Deep Learning Enables Automated Segmentation and Quantification of Ultrastructure from Transmission Electron Microscopy Images

Anqi Zou[1,#], Winston Tan[2,#], Jiayi Ji[2,#], Florencia Rojas-Miguez[3], Laura Dodd[2], Emily Oei[2], Hui Chen[3], Joel M. Henderson[3], Xueping Fan[2,*], Weining Lu[2,*], Chao Zhang[1,*]

[1] Computational Biomedicine Section, Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston Medical Center, Boston, MA, USA

[2] Nephrology Section, Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston Medical Center, Boston, MA, USA

[3] Department of Pathology and Laboratory Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston Medical Center, Boston, MA, USA

[#] These authors contributed equally to this work
* Corresponding authors

E-mail addresses: chz2009@bu.edu (C. Zhang), wlu@bu.edu (W. Lu), xpfan@bu.edu (X. Fan)

Abstract

The widths of kidney glomerular basement membrane (GBM) and podocyte foot processes (FP) are essential ultrastructural markers for assessing kidney function and diagnosing glomerular disease. FP widening reflects podocyte injury, whereas GBM thickening is characteristic of conditions such as diabetic nephropathy and Alport syndrome. Current measurement practices depend on manual tracing and expert review, which are labor intensive, variable between observers, and difficult to reproduce.

We present an automated deep learning dbased framework TEAMKidney for accurate and scalable measurement of GBM and FP widths in transmission electron microscopy (TEM) images of glomeruli. The framework combines a proposed Glom2Mask model for panoramic segmentation, enabling robust identification of GBM and FP across species, including mouse, rat, and human kidney samples. Segmentation outputs are processed by computational pipelines to extract and quantify GBM and FP widths.

The proposed method showed high concordance with expert annotations, substantially reduced measurement time, and improved reproducibility across datasets. This framework supports both experimental research and clinical applications, offering a reproducible and efficient approach to ultrastructural kidney analysis. By reducing dependence on manual measurements, it provides a scalable digital pathology solution with direct relevance to nephrology research, diagnostic practice, and clinical trials.

Introduction

Chronic kidney disease (CKD) affects over 44 million people in the US and 1 billion worldwide[1]. CKD attributed mortality increased 41.5% from 1990-2017[2], now exceeding breast

and prostate cancer deaths combined[1]. Medicare spending for CKD reached $85.4 billion in 2020[3], surpassing the top three cancers combined ($77.9 billion)[4].

Central Importance of Glomerular Filtering Unit Ultrastructural Changes: Glomerular diseases account for more than 20% of CKD cases[5,6] and are the leading cause of kidney failure[7,8]. Although damage to these filtering units of the kidney and a special cell called podocyte can produce a range of symptoms, such as albuminuria which is often an early indicator of CKD[9], a urine test alone cannot diagnose the exact type of glomerular disease. The complexity and variety of conditions leading to glomerular disease mean that each can be characterized by specific damage to ultrastructural features of the glomerulus. For example, minimal change disease (MCD) and focal segmental glomerulosclerosis (FSGS) often result in the effacement of podocyte FP[10,11]. Alport syndrome and thin basement membrane disease are defined by thickening or thinning of the GBM[12]. Immune complex glomerulonephritis is characterized by the deposition of immune complexes within the glomerular capillary wall[13]. Accurate identification of these ultrastructural changes is essential for making a precise diagnosis, guiding treatment, and monitoring disease progression. This level of detailed assessment can have a direct impact on the effectiveness of clinical management and the patient's life.

The glomerular filtration barrier (GFB) is a highly specialized structure within the kidney nephrons that plays a crucial role in maintaining fluid and electrolyte homeostasis by selectively filtering plasma components. It consists of three layers: the glomerular endothelium, the GBM, and podocytes with their interdigitating FP. These components work in concert to regulate the passage of water and solutes while preventing the loss of essential proteins and cellular components into the urine. Disruptions to the structure or function of the GFB contribute to the pathogenesis of CKD and various glomerular disorders[14,15]. The GBM forms the central layer of the GFB, acting as both a size- and charge-selective barrier that regulates plasma filtration[15]. Alterations in GBM thickness or structure are strongly associated with glomerular diseases[16-19]. Podocytes are highly specialized epithelial cells that form the outermost layer of the GFB. They extend FP that interdigitate to form filtration slits, which are covered by a slit diaphragm, a crucial structure that maintains selective permeability and prevents protein loss[20]. Podocyte structural integrity is vital for glomerular function, and their injury correlates with the progression of kidney disease[21,22]. Therefore, accurate GBM and FP width measurement is essential for diagnosing and monitoring kidney disease progression.

Although ultrastructural features can provide crucial evidence for diagnosing glomerular diseases[23], routine light microscopy and immunofluorescence (IF) staining are unable to detect these key changes. In both research and clinical pathology, TEM is commonly used to observe kidney tissue ultrastructure and identify morphological changes at the micrometer scale (Fig. 1A). Because of its unique capabilities, some major medical centers now use TEM routinely for diagnosis, analyzing nearly half of native or transplant renal biopsies[24]. Beyond clinical diagnostics, these measurements are also used as image biomarkers in research, including clinical trials and animal model studies, to quantify ultrastructural changes resulting from podocyte injury, genetic mutations, or therapeutic interventions[25].

The most impactful methods made in the past decade are highlighted in Table 2, with what we considered to be the most important aspects about a model that detects and measures the GBM and FP structures. Laudon et al[26] and Yamashita et al[27] were able to measure both GBM and , with their method only being applied to animal models of disease and limited dataset. Both

approaches used semantic segmentation for width calculation, but this method cannot distinguish individual FPs when they are closely adjacent or overlapping. Such structural overlaps are common in pathological conditions and results in inaccurate measurements. Lin et al[28] proposed a self-supervised representation learning approach for GFB segmentation in TEM images, but their method only provides sematic segmentation without quantitative measurements of structural widths. Smerkous et al[29] developed an automated method for FP width measurement and successfully trained and tested it on human glomerular TEM data. Unnersjö-Jess et al[30] developed AMAP (Automatic Morphological Analysis of Podocytes) for quantifying foot process morphology using confocal and super resolution fluorescence microscopy images. Wang et al[31], Wu et al[32] and Yang et al[33] proposed different algorithms for measuring GBM thickness. Moreover, these existing methods suffer from several fundamental limitations, including restricted training datasets that limit generalizability across diverse disease models and imaging conditions, suboptimal model accuracy in detecting fine ultrastructural details, and the persistent challenge of distinguishing closely adjacent or overlapping FPs in pathological states. These constraints collectively hinder their utility for comprehensive assessment of glomerular ultrastructural changes that require simultaneous and accurate evaluation of both GBM and podocyte structures.

Deep learning models have played a pivotal role in the approaches described above, mirroring the diagnostic workflow of human experts by first identifying ultrastructural components in glomerular TEM images and then performing quantitative measurements. The majority of these methods have been built upon U-Net and its variants as the core segmentation architecture. The Segment Anything Model (SAM)[34] became the most successful deep learning model for general image segmentation. Its fine-tuned application with a large number of medical images named MedSAM was also published recently[35]. Although these two models outperform most models in many biomedical image segmentation scenarios, the performance in TEM image analysis is limited even with manual prompts (Yellow boxes in Fig 1E, 1F). The need for precise prompts highlights the difficulty in applying SAM to TEM images. Significant variance across different diseases, the lack of high quality labeled training data, high resolution images with multi scale and fine grained features, and boundary ambiguity make TEM image analysis extremely challenging for general AI models. Thus, a specially designed AI model is needed to efficiently and accurately solve this challenge.

In this study, we propose a novel deep learning based framework for the automated quantification of GBM and FP widths in glomerular TEM images (Fig 1. G-K). Our approach integrates a self-training strategy to train a high resolution network (HRNet) for initial sematic segmentation, followed by an improved Mask2Former model for panoramic segmentation of TEM images. This framework enables precise delineation of GBM and FP structures across multiple species, including mouse, rat, and human kidney samples, demonstrating its applicability in both preclinical research and clinical diagnostics. Post-processing methods are then employed to extract width measurements, reducing reliance on manual intervention while maintaining high accuracy and efficiency.

Compared with traditional manual or stereological methods, our automated approach provides three main advantages: (i) increased efficiency, enabling large scale studies; (ii) enhanced reproducibility by eliminating observer variability; and (iii) broad applicability, validated across species in both preclinical and clinical contexts.

By integrating deep learning with computational pathology, our method presents a scalable and efficient solution for automated GBM and FP measurement. This approach has the potential to enhance diagnostic accuracy, facilitate large scale kidney disease research, and support clinical decision making in nephrology.

Results

The Overview of the TEAMKidney Framework

We propose a deep learning based automated framework for kidney ultrastructure analysis, designed to precisely quantify GBM and FP widths from TEM images (Fig. 1G). The framework comprises three key components: (1) integrating HRNet with self-training for initial semantic segmentation (Fig. 1H), (2) developing a modified Mask2Former model for panoramic segmentation of glomerular ultrastructure (Fig. 1H), and (3) performing post segmentation analysis to enable robust quantification of GBM and FP across mouse, rat, and human samples (Fig. 1I-K). This framework was implemented to automatically segment, label, and measure GBM and FP widths, thereby eliminating the need for manual tracing and reducing observer variability. This design allows the framework to address diverse measurement tasks, including interspecies comparison, disease associated ultrastructural change, and quantitative evaluation in experimental and clinical datasets.

To train and validate our framework across diverse pathological conditions and species, our study conducted analysis of both animal and human models of normal and diseased states. For mouse models, we used *Integrin linked kinase* podocyte specific conditional knockout (ILK cKO) and *type IV collagen α3* knockout (Col4a3 KO) mice[36-38]. For rats, the passive Heymann nephritis model (PHN)[39] was used, which was originally developed to characterize Membranous nephropathy (MN). For human disease states study, we focused on DKD[40,41] and Fabry[42,43] disease.

Two Stage Segmentation Model for Improved Fine Ultrastructure Recognition

In the semantic segmentation stage (Fig. 2A), we evaluated three backbone networks for comparison: DeepLabV3[44-46], PSPNet[46], and HRNet. All three models are effective in capturing global context and fusing multi scale features. Overall, the models achieved comparable performance (Fig. 2B). Among them, HRNet attained the highest Dice scores across most categories and demonstrated markedly greater training efficiency. Although the average Dice scores of the three methods were similar, HRNet exhibited the most robust and consistent performance. Therefore, we selected HRNet as the backbone for our panoptic segmentation model in the second stage (Fig. 2C). In contrast, DeepLabV3 often failed on low-quality images (Fig. 2C), resulting in unreliable segmentation, while PSPNet tended to misclassify unrelated ultrastructures (Fig. 2C), particularly mistaking epithelial cells on glomerular capillaries for foot processes.

Although our semantic segmentation model accurately recognized major ultrastructures, direct post segmentation analysis to identify individual instances such as FP and slit diaphragms remained highly challenging and often inaccurate, similar to most existing models. To overcome this limitation, we introduced a panoptic segmentation stage (Fig. 2A) based on our proposed Glom2Mask model. Our proposed Glom2Mask model includes the HRNet, trained in the first step, as the backbone for feature extraction and integrated into the modified Mask2Former

architecture. Meanwhile, using the annotations generated during the semantic segmentation stage, together with manually refined corrections, we constructed an expanded and higher quality labeled training dataset.

The panoptic segmentation performance of the proposed Glom2Mask and the baseline Mask2Former models were compared using standard panoptic segmentation metrics: Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (RQ), evaluated for both things (FPs) and stuff (GBM) classes. The results show that Glom2Mask outperformed the baseline Mask2Former (Fig. 2D), achieving higher PQ, SQ, and RQ scores for both FP and GBM while requiring fewer training iterations (66,000 vs. 360,000). These findings demonstrate superior segmentation accuracy and training efficiency of the proposed model. Fig. 2E shows representative segmentation results from Glom2Mask and Mask2Former. The visualizations demonstrate that Glom2Mask accurately distinguishes continuous GBM structures (blue) from discrete FP instances (various shades of red), further supporting its robustness and generalizability across species.

Concordance of TEAMKidney Predictions with Ground Truth Across Species

To evaluate measurement accuracy, we compared GBM and FP width values obtained from the TEAMKidney method with ground truth measurements across mouse, rat, and human samples. All ground truth images came from the model validation set. The GBM and FP annotation was first manually labeled by three human experts to generate masks of the ultrastructures. By integrating manual annotation with expert and standardized measurement, this process produced ground truth data that closely approximated true biological values with high consistency and efficiency. No statistically significant differences were observed between the ground truth and TEAMKidney predicted measurements for either GBM or FP width across all three species (Fig. 2F), demonstrating strong concordance and reliability of the proposed framework. Overall, these results indicate that the proposed method provides highly reliable GBM and FP width measurements across different datasets.

Cross species Validation of GBM and FP Measurements in Experimental Kidney Diseases

To test the robustness of our framework in detecting pathophysiological changes in the glomerular filtration barrier, we applied our automated method to measure two well established mouse models and a rat model of glomerular diseases (Fig. 3A and B). In the mouse models, we used type IV collagen subunit alpha 3 knockout (Col4a3 KO) and integrin linked kinase podocyte specific knockout (ILK cKO) mice to represent two distinct mechanisms of kidney glomerular injury. Col4a3 KO mice exhibit severe GBM disorganization due to the lack of type IV collagen, which is a crucial component of the GBM, whereas ILK cKO mice develop podocyte detachment and cytoskeletal abnormalities with absence of ILK that is required for podocyte adhesion to the GBM (Fig. 3A and C). Consistent with these pathological features, our automated analysis revealed significantly increased GBM widths in both models compared with wild type controls. Similarly, FP widths were markedly widened. These quantitative differences align with the expected ultrastructural alterations associated with podocyte injury (Fig. 3C and D).

We next examined the rat PHN model of immune mediated podocyte injury. In this model, rats were injected with anti-Fx1a antibody to induce the formation of immune complex that deposits beneath podocytes and trigger complement activation, leading to progressive FP effacement and GBM thickening (Fig. 3B and E). Automated measurements faithfully captured

these ultrastructural changes: compared with wild type rats, PHN animals showed a striking increase in GBM width (Fig. 3F). These differences are consistent with classical TEM based characterizations of PHN, confirming that our approach is sensitive to detecting pathological remodeling of the glomerular barrier.

Automated Quantification of GBM and FP Widths in Human Kidney Diseases

To investigate the translational utility of our framework in human kidney disease states, we analyzed TEM images from patients with DKD and Fabry disease (Fig. 4A). DKD is characterized by hyperglycemia induced hyperfiltration, leading to podocyte stress and progressive glomerular injury, whereas Fabry disease results from GLA mutations that cause accumulation of globotriaosylceramide (Gb3) in podocytes and subsequent structural damage (Fig. 4A). Representative TEM images predicted by our TEAMKidney model demonstrate marked distinct pathological features in both disease cohorts compared with normal controls (Fig. 4B).

Quantitative measurements revealed disease specific ultrastructural alterations (Fig. 4C). In DKD patients, GBM thickness was significantly increased. Notably, mean FP width in DKD showed only a slight increase relative to normal samples. The relatively preserved FP width in our DKD cohort warrants further investigation, as this may reflect the disease stage of the included patients or capture an intermediate phase where GBM pathology precedes overt podocyte architectural disruption. Longitudinal studies correlating FP measurements with clinical disease staging would help clarify the temporal evolution of these ultrastructural changes. In contrast, Fabry disease patients exhibited significant increases in both GBM thickness and FP width, reflecting the more severe podocyte injury characteristic of lysosomal storage disorders.

Furthermore, sex stratified analysis of Fabry patients revealed that both male and female patients exhibited substantial GBM and FP widening, with more pronounced changes observed in Fabry males. These findings are consistent with the known pathology of Fabry disease, where hemizygous males typically display more severe phenotypes due to the X-linked inheritance pattern.

These results highlight the ability of our automated framework to detect clinically relevant and disease specific ultrastructural alterations in human glomerular diseases. By accurately capturing the differential patterns of GBM thickening and FP effacement in DKD and Fabry disease, the method demonstrates strong potential for application in both diagnostic pathology and clinical research, where reproducible and scalable measurements are critical for disease characterization, staging, and patient stratification.

Comparison of FP quantification between TEAMKidney and the Benchmark FPW-DL Models

We next benchmarked our proposed framework against existing automated approaches for quantifying FP widths across multiple species including human, mouse and rat kidney samples (Fig. 5). Our method TEAMKidney was compared to FPW-DL, a model published and publicly released by Smerkous et al. in 2023[29]. It is to note again, that FPW-DL is only able to measure FP widths, and trained on human dataset.

In human samples, TEAMKidney consistently outperformed FPW-DL across both normal and disease cohorts, with data points clustering more tightly along the identity line in correlation

plots (Fig. 5A). For DKD human samples, the mean error of FP width was reduced from 24.80 pixels (FPW-DL) to 4.46 pixels using our framework. This difference can be attributed to FPW-DL's approach of calculating FP widths through split diaphragm detection. In disease conditions or under varied imaging conditions, FP effacement and overlapping can obscure split diaphragms, leading to either overestimation of FP widths when diaphragms are missed, or selective detection of only relatively normal appearing foot processes while bypassing pathologically altered regions.

The performance gap between the two methods became more pronounced in mouse ILK-KO and rat PHN models, where FPW-DL, which was trained exclusively on human kidney data, showed substantially larger deviations from ground truth (Fig. 5A). This limitation stems from two primary factors: first, morphological differences between species, particularly the thinner GBM in rodents compared to humans, pose additional challenges for accurate structure identification; second, differences in imaging resolution and scale between our rodent datasets and FPW-DL's training data further compromised its generalizability. In contrast, TEAMKidney maintained robust performance across species, demonstrating superior adaptability to diverse morphological features and imaging conditions.

In addition to error reduction in FP width measurements, our method improved the accuracy of FP counts, which are highly sensitive to segmentation quality. The scatter plots revealed that while FPW-DL tended to underestimate FP number with greater deviation from ground truth, our framework achieved close concordance with ground truth annotations across human disease states and animal models (Fig. 5B). Representative TEM images further illustrate that our approach provided more continuous and anatomically faithful segmentation of podocyte foot processes compared with alternative methods (Fig. 5C).

Concordance of automated and manual measurements in human kidney disease

To further assess the reliability of our framework, we compared automated measurements with manual measurements across normal, DKD, and Fabry disease (male and female) cohorts (Fig. 6). For GBM widths, automated quantification showed no significant differences from manual measurements in all groups (Fig. 6A-D, top panels), confirming that the framework preserves accuracy for this parameter.

For FP widths, although the mean values of our predictions are lower than those of manual measurements, the two distributions exhibit a similar shape and range, indicating consistent measurement trends. In the manual approach, FP width is obtained by dividing the GBM length by the number of FPs covering it, so the slit diaphragm is included in the FP. In contrast, the automated measurement strategy segments individual FPs and the slit diaphragm separately; therefore, the slit diaphragm is excluded from the FP measurement, resulting in higher accuracy compared with the manual method. Similar concordance was observed in DKD, Fabry male, and Fabry female samples (Fig. 6, middle panels).

In terms of FP counts, no significant differences were detected between automated and manual annotations across all groups (Fig. 6A-D, bottom panels), indicating that the framework maintains biological fidelity in enumerating podocyte processes.

Together, these results demonstrate strong concordance between automated and manual measurements, establishing the framework as a reliable substitute for labor intensive tracing. By accurately replicating expert level quantification across both healthy and diseased human

samples, the method enhances reproducibility and offers a scalable solution for ultrastructural kidney assessment.

Comparison of clinical protocol and automated framework for GBM and FP quantification

To contextualize the advantages of our automated approach, we compared it with the conventional diagnostic manual workflow commonly used for ultrastructural measurements of GBM and FP widths (Fig. 7). Pathologists rely on ImageJ (or FIJI) based protocols, which require calibration with scale bars, grid overlays, and labor intensive tracing of GBM thickness at multiple grid intersections, followed by foot process identification and marking of lateral edges (Fig. 7A). This procedure demands at least 200 measurements from three independent glomeruli per sample to achieve statistical robustness, making it time consuming and highly dependent on operator expertise.

Representative TEM images further illustrate the differences in measurement strategies (Fig. 7B-E). Manual quantification requires repetitive grid based annotation and point by point width extraction (Fig. 7C and D), whereas our automated framework directly performs panoptic segmentation and downstream computational measurement of GBM and FP widths (Fig. 7E). This not only eliminates observer variability but also enables high throughput analysis of large datasets with consistent reproducibility.

Collectively, these comparisons underscore the limitations of traditional manual workflows, including their subjectivity, inefficiency, and limited scalability, and demonstrate that the proposed automated framework effectively overcomes these challenges, providing a practical and objective solution for ultrastructural kidney analysis.

Discussion

Nephrology research and clinical pathology have long awaited tools to alleviate the labor intensive and often inaccurate assessment of glomerular diseases. Accordingly, this study was motivated by the urgent need for a reliable, efficient, and reproducible method to quantify ultrastructural alterations in the kidney filtration barrier. In both research and clinical settings, accurate measurement of the GBM and podocyte FP is essential for diagnosing and monitoring proteinuric kidney diseases. However, existing manual workflows are labor intensive, subjective, and poorly scalable, while current automated tools remain limited in accuracy, generalization, and interpretability. Therefore, we aimed to develop a robust deep learning framework that integrates semantic and panoptic segmentation to achieve precise and automated quantification of GBM and FP widths.

Previous models have primarily focused on segmenting a single structure, often lacking the ability to measure both GBM and FP quantitatively or to generalize across species. As summarized in Table 2, most U-Net based or semi-automated approaches fail to preserve fine spatial details and multi scale contextual information. To address these limitations, we designed a fully automated two stage pipeline. In the first stage, our proposed Glom2Mask panoptic segmentation network identifies individual instances of the GBM and FP with high precision. In the second stage, quantitative morphometric analysis is performed through an image processing module that automatically measures the widths of both GBM and FP based on the segmentation outputs. This two stage workflow requires minimal manual intervention and very few input parameters, ensuring full automation, high reproducibility, and robustness across species and

imaging conditions. The integration of instance level segmentation and quantitative measurement enables standardized, objective, and scalable ultrastructural assessment, representing a substantial advancement over prior methods that relied on manual or semi-automated measurement procedures.

The development of this model required iterative refinement. We initially adopted a semantic segmentation strategy, a concept that has also been explored in our previous studies[47], to delineate the GBM and FP. Although current semantic segmentation models such as HRNet can achieve relatively accurate boundary detection for both structures, they cannot distinguish individual FP instances. Due to variations in TEM imaging angles and pathological changes, many FPs, particularly in disease models—tend to cluster closely together, making it difficult to separate adjacent FPs after semantic segmentation. This limitation often results in overestimation of FP width in subsequent quantitative analysis. To address this issue, we transitioned to a panoptic segmentation framework. This design enables precise identification of individual FP instances and substantially improves the accuracy of FP width measurement.

Developing accurate ground truth labels for GBM and FP segmentation posed substantial challenges. Manual annotation of TEM images is extremely time consuming and requires expert knowledge due to the subtle and often ambiguous boundaries between ultrastructural components. Differences in image contrast, resolution, and tissue morphology further complicate the labeling process. To overcome these challenges, we established a standardized annotation protocol in collaboration with experienced renal pathologists, ensuring consistency and accuracy across datasets. In addition, we adopted semi-automated prelabeling strategies to accelerate the annotation workflow and implemented a multi stage quality control process, where each label was independently reviewed and refined. These steps significantly reduced inter operator variability and improved the overall reliability of the training data. We further employed a two stage self-training strategy, in which a high resolution semantic segmentation model (HRNet) was first used to generate pseudo labels of GBM and FP regions. These pseudo labels were subsequently refined by expert review and then used to train the panoptic segmentation model (Glom2Mask). This design offers two advantages over directly applying self-training on the panoptic model: (1) semantic segmentation provides spatially consistent region level guidance, reducing noisy or fragmented instance predictions during pseudo labeling; and (2) expert correction of semantic maps is more efficient than instance level annotation, substantially improving label quality and consistency. Moreover, the trained HRNet model can also serve as the backbone for the panoptic segmentation network, allowing feature representations learned during semantic segmentation to be directly transferred and reused. This not only accelerates model convergence but also enhances feature consistency between semantic and panoptic tasks, leading to improved boundary precision and instance recognition. Overall, this hierarchical training strategy achieves higher stability and accuracy, particularly for densely packed ultrastructural components such as podocyte foot processes.

TEM images inherently exhibit substantial variation in magnification and spatial resolution across different species, disease models, and experimental setups. Such differences often lead to inconsistencies in the visual scale and morphological representation of ultrastructural components, posing a major challenge for deep learning models trained on limited magnification ranges. To address this issue, we employed a series of strategies at multiple stages to enhance the model's robustness to magnification variability. First, we collected TEM samples acquired at different magnifications across species and disease models to ensure comprehensive

coverage of structural and contextual diversity. Second, we organized and standardized these multi magnification samples by applying systematic cropping and rescaling operations to generate datasets representing diverse spatial scales. This step provided the model with balanced exposure to both high and low resolution features of the glomerular ultrastructure. Third, we incorporated random scaling, cropping, and other image augmentation techniques during model training to simulate continuous magnification changes and prevent overfitting to specific resolutions. Collectively, these strategies substantially improved the model's ability to generalize across heterogeneous imaging conditions, resulting in consistent segmentation quality and reliable quantitative measurements in a wide range of magnification differences.

Our proposed model demonstrates three major advantages over existing approaches. First, it provides high accuracy and reproducibility in segmenting and measuring GBM and FP structures. Second, it generalizes effectively across species and diseases, accommodating significant morphological and magnification differences. Third, it integrates quantitative measurement and visual interpretation within one unified pipeline, facilitating standardized assessment for both experimental research and diagnostic pathology. Compared with prior frameworks that can only segment or qualitatively analyze these structures, our model is the first to achieve both segmentation and quantitative measurement with expert level precision.

It is again important to note that a model that can accurately map and measure normal and diseased GFB structure could be highly beneficial to renal pathologists. A model that can analyze highly prevalent disease states such as DKD, as well as rare diseases like Fabry could serve useful to clinicians that need a tool to accurately handle any disease state provided. This leads to our primary future direction — applying this model to TEM images of other disease states. Fine tuning the model to recognize the nuances of different diseases, such as the GBMs of thin basement membrane disease or the sclerotic and segmental progression of focal segmental glomerulosclerosis, would take further improvement on both the manual and model based aspects of this project. Although the ability to handle all potential disease states requiring GFB TEM imaging remains a challenge, it holds great promise given the progress this model has achieved.

Future work can further refine this system. Incorporating FP orientation (angle) analysis, distinguishing immune deposits embedded within the GBM, and excluding parietal GBM regions would further improve biological accuracy. Expanding training data with higher quality clinical TEM images and a wider magnification range will also enhance model robustness. Ultimately, integrating multimodal imaging data such as histology and immunofluorescence, or even involving clinical reports and data, could establish a comprehensive diagnostic platform for glomerular diseases.

In conclusion, our model represents a significant advancement in ultrastructural kidney analysis. By enabling precise, automated, and scalable quantification of GBM and FP widths, it addresses longstanding technical barriers in nephrology research and pathology. This framework holds strong potential to standardize ultrastructural evaluation across laboratories and to accelerate translational kidney research.

Reference

1 Assady, S., Benzing, T., Kretzler, M. & Skorecki, K. L. Glomerular podocytes in kidney health and disease. *Lancet* **393**, 856-858 (2019). https://doi.org:10.1016/s0140-6736(18)33000-9

2 Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **395**, 709-733 (2020). https://doi.org:10.1016/s0140-6736(20)30045-3

3 Johansen, K. L. *et al*. US Renal Data System 2022 Annual Data Report: Epidemiology of Kidney Disease in the United States. *Am J Kidney Dis* **81**, A8-a11 (2023). https://doi.org:10.1053/j.ajkd.2022.12.001

4 Mariotto, A. B., Enewold, L., Zhao, J., Zeruto, C. A. & Yabroff, K. R. Medical Care Costs Associated with Cancer Survivorship in the United States. *Cancer Epidemiol Biomarkers Prev* **29**, 1304-1312 (2020). https://doi.org:10.1158/1055-9965.Epi-19-1534

5 Ramachandran, R. *et al.* Challenges in Diagnosis and Management of Glomerular Disease in Resource-Limited Settings. *Kidney Int Rep* **7**, 2141-2149 (2022). https://doi.org:10.1016/j.ekir.2022.07.002

6 George, C., Mogueo, A., Okpechi, I., Echouffo-Tcheugui, J. B. & Kengne, A. P. Chronic kidney disease in low-income to middle-income countries: the case for increased screening. *BMJ Glob Health* **2**, e000256 (2017). https://doi.org:10.1136/bmjgh-2016-000256

7 Rajapurkar, M. M. *et al.* What do we know about chronic kidney disease in India: first report of the Indian CKD registry. *BMC Nephrol* **13**, 10 (2012). https://doi.org:10.1186/1471-2369-13-10

8 Sharma, M., Doley, P. & Das, H. J. Etiological profile of chronic kidney disease: A single-center retrospective hospital-based study. *Saudi J Kidney Dis Transpl* **29**, 409-413 (2018). https://doi.org:10.4103/1319-2442.229297

9 Tian, X. & Ishibe, S. Targeting the podocyte cytoskeleton: from pathogenesis to therapy in proteinuric kidney disease. *Nephrol Dial Transplant* **31**, 1577-1583 (2016). https://doi.org:10.1093/ndt/gfw021

10 Fogo, A. B. Causes and pathogenesis of focal segmental glomerulosclerosis. *Nat Rev Nephrol* **11**, 76-87 (2015). https://doi.org:10.1038/nrneph.2014.216

11 Rosenberg, A. Z. & Kopp, J. B. Focal Segmental Glomerulosclerosis. *Clin J Am Soc Nephrol* **12**, 502-517 (2017). https://doi.org:10.2215/cjn.05960616

12 Thorner, P. S. Alport syndrome and thin basement membrane nephropathy. *Nephron Clin Pract* **106**, c82-88 (2007). https://doi.org:10.1159/000101802

13 Haas, M. & Eustace, J. A. Immune complex deposits in ANCA-associated crescentic glomerulonephritis: a study of 126 cases. *Kidney Int* **65**, 2145-2152 (2004). https://doi.org:10.1111/j.1523-1755.2004.00632.x

14 Benzing, T. & Salant, D. Insights into Glomerular Filtration and Albuminuria. *N Engl J Med* **384**, 1437-1446 (2021). https://doi.org:10.1056/NEJMra1808786

15 Miner, J. H. The glomerular basement membrane. *Exp Cell Res* **318**, 973-978 (2012). https://doi.org:10.1016/j.yexcr.2012.02.031

16 Najafian, B., Alpers, C. E. & Fogo, A. B. Pathology of human diabetic nephropathy. *Contrib Nephrol* **170**, 36-47 (2011). https://doi.org:10.1159/000324942

17 Menon, R. *et al.* Single-cell analysis of progenitor cell dynamics and lineage specification in the human fetal kidney. *Development* **145** (2018). https://doi.org:10.1242/dev.164038
18 Hudson, B. G., Tryggvason, K., Sundaramoorthy, M. & Neilson, E. G. Alport's syndrome, Goodpasture's syndrome, and type IV collagen. *N Engl J Med* **348**, 2543-2556 (2003). https://doi.org:10.1056/NEJMra022296
19 Kashtan, C. Alport syndrome: facts and opinions. *F1000Res* **6**, 50 (2017). https://doi.org:10.12688/f1000research.9636.1
20 Pavenstädt, H., Kriz, W. & Kretzler, M. Cell biology of the glomerular podocyte. *Physiol Rev* **83**, 253-307 (2003). https://doi.org:10.1152/physrev.00020.2002
21 Reiser, J. & Altintas, M. M. Podocytes. *F1000Res* **5** (2016). https://doi.org:10.12688/f1000research.7255.1
22 Yu, D. *et al.* Urinary podocyte loss is a more specific marker of ongoing glomerular damage than proteinuria. *J Am Soc Nephrol* **16**, 1733-1741 (2005). https://doi.org:10.1681/asn.2005020159
23 Yamashita, M., Lin, M. Y., Hou, J., Ren, K. Y. M. & Haas, M. The Continuing Need for Electron Microscopy in Examination of Medical Renal Biopsies: Examples in Practice. *Glomerular Dis* **1**, 145-159 (2021). https://doi.org:10.1159/000516831
24 Haas, M. A reevaluation of routine electron microscopy in the examination of native renal biopsies. *J Am Soc Nephrol* **8**, 70-76 (1997). https://doi.org:10.1681/asn.V8170
25 Beck, L. H., Jr. *et al.* PODO: Trial Design: Phase 2 Study of PF-06730512 in Focal Segmental Glomerulosclerosis. *Kidney Int Rep* **6**, 1629-1633 (2021). https://doi.org:10.1016/j.ekir.2021.03.892
26 Laudon, A. *et al.* Digital pathology assessment of kidney glomerular filtration barrier ultrastructure in an animal model of podocytopathy. *Biology Methods and Protocols* **10**, bpaf024 (2025). https://doi.org:10.1093/biomethods/bpaf024
27 Yamashita, M. *et al.* AI-driven glomerular morphology quantification: a novel pipeline for assessing basement membrane thickness and podocyte foot process effacement in kidney diseases. *Computer Methods and Programs in Biomedicine* **268**, 108842 (2025). https://doi.org:https://doi.org/10.1016/j.cmpb.2025.108842
28 Lin, G. *et al.* GCLR: A self-supervised representation learning pretext task for glomerular filtration barrier segmentation in TEM images. *Artificial Intelligence in Medicine* **146**, 102720 (2023). https://doi.org:https://doi.org/10.1016/j.artmed.2023.102720
29 Smerkous, D. *et al.* Development of an automated estimation of foot process width using deep learning in kidney biopsies from patients with Fabry, minimal change, and diabetic kidney diseases. *Kidney Int* **105**, 165-176 (2024). https://doi.org:10.1016/j.kint.2023.09.011
30 Unnersjö-Jess, D. *et al.* Deep learning-based segmentation and quantification of podocyte foot process morphology suggests differential patterns of foot process effacement across kidney pathologies. *Kidney Int* **103**, 1120-1130 (2023). https://doi.org:10.1016/j.kint.2023.03.013
31 Wang, Y. *et al.* Segmentation and thickness calculation of glomerular basement membrane using RADS-Net in glomerular microscopic images. *Biomedical Signal Processing and Control* **88**, 105557 (2024). https://doi.org:https://doi.org/10.1016/j.bspc.2023.105557

32 Wu, H. S., Dikman, S. & Gil, J. A semi-automatic algorithm for measurement of basement membrane thickness in kidneys in electron microscopy images. *Comput Methods Programs Biomed* **97**, 223-231 (2010). https://doi.org:10.1016/j.cmpb.2009.07.002

33 Yang, J., Hu, X., Pan, H., Chen, P. & Xia, S. Multi-scale attention network for segmentation of electron dense deposits in glomerular microscopic images. *Microsc Res Tech* **85**, 3256-3264 (2022). https://doi.org:10.1002/jemt.24182

34 Kirillov, A. *et al.* in *Proceedings of the IEEE/CVF international conference on computer vision.* 4015-4026.

35 Ma, J. *et al.* Segment anything in medical images. *Nat Commun* **15**, 654 (2024). https://doi.org:10.1038/s41467-024-44824-z

36 Huang, M. *et al.* Integrin-Linked Kinase Deficiency in Collecting Duct Principal Cell Promotes Necroptosis of Principal Cell and Contributes to Kidney Inflammation and Fibrosis. *J Am Soc Nephrol* **30**, 2073-2090 (2019). https://doi.org:10.1681/asn.2018111162

37 El-Aouni, C. *et al.* Podocyte-specific deletion of integrin-linked kinase results in severe glomerular basement membrane alterations and progressive glomerulosclerosis. *J Am Soc Nephrol* **17**, 1334-1344 (2006). https://doi.org:10.1681/asn.2005090921

38 Cosgrove, D. *et al.* Collagen COL4A3 knockout: a mouse model for autosomal Alport syndrome. *Genes Dev* **10**, 2981-2992 (1996). https://doi.org:10.1101/gad.10.23.2981

39 Jefferson, J. A., Pippin, J. W. & Shankland, S. J. Experimental Models of Membranous Nephropathy. *Drug Discov Today Dis Models* **7**, 27-33 (2010). https://doi.org:10.1016/j.ddmod.2010.11.001

40 Yang, H. *et al.* Podocyte programmed cell death in diabetic kidney disease: Molecular mechanisms and therapeutic prospects. *Biomedicine & Pharmacotherapy* **177**, 117140 (2024). https://doi.org:https://doi.org/10.1016/j.biopha.2024.117140

41 Marshall, C. B. Rethinking glomerular basement membrane thickening in diabetic nephropathy: adaptive or pathogenic? *Am J Physiol Renal Physiol* **311**, F831-f843 (2016). https://doi.org:10.1152/ajprenal.00313.2016

42 Bokhari, S. R. A., Zulfiqar, H. & Hariz, A. in *StatPearls* (StatPearls Publishing Copyright © 2025, StatPearls Publishing LLC., 2025).

43 Levstek, T. *et al.* Identification of genetic variants associated with Fabry nephropathy progression using whole-exome sequencing. *Mol Genet Metab* **145**, 109191 (2025). https://doi.org:10.1016/j.ymgme.2025.109191

44 Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

45 Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* **40**, 834-848 (2017).

46 Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. in *Proceedings of the European conference on computer vision (ECCV).* 801-818.

47 Zou, A. *et al.* Developing a Novel Deep Learning-Based Method for Kidney Transmission Electron Microscopy Image Analysis: TH-PO029. *Journal of the American Society of Nephrology* **35** (2024). https://doi.org:10.1681/ASN.20246p57j8k8

48 Graham, L. & Orenstein, J. M. Processing tissue and cells for transmission electron microscopy in diagnostic pathology and research. *Nat Protoc* **2**, 2439-2450 (2007). https://doi.org:10.1038/nprot.2007.304

49 Pisarek-Horowitz, A. *et al.* Loss of Roundabout Guidance Receptor 2 (Robo2) in Podocytes Protects Adult Mice from Glomerular Injury by Maintaining Podocyte Foot Process Structure. *Am J Pathol* **190**, 799-816 (2020). https://doi.org:10.1016/j.ajpath.2019.12.009

50 Salant, D. J. & Cybulsky, A. V. Experimental glomerulonephritis. *Methods Enzymol* **162**, 421-461 (1988). https://doi.org:10.1016/0076-6879(88)62096-9

51 Wada, K. Labelme: Image Polygonal Annotation with Python. https://doi.org:0.5281/zenodo.5711226

52 Veron, D. *et al.* Overexpression of VEGF-A in podocytes of adult mice causes glomerular disease. *Kidney Int* **77**, 989-999 (2010). https://doi.org:10.1038/ki.2010.64

53 Lahdenkari, A. T. *et al.* Podocytes are firmly attached to glomerular basement membrane in kidneys with heavy proteinuria. *J Am Soc Nephrol* **15**, 2611-2618 (2004). https://doi.org:10.1097/01.Asn.0000139478.03463.D9

54 Shea, S. M. & Morrison, A. B. A stereological study of the glomerular filter in the rat. Morphometry of the slit diaphragm and basement membrane. *J Cell Biol* **67**, 436-443 (1975). https://doi.org:10.1083/jcb.67.2.436

Methods

Sample Preparation and Image Acquisition

ILK podocyte specific knockout (ILK cKO) and Col4a3 knockout (Col4a3 KO) models. The ILK cKO mice were generated by crossing ILKflox/flox mice with Podocin-Cre mice[37]. Col4a3 cKO mice were purchased from the Jackson laboratory (RRID:IMSR_JAX:002908). The littermate wild type mice were used as healthy controls. Mice were sacrificed at 16 weeks old of age and kidneys were dissected. The fresh kidney samples were fixed in Karnovsky's fixative and processed according to the published protocol[48]. In brief, small pieces of kidney samples (<1 mm thick) were fixed in buffered 2.5% glutaraldehyde. Kidney specimens were then post fixed in 1% osmium tetroxide (Electron Microscopy Sciences), dehydrated, embedded in propylene oxide/Epon resin, and sectioned (60 to 90 nm thick) using the ultramicrotome. Collected sections were post stained with uranyl acetate and lead citrate (Electron Microscopy Sciences), according to the published protocol[48]. Images were obtained using the Jeol JEM-1011 transmission electron microscope (JEOL USA Inc.) equipped with an Erlangshen ES100W digital camera (Gatan, Pleasanton, CA). The operator was blinded as to the genotype and the injury status of the animals at the time of image collection and analyses[49].

Rat PHN injury model. A modified PHN injury model was induced in Lewis rats (strain code 004; Charles River Laboratories, Wilmington, MA) by one dose of i.v. injection [1.5 to 2 mL/kg sheep anti-rat Fx1A serum (PTX-002S; Probetex, San Antonio, TX)], according to Probetex and published protocol[50]. Rats were euthanized at day 16 after anti-rat Fx1A serum injection, and kidneys were collected and fixed in Karnovsky's fixative for TEM analysis. Sterile 0.9% sodium chloride (normal saline) was used in control rats. A total of six animals in each experimental group were studied. Kidney samples were prepared, and TEM imaging was performed as described above.

Human diseases: Electron microscopy (EM) images used in this study were obtained from an open source dataset described in the publication of Smekous et al[29].The dataset includes kidney biopsy samples from patients with Fabry disease, type 2 diabetes, and minimal change disease, as well as healthy controls from living kidney transplant donors. All biopsies were processed at eight EM laboratories following standardized protocols. Ultrathin glutaraldehyde fixed sections were stained with toluidine blue for glomerular identification, and high magnification transmission EM images (original magnification ×30,000) were captured using a systematic uniform random sampling strategy. Each glomerulus was reviewed by expert technicians under masked conditions to estimate average FP width and GBM width.

Label Generation, Manual Measurement, and Image Prepressing

Ground Truth Label. All TEM images were divided into training, validation, and test sets as summarized in Table 1. The training/validation and test sets were intentionally sourced from different cohorts, encompassing distinct species and disease types, to ensure that the test set could robustly evaluate the generalization capability of our algorithm. For the training and validation sets, regions corresponding to the GBM and FP were manually annotated using LabelMe software[51]. The annotations were performed by trained researchers using high resolution TEM images and subsequently reviewed by renal pathologists to ensure consistency and structural accuracy. This manual labeling process served as the foundation for model training and validation, while the independently collected test data were reserved for assessing model performance across unseen imaging and disease conditions.

Manual measurement for GBM and FP widths. Manual podocyte foot process width was traced and measured on ×12,000 high magnification transmission EM images using ImageJ software version 1.47v, according to published methods[49]. Briefly, the total length of the GBM spanning the area of the slit diaphragms was measured and combined in each image using ImageJ, according to published methods[52-54]. The results were then pasted on Excel sheet and adjusted for scale bar. After that, all the slit diaphragms covering the GBM were counted, and FP width was calculated using the total length of GBM divided by total number of slit diaphragms in each image. The GBM width was obtained by measuring the cross length from the podocyte membrane to the endothelial membrane. Each image was randomly measured 20 areas around the GBM and the harmonic mean was calculated for each image.

Preparing training and testing datasets for the deep learning framework. Images were normalized by mean-std normalization and tiled into [size, e.g., 1024×1024, 2624*2624] with 50% overlap. Scale bars and metadata were parsed to record the pixel to nm factor (e.g.,4.122 nm/pixel for mice, 3.88 nm/pixel for mice) used for unit conversion. To avoid leakage, splits were patient/animal level stratified into train/val with ratios 80/20 while preserving class balance per species and disease model. All preprocessing scripts are provided (see Code availability).

Hybrid Data Preparation. To improve the model's adaptability across images of different resolutions, we adopted a stage wise data preparation and training strategy. High resolution TEM images were first used to pretrain the backbone network, enabling the model to learn fine grained structural features of GBM and FP. Subsequently, lower resolution images were used for fine tuning to encourage scale invariant feature learning and improve cross magnification generalization. During both training stages, various data augmentation techniques were applied, including random scaling, random cropping, horizontal and vertical flipping, and random adjustments of brightness and contrast. These augmentations helped prevent overfitting and further enhanced the model's robustness to variations in image scale and quality.

Ultrastructure Quantification Based on Clinical Protocols. GBM thickness and FP width were measured using TEM images analyzed in ImageJ (FIJI). Images were calibrated using reference scale bars to ensure accurate morphometric analysis. For GBM thickness measurement, a standardized 1.5 µm × 1.5 µm grid was superimposed on micrographs for systematic sampling. At each grid intersection, perpendicular lines were drawn from the endothelial cell membrane to the outer margin of the lamina rara externa beneath epithelial foot processes. Measurements were taken only in well preserved regions, excluding artifacts from sectioning angle, folding, or detachment. The Orthogonal Intercept Method (OIM) was applied to calculate mean GBM thickness. For FPW measurement, the lateral edges of individual foot processes were manually identified, and width was measured as the linear distance between adjacent borders. Only clearly defined and intact foot processes were included in the analysis.

GLome2Mask: TEM Image Panoptic Segmentation Model Architecture

Our proposed Glom2Mask model is designed based on the Mask2Former panoptic segmentation framework, enabling comprehensive segmentation of GBM and FP in TEM images. Given the distinct structural characteristics of these components, we categorize GBM as a "stuff" class for semantic segmentation and FP as a "things" class for instance segmentation. This formulation ensures precise differentiation between the continuous structure of the GBM and the discrete FP instances within the glomerular d barrier.

To enhance segmentation accuracy, we integrate HRNet as the backbone for high resolution feature extraction. HRNet is well suited for this task due to its ability to maintain spatial details throughout the network, which is essential for capturing the fine structural differences in TEM images. Unlike traditional backbones that down sample feature maps at early stages, HRNet maintains high resolution representations, allowing for better delineation of GBM boundaries and accurate identification of individual FP structures.

Our model follows a transformer based approach, where multi scale features extracted by HRNet are processed through multi head attention mechanisms to generate high quality segmentation masks. The network outputs a set of learned queries, each corresponding to a GBM region or an FP instance, enabling robust panoptic segmentation across diverse TEM datasets, including mouse, rat, and human kidney samples.

By leveraging HRNet for high resolution feature extraction and Mask2Former for unified panoptic segmentation, our model effectively segments GBM and FP structures, providing the necessary groundwork for precise automated width measurement in kidney disease analysis.

Model Training for the Deep Learning Framework

Our training process consists of two stages: a self-training based semantic segmentation stage and a panoptic segmentation stage using Glom2Mask. The first stage aims to generate additional pseudo labeled data while simultaneously learning a high resolution backbone network. In the second stage, we utilize the HRNet trained in the first stage as the backbone within a modified Mask2Former model for panoptic segmentation, improving both training efficiency and segmentation accuracy.

Self-Training for Semantic Segmentation. In the first stage, we employ a self-supervised learning approach for semantic segmentation based on HRNet.

For cases with a limited number of annotated images (e.g., 19, 48, or 64 labeled samples), we employed a self-training strategy to improve segmentation performance. Self-training is a semi-supervised learning technique that utilizes both labeled and unlabeled data. Initially, the model is trained using only the labeled dataset to capture fundamental feature representations. After this initial training, the trained model generates pseudo labels for the unlabeled data. These pseudo labeled samples, combined with the original labeled dataset, are then used to retrain the model iteratively, progressively refining its segmentation accuracy and generalization capability.

The model performance was evaluated using mean Dice score. Each manual segmentation mask, per category, was compared to the prediction mask of the model. Dice score was calculated by,

$$Dice = \frac{2|A_i \cap P_i|}{|A_i| + |P_i|}$$

Where A_i as the ground truth segmentation result for class i, P_i as the predicted segmentation result for class i. Dice score values range from 0 to 1, where 0 represents no pixel overlap and 1 represents perfect pixel overlap between the two masks.

During training, we applied a series of data augmentation techniques to improve model robustness, including random scaling within a range of 0.5–2.0, random rotation within ±10°, horizontal and vertical flipping with a probability of 0.5 each, random cropping to a size of 1024×1024 pixels, and random intensity jitter with brightness and contrast factors between 0.8

and 1.2. The model was optimized using the AdamW optimizer with a learning rate of 6×10□□ and a weight decay of 0.05. A polynomial learning rate decay schedule with a linear warm up phase of 1,500 iterations was employed. Training was conducted with a batch size of 16 for a total of 80,000 iterations. To address class imbalance, a combination of cross entropy loss and focal loss ($\gamma = 2.0$, $\alpha = 0.25$) was used.

Panoptic Segmentation with Glom2Mask. In the second stage, The training setup follows the same training validation split as in the first step, ensuring consistency in dataset partitioning. Data augmentation techniques, including random scaling, cropping, rotation, flipping, and contrast adjustments, are applied during training to enhance generalization similar with step 1. The model was optimized using the AdamW optimizer with a learning rate of 1 × 10□□ and a weight decay of 0.05 for a total of 160,000 iterations. The training objective combined three loss functions including classification cross entropy, mask cross entropy, and Dice loss with relative weights of 2.0: 5.0: 1.0. Early stopping and checkpoint selection were determined based on the PQ metric evaluated on the validation set.

The loss functions for training are as follows: Segmentation classification head: Cross entropy loss, ensuring accurate categorical predictions for GBM and FP. Mask prediction head: A combination of cross entropy loss and Dice loss, improving both pixel wise accuracy and mask shape consistency. To quantitatively assess model performance, we employ the following panoptic segmentation metrics:

- Panoptic Quality (PQ): Measures the overall segmentation accuracy, balancing segmentation precision (true positive detection) and quality (IoU based mask accuracy). Defined as:

$$PQ = \frac{\sum_{(p,g)\in TP} IoU(p,g)}{|TP| + 0.5|FP| + 0.5|FN|}$$

where TP (true positives) are correctly matched masks, FP (false positives) are extra masks, and FN (false negatives) are missed masks.

- Segmentation Quality (SQ): Measures the average mask accuracy of correctly predicted objects:

$$SQ = \frac{\sum_{(p,g)\in TP} IoU(p,g)}{|TP|}$$

- Recognition Quality (RQ): Measures the correctness of mask instance association and is computed as:

$$RQ = \frac{|TP|}{|TP| + 0.5|FP| + 0.5|FN|}$$

Training Environment and Reproducibility

All experiments were run on Ubuntu 22.04 with Python 3.9, PyTorch 2.1.0, CUDA 12.1, cuDNN 8.9, MMDetection 2.0.1 and MMCV. Hardware:1*NVIDIA A100-80G GPU, Inter Xeon Gold 6338 CPU (40 cores). Deterministic behavior was enforced via fixed seeds (seed=42) and disabling nondeterministic cuDNN ops. Full environmental files, exact config YAMLs, and trained checkpoints are provided (see Code availability).

Post Segmentation Processing and Quantification

GBM Width Measurement. The second stage of our proposed method consists of a Python image processing algorithm that takes input of the GBM and FP segmentation tiles predicted by the deep learning model to measure mean GBM width and FPW. We will describe GBM width measurement first here. The mean GBM width for each tile is then calculated as its total GBM area divided by its total GBM central length according to Equation 1:

$$GBM\ width = \frac{GBM\ area}{GBM\ central\ length}$$

Where, the area of the GBM is straightforward to compute; it involves directly counting the number of pixels in a section of the GBM. The process to calculate the length of the GBM is as follows: First, we obtain the skeleton of the GBM using the skeletonize method with Python package scikit image. The general skeleton obtained through the skeletonize method, in addition to the central length of the GBM, also includes many small branching points. To eliminate these branching points, we construct an undirected cyclic graph of the GBM skeleton using the Python package NetworkX. Here, there are two scenarios: if the GBM forms a loop structure, we iteratively prune nodes with an in degree of one from the undirected cyclic graph of the GBM skeleton until no nodes with an in degree of one remain. We then calculate the perimeter of the resulting graph to determine the central length of the GBM; if the GBM forms a strip like structure, we calculate the shortest distance between any two endpoints and identify the two endpoints with the greatest distance as the central length of the GBM. Lastly, the specimen mean GBM width was converted from pixels to distance with a conversion factor derived from the TEM image scale bar, for example, 3.88 nm/pixel for rat model data.

FP Width Measurement. Following this, the FP width calculation is more direct. Each FP instance is represented by a unique integer starting from 1, with the background assigned a value of 0.

Next, we utilize morphological operations in OpenCV to dilate the GBM, allowing us to derive the contour of the dilated GBM. In this processed image, the background remains coded as 0, and the contour positions are coded as 1.

Finally, we calculate the width of each FP by multiplying the FP instance matrix with the matrix of the dilated GBM contours. This operation gives us the width by calculating the pixel length of FPs. Similarly, the specimen FP widths were converted from pixels to distance with a conversion factor derived from the TEM image scale bar, for example, 3.88 nm/pixel for rat model data.

Contributions

W.L., C.Z., A.Z., and X.F. conceived the study. A.Z. developed the deep learning model. X.F. and H.C. acquired the images from animal models. X.F., A.Z., W.T., J.J., L.D., and E.O. labeled, curated, and measured the images. A.Z., W.T., J.J., and C.Z. performed the data analysis. F.R.M. and J.M.H. provided clinical consultation and measurements. A.Z., C.Z., X.F., W.T., and J.J. generated the figures and wrote the original draft. C.Z., W.L., X.F., and A.Z. supervised the work.

Data availability

The datasets supporting this study's findings are publicly accessible and include three infrared and visible image datasets: human disease dataset (https://github.com/najafian-lab/fpw-dl-v1), mice and rat disease model dataset (https://github.com/CZCBLab/TEAMKidney)

Code availability

The code is available at https://github.com/CZCBLab/TEAMKidney.

Figure legends

Fig. 1. Lack of existing models for accurate TEM image analysis and overview of the proposed framework. (A) Example TEM image of a podocyte specific ILK conditional knockout mouse. (B) The boxed region of the glomerular filtration barrier in (A) with manually measured FP in red and GBM in blue. (C) Ground truth segmentation of FP and GBM. Segmentation results obtained using (D) FPW-DL, (E) SAM, and (F) MedSAM models. (G) Workflow of the proposed TEAMKidney framework for augmentation, segmentation, and quantitative analysis of mouse, rat, and human TEM images. (H) The proposed Glom2Mask network architecture for segmentation. (I) Example image showing post segmentation of GBM and FP. (J) Boxed region in (I) illustrating the initial skeleton extraction. (K) Demonstration of skeleton extraction error correction by the post segmentation analysis algorithm.

Fig. 2. Detailed structure and evaluation of the TEAMKidney framework. (A) Workflow of the segmentation stage incorporating a self-training strategy. (B) Performance and training time comparison of different backbone networks for semantic segmentation across datasets. (C) Representative TEM image segmented by the three backbone networks for qualitative comparison. (D) Performance and training time comparison between the proposed Glom2Mask model and the benchmark Mask2Former for panoptic segmentation. (E) Representative image showing segmentation results from Glom2Mask and Mask2Former. The GBM is segmented as a stuff class (blue), and FP as thing instances (various shades of red). (F) Comparison of GBM and FP measurements between the TEAMKidney algorithm and ground truth across different species.

Fig. 3. Application of TEAMKidney to three animal models of kidney disease. (A) Schematic illustrating two mouse models of glomerular disease: knocking out Col4a3 or specific podocyte ILK knocking out. (B) Schematic of a rat model of glomerular disease (PHN) induced by anti-Fx1a antibody injection. (C) Representative TEM images analyzed by the TEAMKidney model for WT, Col4a3 KO, and ILK cKO mice. (D) Violin plots showing quantitative measurements of FP and GBM widths across WT, Col4a3 KO, and ILK cKO mice. (E) Representative TEAMKidney predicted images for WT and PHN rat samples. (F) Violin plots comparing FP and GBM widths between WT and PHN groups.

Fig. 4. Application of TEAMKidney to human kidney diseases. (A) Schematic illustration of the pathogenesis of DKD and Fabry disease. (B) Representative clinical TEM images from normal, Fabry disease, and DKD patients analyzed using the TEAMKidney model. (C) Violin plots showing the average GBM and FP widths per glomerulus across normal, Fabry disease, and

DKD samples. Each point represents the mean width for one glomerulus. (D) Representative TEAMKidney predicted images from male and female Fabry disease patients. (E) Violin plots comparing GBM and FP widths between male and female Fabry disease groups. (F) Violin plots illustrating the distributions of GBM and FP widths across different magnifications in human kidney TEM images.

Fig. 5. Comparison of FP quantification between TEAMKidney and FPW-DL methods. Scatter plots with diagonal reference lines show the agreement between model predictions and ground truth measurements for (A) average FP width and (B) FP counts. (C) Representative TEM images with segmentation masks generated by TEAMKidney (left) and FPW-DL (right). From top to bottom: normal human kidney, human DKD, ILK cKO mouse model, and PHN rat model.

Fig. 6. Evaluation of concordance between expert manual measurements and TEAMKidney predictions on human clinical TEM images. (A) Normal kidney samples. (B) DKD samples. (C) Fabry disease samples from male patients. (D) Fabry disease samples from female patients. From top to bottom: GBM mean width, FP mean width, and FP count.

Fig. 7. TEAMKidney aligned with clinical routine measurement protocols. (A) Schematic illustration of the clinical manual annotation process for GBM and FP. Representative example showing (B) the original TEM image, (C) GBM measurement following the clinical manual procedure, (D) FP counting following the clinical manual procedure, and (E) segmentation masks predicted by TEAMKidney.

Table 1. Selection Data and Criteria.

Table 2. Comparison Between Proposed Model and Other Published Groups.

Figure 1

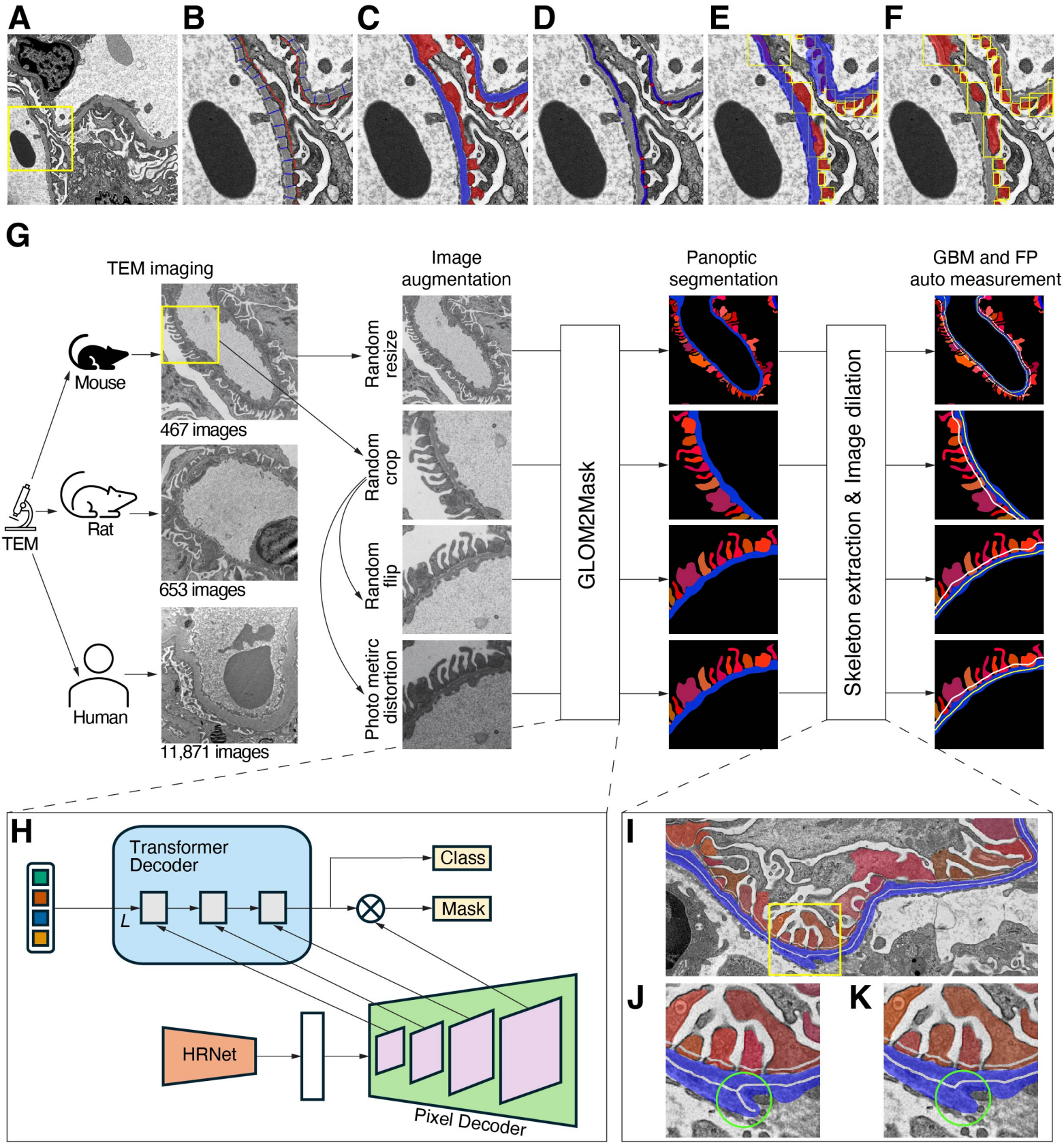
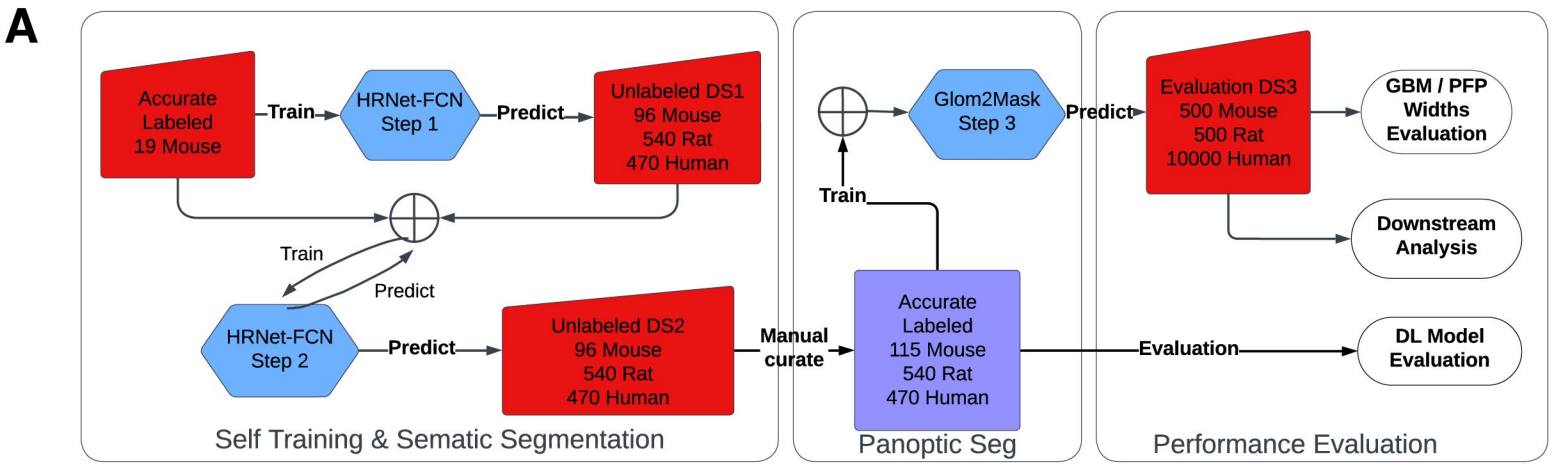
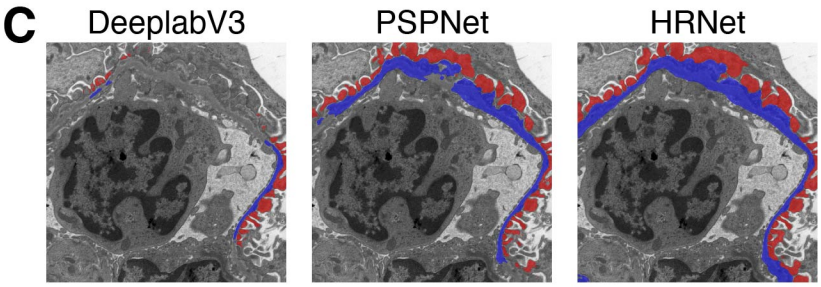

A
B
C
D
E
F
G
TEM imaging
Image augmentation
Panoptic segmentation
GBM and FP auto measurement
Mouse
467 images
Rat
653 images
TEM
Human
11,871 images
Random resize
Random crop
Random flip
Photo metirc distortion
GLOM2Mask
Skeleton extraction & Image dilation
H
Transformer Decoder
L
Class
Mask
HRNet
Pixel Decoder
I
J
K

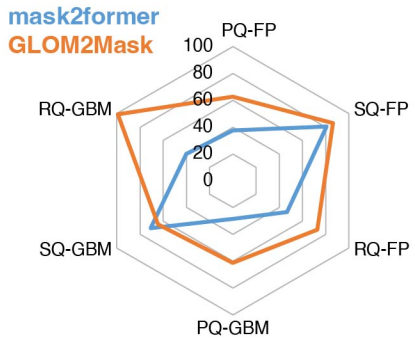
Figure 2

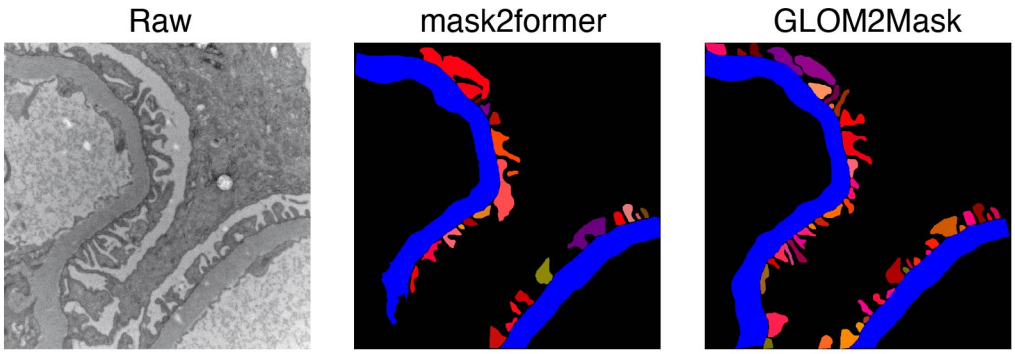


B

Model	Mouse		Rat		Human		Time (h)
	GBM	FP	GBM	FP	GBM	FP	
DeeplabV3	0.815	0.882	0.861	0.912	0.768	0.886	34.3
PSPNet	0.833	0.891	0.882	0.951	0.771	0.893	39.47
HRNet	0.838	0.892	0.917	0.96	0.768	0.892	22







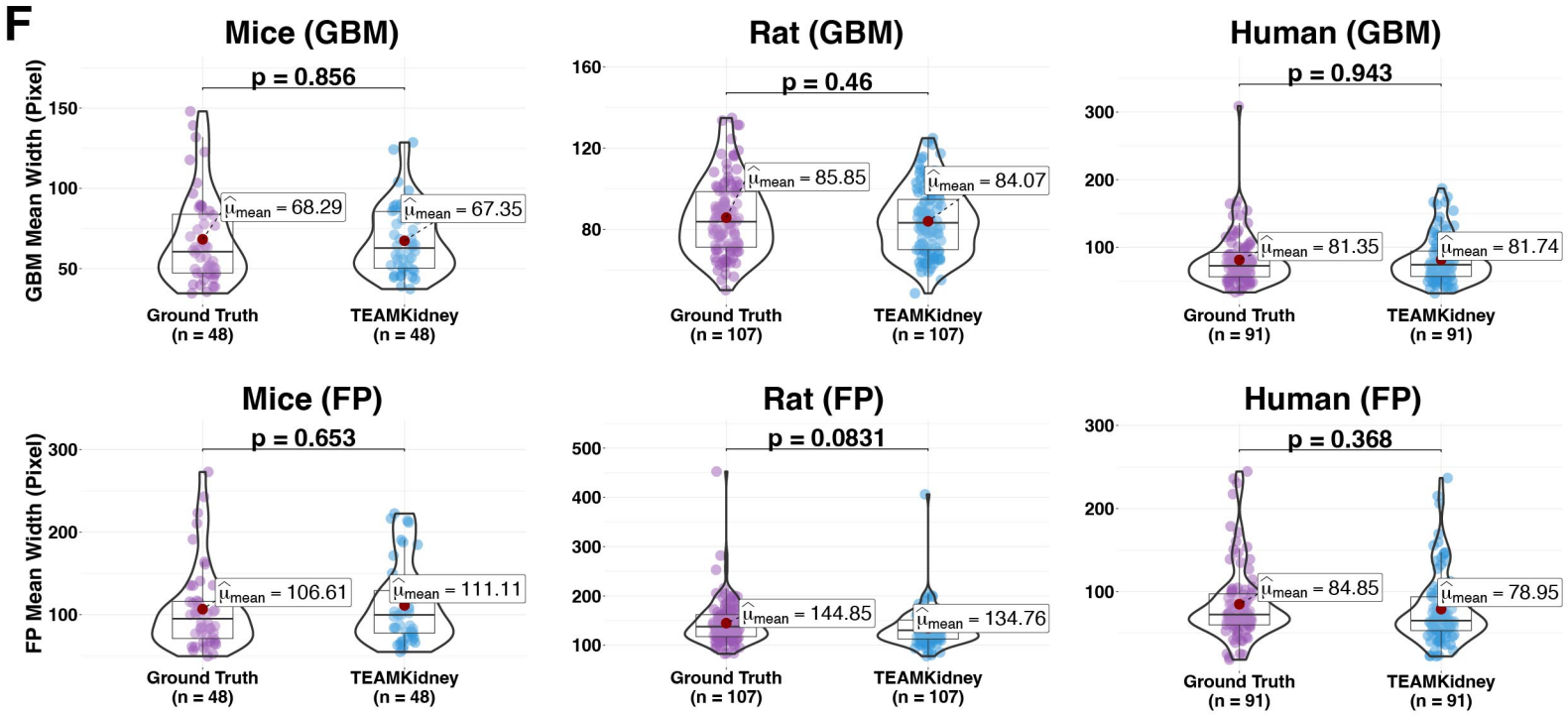


Figure 3

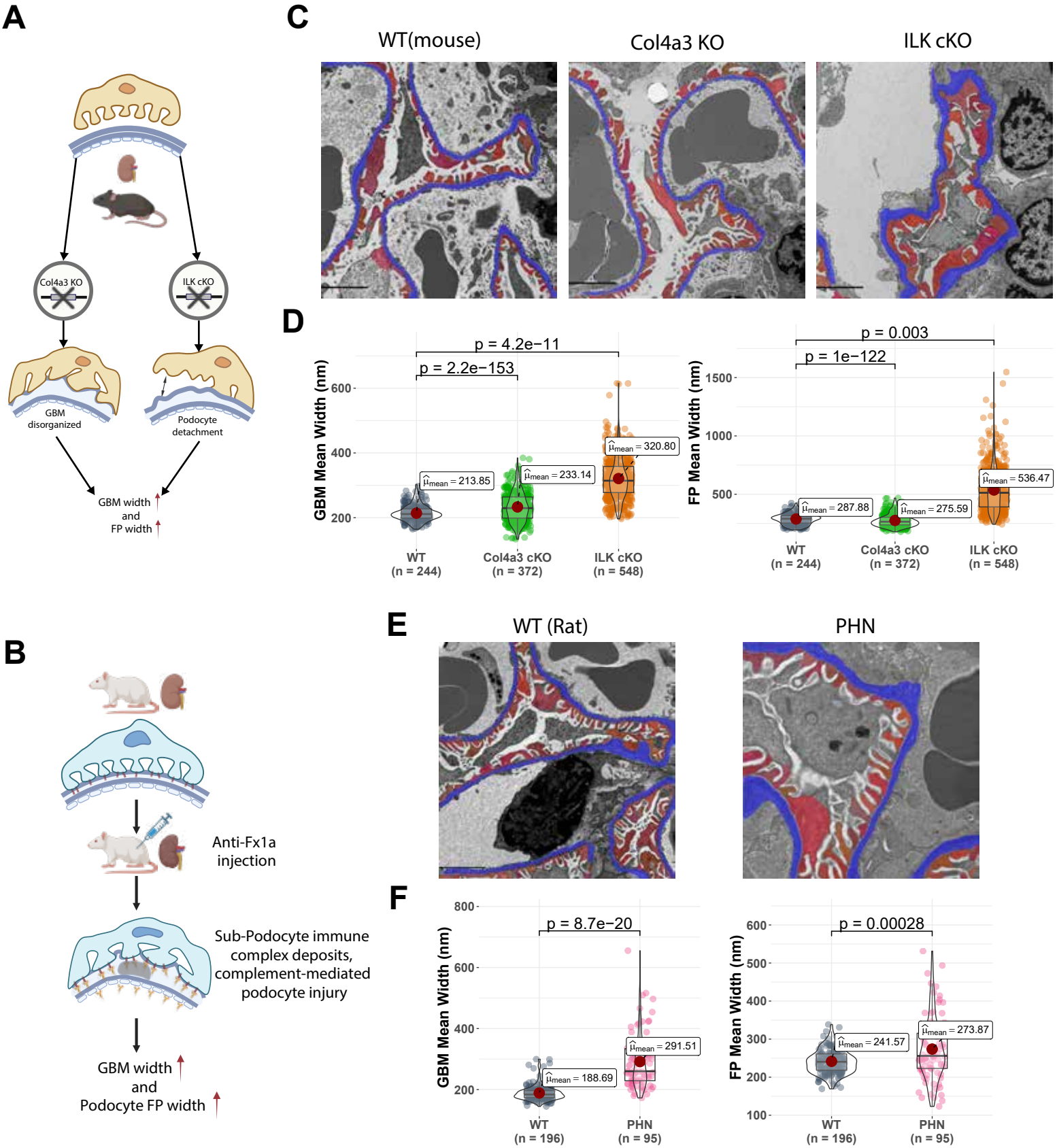

A
Col4a3 KO
ILK cKO
GBM disorganized
Podocyte detachment
GBM width
and
FP width
B
Anti-Fx1a injection
Sub-Podocyte immune complex deposits, complement-mediated podocyte injury
GBM width
and
Podocyte FP width
C
WT(mouse)
Col4a3 KO
ILK cKO
D
GBM Mean Width (nm)
p = 4.2e−11
p = 2.2e−153
μ̂mean = 213.85
μ̂mean = 233.14
μ̂mean = 320.80
WT (n = 244)
Col4a3 cKO (n = 372)
ILK cKO (n = 548)
FP Mean Width (nm)
p = 0.003
p = 1e−122
μ̂mean = 287.88
μ̂mean = 275.59
μ̂mean = 536.47
E
WT (Rat)
PHN
F
GBM Mean Width (nm)
p = 8.7e−20
μ̂mean = 188.69
μ̂mean = 291.51
WT (n = 196)
PHN (n = 95)
FP Mean Width (nm)
p = 0.00028
μ̂mean = 241.57
μ̂mean = 273.87

Figure 4

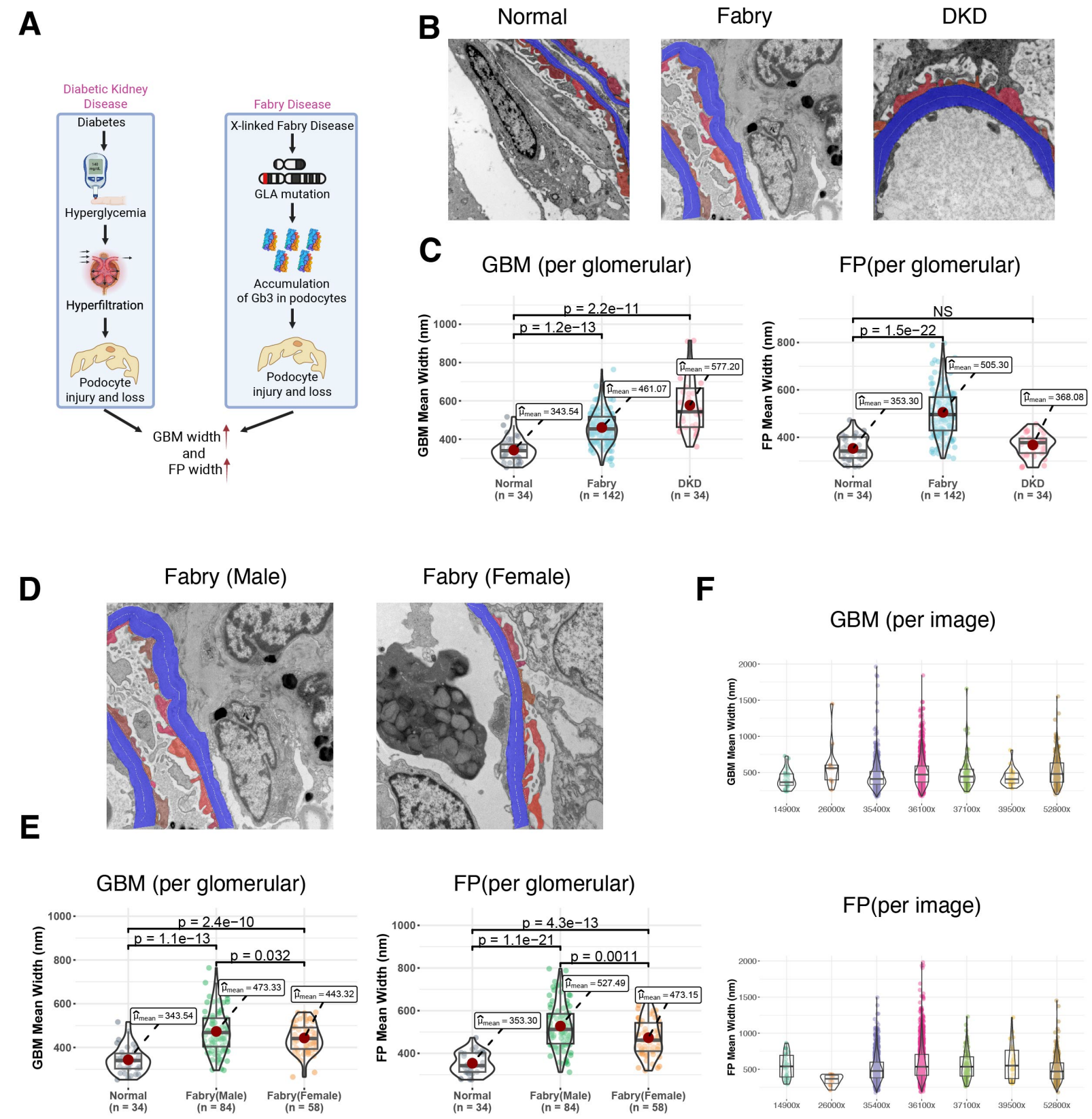

A
Diabetic Kidney Disease
Diabetes
Hyperglycemia
Hyperfiltration
Podocyte injury and loss
Fabry Disease
X-linked Fabry Disease
GLA mutation
Accumulation of Gb3 in podocytes
Podocyte injury and loss
GBM width and FP width
B
Normal
Fabry
DKD
C
GBM (per glomerular)
GBM Mean Width (nm)
p = 2.2e−11
p = 1.2e−13
$\hat{\mu}_{mean}$ = 343.54
$\hat{\mu}_{mean}$ = 461.07
$\hat{\mu}_{mean}$ = 577.20
Normal (n = 34)
Fabry (n = 142)
DKD (n = 34)
FP(per glomerular)
FP Mean Width (nm)
NS
p = 1.5e−22
$\hat{\mu}_{mean}$ = 353.30
$\hat{\mu}_{mean}$ = 505.30
$\hat{\mu}_{mean}$ = 368.08
Normal (n = 34)
Fabry (n = 142)
DKD (n = 34)
D
Fabry (Male)
Fabry (Female)
E
GBM (per glomerular)
GBM Mean Width (nm)
p = 2.4e−10
p = 1.1e−13
p = 0.032
$\hat{\mu}_{mean}$ = 343.54
$\hat{\mu}_{mean}$ = 473.33
$\hat{\mu}_{mean}$ = 443.32
Normal (n = 34)
Fabry(Male) (n = 84)
Fabry(Female) (n = 58)
FP(per glomerular)
FP Mean Width (nm)
p = 4.3e−13
p = 1.1e−21
p = 0.0011
$\hat{\mu}_{mean}$ = 353.30
$\hat{\mu}_{mean}$ = 527.49
$\hat{\mu}_{mean}$ = 473.15
Normal (n = 34)
Fabry(Male) (n = 84)
Fabry(Female) (n = 58)
F
GBM (per image)
GBM Mean Width (nm)
14900x
26000x
35400x
36100x
37100x
39500x
52800x
FP(per image)
FP Mean Width (nm)
14900x
26000x
35400x
36100x
37100x
39500x
52800x

Figure 5

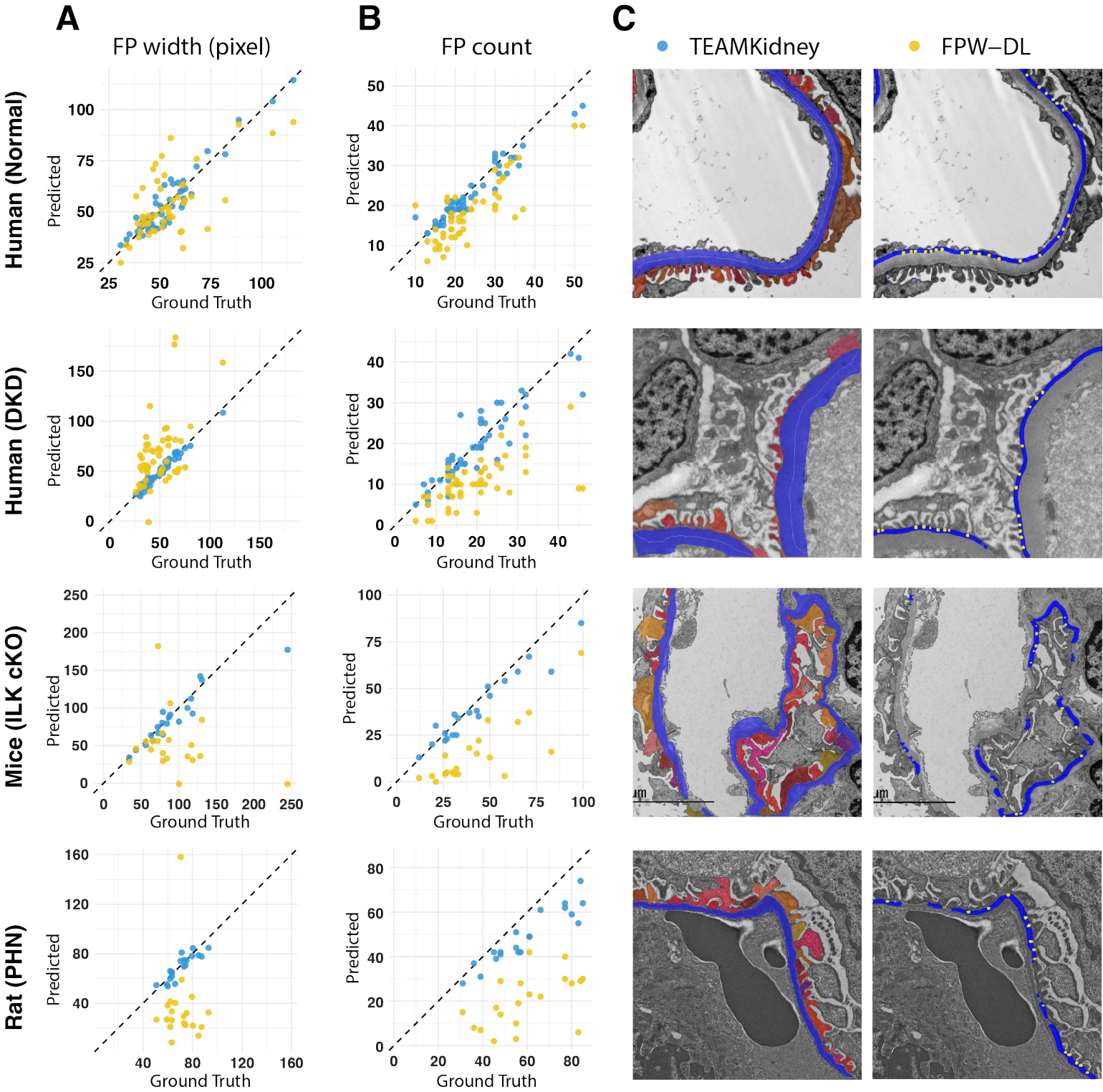

A
FP width (pixel)
B
FP count
C
TEAMKidney
FPW–DL
Human (Normal)
Human (DKD)
Mice (ILK cKO)
Rat (PHN)
Predicted
Ground Truth

Figure 6

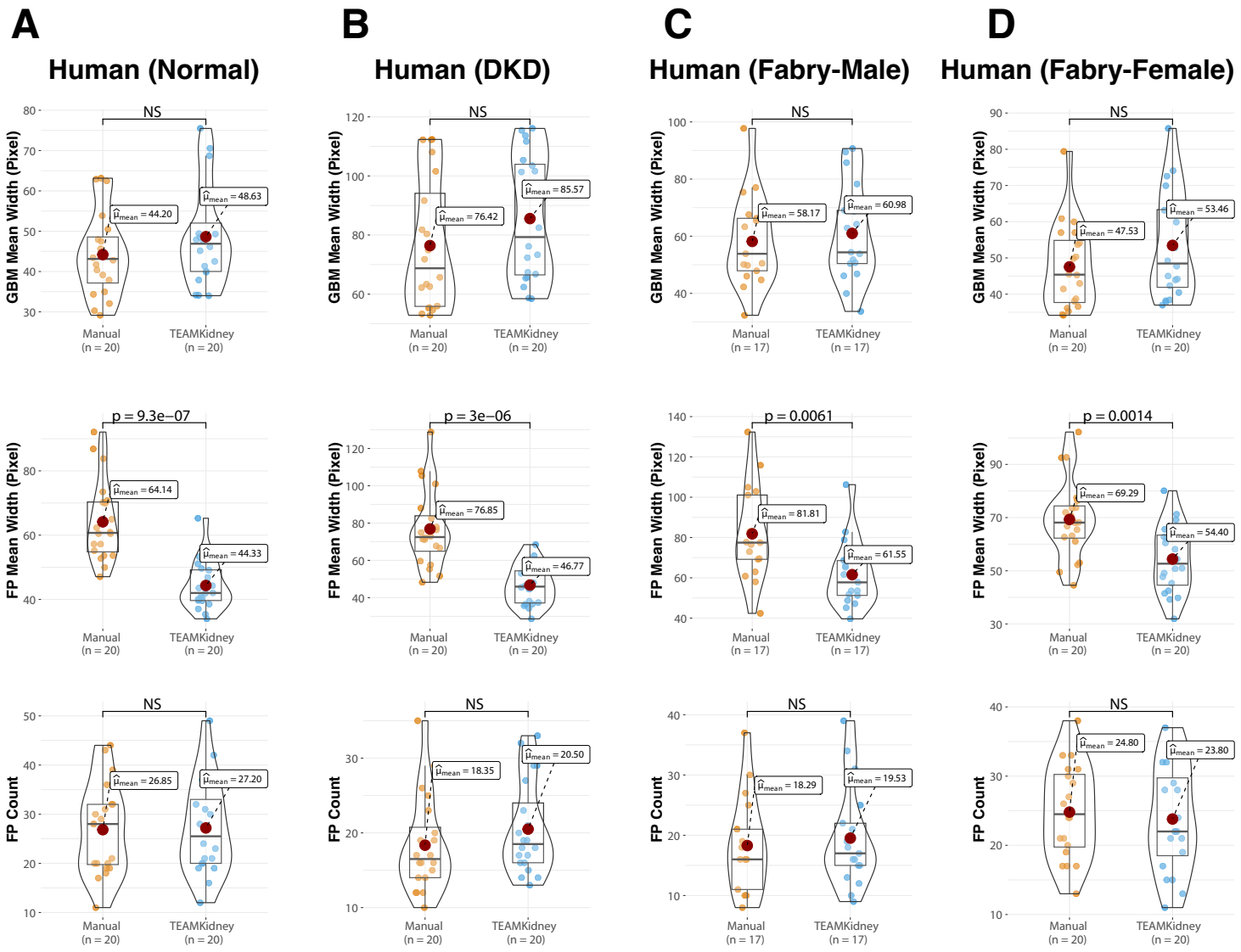
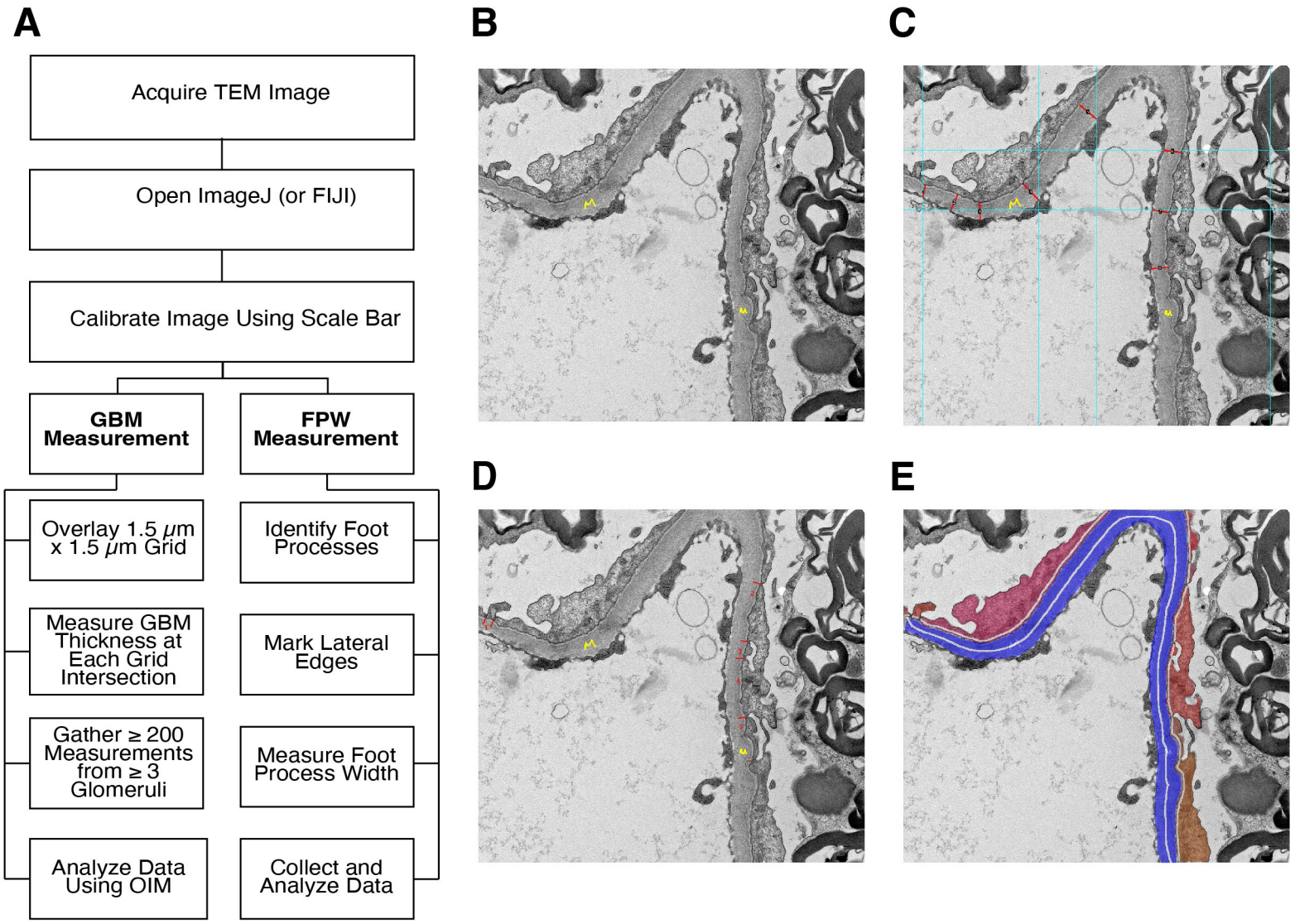

A
Human (Normal)
B
Human (DKD)
C
Human (Fabry-Male)
D
Human (Fabry-Female)
GBM Mean Width (Pixel)
FP Mean Width (Pixel)
FP Count
NS
p = 9.3e−07
p = 3e−06
p = 0.0061
p = 0.0014
Manual (n = 20)
TEAMKidney (n = 20)
Manual (n = 17)
TEAMKidney (n = 17)
μ̂mean = 44.20
μ̂mean = 48.63
μ̂mean = 76.42
μ̂mean = 85.57
μ̂mean = 58.17
μ̂mean = 60.98
μ̂mean = 47.53
μ̂mean = 53.46
μ̂mean = 64.14
μ̂mean = 44.33
μ̂mean = 76.85
μ̂mean = 46.77
μ̂mean = 81.81
μ̂mean = 61.55
μ̂mean = 69.29
μ̂mean = 54.40
μ̂mean = 26.85
μ̂mean = 27.20
μ̂mean = 18.35
μ̂mean = 20.50
μ̂mean = 18.29
μ̂mean = 19.53
μ̂mean = 24.80
μ̂mean = 23.80

Figure 7


A
Acquire TEM Image
Open ImageJ (or FIJI)
Calibrate Image Using Scale Bar
GBM Measurement
FPW Measurement
Overlay 1.5 μm x 1.5 μm Grid
Measure GBM Thickness at Each Grid Intersection
Gather ≥ 200 Measurements from ≥ 3 Glomeruli
Analyze Data Using OIM
Identify Foot Processes
Mark Lateral Edges
Measure Foot Process Width
Collect and Analyze Data
B
C
D
E