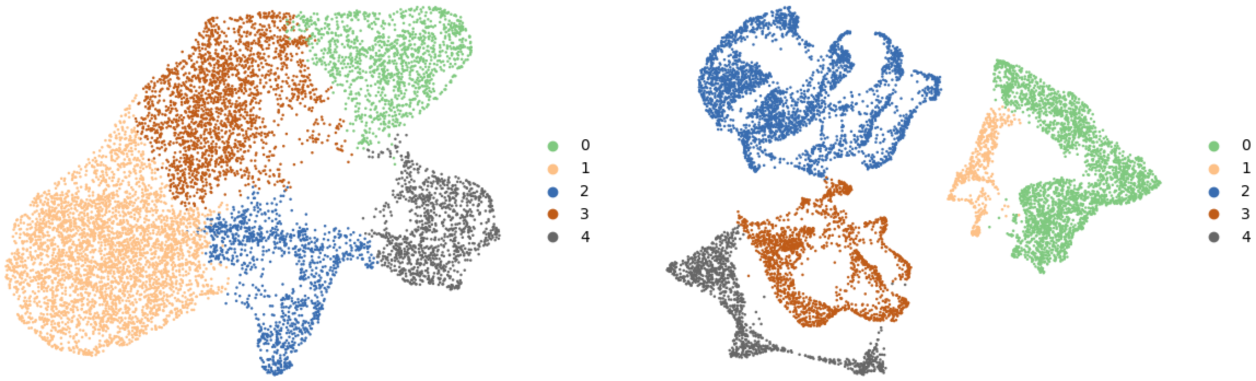


Figure D.1. UMAP of CLIP image embedding space

E. ARI Reference Clusters



(a) GEX Embedding

(b) GEX

Figure E.1. Clusters found by clustering jointly across samples in (a) gene expression embedding space (b) gene expression space. Used as reference to compute Adjusted Rand Index.

F. Gene Attributions

F.1. Interpretation of GO:CC Terms enriched in Sample 4

Table 3 shows an enrichment of the Lipid Droplet term in brown adipose tissue as compared to other renal regions, which is expected given the high concentration of fat-storing adipocytes in this region. The embeddings of Basolateral Plasma Membrane genes distinguish the cortex from both the OSOM and brown adipose tissue. This reflects the cortex's abundance of proximal and distal convoluted tubules, whose extensive basolateral membranes support active secretion and absorption of solutes. In contrast, the OSOM has fewer of these tubule segments, while the brown adipose tissue is composed mainly of adipocytes which do not directly participate in solute transport. Consequently, basolateral membrane–related genes are more characteristic of the cortex's morphology than those of the OSOM or brown adipose tissue.

Differences between the inner medulla/papilla and the OSOM emerge from terms linked to the extracellular matrix. The inner medulla and papilla contain specialized connective tissue that maintains structural integrity under the high osmotic pressure generated when urine is concentrated in the collecting ducts. In line with this function, renal interstitial cells in these regions produce collagen and other extracellular matrix components, which rationalizes the enrichment of the collagen-containing extracellular matrix term. Additionally, enrichment of genes associated with the endoplasmic reticulum (ER) lumen and intracellular organelle lumen further differentiates the inner medulla/papilla from the OSOM, reflecting higher levels of protein synthesis and post-translational modification. In particular, the collecting ducts—key structures within the inner medulla and papilla—rely on these processes to carry out their specialized functions.

The derived attribution scores offer an opportunity to identify novel gene sets that drive these morphological differences, providing a valuable direction for future research to pinpoint genes whose loss or alteration may contribute to disease or drug-associated toxicity.

F.2. Significant GO:CC Terms

Table F.1. Significantly enriched GO:CC terms in sample 2 with more than 4 leading genes

Term	NES ↑	Query Cluster	Reference Cluster
Intracellular Membrane-Bounded Organelle	1.852065	Inner Medulla & Papilla	Cortex
Basolateral Plasma Membrane	1.631743	Inner Medulla & Papilla	Brown Adipose Tissue
Basolateral Plasma Membrane	1.661647	Cortex	OSOM
Endoplasmatic Reticulum Lumen	1.951187	Inner Medulla & Papilla	OSOM
Intracellular Organelle Lumen	1.729963	Inner Medulla & Papilla	OSOM

Table F.2. Significantly enriched GO:CC terms in sample 3 with more than 4 leading genes

Term	NES ↑	Query Cluster	Reference Cluster
Vesicle	-1.853656	Brown Adipose Tissue	OSOM
Intracellular Organelle Lumen	1.78486	Brown Adipose Tissue	OSOM
Endoplasmatic Reticulum Lumen	1.703005	Inner Medulla & Papilla	OSOM
Basolateral Plasma Membrane	1.918574	Inner Medulla & Papilla	Cortex
Nucleus	1.852569	Inner Medulla & Papilla	Cortex
Intracellular Membrane-Bounded Organelle	1.700012	Inner Medulla & Papilla	Cortex
Basolateral Plasma Membrane	1.701396	Inner Medulla & Papilla	Brown Adipose Tissue
Nucleus	-1.687745	ISOM	Inner Medulla & Papilla

Table F.3. Significantly enriched GO:CC terms in sample 5 with more than 4 leading genes

Term	NES ↑	Query Cluster	Reference Cluster
Endoplasmatic Reticulum Lumen	1.937943	Inner Medulla & Papilla	OSOM
Intracellular Organelle Lumen	1.72107	Inner Medulla & Papilla	OSOM
Basolateral Plasma Membrane	1.938538	Inner Medulla & Papilla	Cortex
Nucleus	-1.677389	ISOM	Inner Medulla & Papilla