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], Sarah R. Vargas[4], Stephen P. Berasi[4], Hongying Yang[4], Hui Chen[3], Joel M. Henderson[3,5,6], Xueping Fan[2,*], Weining Lu[2,3,*], Chao Zhang[1,*]

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

[4]Global Pathology and Centers for Therapeutic Innovation, Pfizer, Inc., Groton, CT, and Cambridge, MA, USA.

[5]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA, USA

[6]Department of Pathology, Harvard Medical School, 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

Transmission electron microscopy (TEM) has become an essential technique for observing subcellular ultrastructure, and is widely used in both clinical diagnosis and biomedical research. However, analysis of TEM data remains extremely labor-intensive and often inconsistent across operators due to the lack of dedicated computational methods. Here, we present TEAMKidney, a deep learning framework for accurate and scalable measurement of ultrastructures in TEM images across species, magnifications, and instrument platforms. We collected 12,991 TEM images from patients with multiple kidney diseases and from different animal models. By combining a self-training-based semantic segmentation stage with a TEM-tailored panoptic segmentation model, we address two major challenges in TEM data analysis: the lack of accurately labeled training data and the difficulty of achieving high segmentation accuracy for complex ultrastructure. Application of TEAMKidney to both human and animal images successfully reveals disease-associated changes in two critical glomerular ultrastructures: the glomerular basement membrane and podocyte foot processes. In addition to significantly outperforming existing tools, TEAMKidney shows close agreement with pathological expert measurements used in clinical assessment protocols. By reducing dependence on manual tracing while preserving expert-level accuracy, TEAMKidney demonstrates that deep learning can substantially reduce the burden of image analysis in both clinical pathology and biomedical research settings.

Introduction

Transmission electron microscopy (TEM) provides high-resolution visualization of cellular ultrastructure beyond the resolving capabilities of light microscopy, immunofluorescence (IF) staining, immunohistochemistry, or molecular assays[1, 2] (Fig 1A). Beyond its extensive use in biomedical research and drug discovery, TEM remains an important diagnostic tool in many clinical areas, including infectious diseases, oncology, ophthalmology, neurology, and especially renal pathology[3-5].

Chronic kidney disease (CKD) affects over 37 million people in the US[6] and 850 million worldwide[7, 8]. It has been recognized as a leading public health burden in the US. By 2027, the direct healthcare costs for kidney diseases in the US were projected to reach $173.4 billion[9]. Diabetic kidney disease and glomerular disease account for more than 50% of CKD cases[10, 11] and are the leading cause of kidney failure[12, 13]. These disorders encompass a broad spectrum of conditions characterized by distinct patterns of glomerular ultrastructural damage, which are associated with different therapeutic strategies and clinical outcomes. In the past decade, TEM has become an essential and often decisive component of the diagnostic modalities for kidney diseases[14]. Some major nephrology and pathology centers now utilize TEM as a routine diagnostic tool, applying it to more than 80% renal biopsies[14, 15].

Despite its value, significant barriers limit the broader adoption of TEM. The technique remains highly labor-intensive[14] and producing a full pathology report can take up to 26 working days on average[5], primarily due to the need for expert renal pathologists to analyze the ultrastructure, such as glomerular basement membrane (GBM) and podocyte foot processes (PFP). Measurement of GBM and PFP widths is widely used for both research and clinical diagnosis, yet researchers and renal pathologists still rely primarily on manual measurement of TEM images (Fig

1B). This process is time-consuming and restricts the advancement of CKD research and clinical diagnosis.[16-19]. Manual measurement also introduces significant inter-operator variation and bias[20-22]. Due to the inherent characteristics of TEM data, such as grayscale-only imaging, low contrast, and fuzzy boundaries between ultrastructural features, which make conventional artificial intelligence (AI) models struggle to delineate fine details accurately. Currently, no available tools achieve results as accurate as manual labeling (Fig 1C).

During the past decade, semi-automated methods have been piloted for GBM and PFP measurements[20, 23, 24] (Table 1). However, these methods have suffered from low accuracy and still require expert user intervention for each image. In parallel, several fully automated deep learning-based approaches have been proposed for TEM image analysis, targeting the segmentation of GBM[25, 26], PFP[21], or both[27-29], as well as other related ultrastructural components[25, 30, 31]. Most of these methods adopt U-Net-based semantic segmentation frameworks. While such models demonstrate improved efficiency over traditional pipelines, they are inherently limited in their ability to distinguish individual PFP instances, a critical requirement for accurate morphometric analysis. Moreover, many of these studies are trained and evaluated on single-center or disease-specific datasets, limiting generalizability across species, disease types, and imaging protocols[25, 26, 28]. In addition, suboptimal segmentation accuracy further constrains their applicability for precise downstream ultrastructural measurements. More recently, the automated model FPW-DL was developed to measure PFP width[31], with publicly released code and a valuable human disease TEM dataset. The method demonstrates stable performance when applied to human samples during testing. However, its reliance on identifying slit diaphragms as proxies for PFP width measurement imposes intrinsic methodological limitations, leaving substantial room for improvement in measurement accuracy (Fig. 1D). The Segment Anything Model (SAM)[32], the

recently published model, has become one of the most successful deep learning models for general image segmentation. A variant fine-tuned for medical images, called MedSAM[33], has also been released. While both models outperform most others in many biomedical image segmentation tasks, their performance in TEM image analysis is unreliable, even when assisted with manual prompting (Yellow boxes in Fig 1E, 1F). The need for precise prompts underscores the challenge of applying SAM to TEM images. Building upon these developments, a recent study proposed TEM-AID[34], a multi-stage TEM analysis framework integrating detection, prompt-based SAM, measurement, and classification. However, its reliance on sequential processing and human-in-the-loop refinement increases system complexity and limits scalability, and its quantitative analysis primarily focuses on GBM, with limited support for simultaneous PFP quantification. 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. Together, these challenges highlight the urgent need for a specialized digital pathology platform capable of efficient and precise TEM image analysis.

In response to these challenges, we introduce TEAMKidney, a novel deep-learning framework designed for precise segmentation and quantification of TEM ultrastructure. Through extensive generation, collection, and curation of approximately 13,000 TEM images representing various kidney diseases across human, mouse, and rat samples, TEAMKidney is the first deep learning model capable of robustly analyzing TEM images across different magnifications, disease states, and species. To address the challenge of limited high-quality labeled data, in addition to having experienced pathologists meticulously annotating the ultrastructure of glomeruli in 315 TEM images, we incorporated a self-training strategy to leverage large amounts of unlabeled data.

Furthermore, our specifically designed panoptic segmentation model can accurately distinguish individual ultrastructural components even in the absence of clear boundaries, which represents a major challenge for existing methods.

To validate the framework, we conducted extensive quantitative evaluations. In addition to comparisons with expert manual annotations, TEAMKidney significantly outperforms existing methods. Robust performance across data from different sources and species demonstrates the strong generalizability of TEAMKidney. Successful application to both animal models and patient datasets with diverse disease states further indicates its ability to capture subtle ultrastructural changes in real-world TEM data. Overall, this approach has the potential to enhance diagnostic accuracy, facilitate large-scale research on kidney disease, and support clinical decision-making in nephrology. Moreover, it can readily be extended to other clinical domains, making it a promising step toward practical deep learning-based analysis for general TEM imaging.

Results

The Overview of the TEAMKidney Framework

TEAMKidney is a deep learning-based automated framework for TEM image analysis, designed to precisely quantify ultrastructure, such as GBM and PFP widths (Fig. 1G). The framework comprises three key components: (1) integrating HRNet with self-training for initial semantic segmentation (Fig. 1H), (2) developing a novel model for panoptic segmentation of glomerular ultrastructure (Fig. 1H), and (3) performing post-segmentation analysis to robustly quantify detailed ultrastructure, such as GBM width and PFP width (Fig. 1I-K).

To train and validate our framework across diverse pathological conditions and species, our study conducted an analysis of both animal and human models under normal and diseased states. For the mouse, the *Integrin linked kinase* podocyte-specific conditional knockout (ILK cKO)[35, 36] was used for model training, whereas the type IV collagen subunit alpha 3 knockout (Col4a3 KO) model was reserved as an independent evaluation dataset to assess model generalizability. For rats, the passive Heymann nephritis model (PHN)[37] was used, which was originally developed to characterize Membranous nephropathy (MN). For human samples, we focused on the comparison among healthy donors, diabetic kidney disease (DKD)[38, 39] and Fabry disease[40, 41]. For each species and disease model, a corresponding normal control dataset was included to enable both training and validation.

This framework was implemented to automatically segment, label, and measure ultrastructure in TEM images, thereby eliminating the need for manual tracing and reducing observer variability. This design allows the framework to address diverse measurement tasks, including interspecies comparisons, disease-associated ultrastructural changes, and quantitative evaluation of experimental and clinical datasets.

147

148 **Two-Stage Segmentation Model for Improved Fine Ultrastructure Recognition**

149 Due to the high magnification and resolution of grayscale TEM images, ultrastructural features
150 often appear irregular and have unclear boundaries. In addition, TEM images are acquired across
151 multiple magnification levels and spatial resolutions, resulting in substantial scale variation of
152 ultrastructural components within and across datasets. These challenges prevent the direct
153 adaptation of existing segmentation software for TEM image analysis (Fig. 1D-F)[31-33]. To address
154 this, we design a novel deep learning model, Glom2Mask, that fuses both semantic and panoptic
155 segmentation approaches, enabling precise segmentation of ultrastructural features across species
156 and imaging scales.

157 In the semantic segmentation stage (Fig. 2A), we evaluated three backbone networks for
158 comparison: DeepLabV3[42-44], PSPNet[45], and HRNet[46]. All three models are effective in capturing
159 global context and fusing multi-scale features. Overall, the models achieved comparable
160 performance (Fig. 2B). Among them, HRNet attained the highest Dice scores across most
161 categories and demonstrated markedly greater training efficiency. Although the average Dice
162 scores of the three methods were similar, HRNet exhibited the most robust and consistent
163 performance. Therefore, we selected HRNet as the backbone for our panoptic segmentation model
164 in the second stage (Fig. 2C). In contrast, DeepLabV3 often failed on low-quality images (Fig. 2C,
165 Supplementary Fig. 1), resulting in unreliable segmentation, while PSPNet tended to misclassify
166 unrelated ultrastructure (Fig. 2C, Supplementary Fig. 1), particularly mistaking epithelial cells on
167 glomerular capillaries for foot processes.

168 Although our semantic segmentation model accurately recognized major ultrastructure,
169 direct post-segmentation analysis to identify individual instances, such as PFP and slit diaphragms,

remained highly challenging and often inaccurate, similar to most existing models[28, 29]. To overcome this limitation, we introduced a panoptic segmentation stage (Fig. 1H) described in our proposed Glom2Mask model. This architecture utilizes the extracted features from HRNet trained in the first step. In addition, using annotations generated during the semantic segmentation stage, along with manually refined corrections, we constructed an expanded and higher-quality labeled training dataset to improve panoptic segmentation accuracy.

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 (PFPs) and stuff (GBM) classes. The results show that Glom2Mask outperformed the baseline Mask2Former (Fig. 2D), achieving higher PQ, SQ, and RQ scores for both PFP and GBM while requiring fewer training iterations (66,000 vs. 360,000). The quantitative results corresponding to Fig. 2D are summarized in Table S1. 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 PFP instances (various shades of red), further supporting its robustness and generalizability across species.

To evaluate the robustness of our method to variations in image resolution, which may mimic the differences in images generated by different institutes or instruments, we conducted a controlled downsampling experiment using TEM images at multiple scales. As shown in Supplementary Fig. 2A-B, representative segmentation results and corresponding input images demonstrate that meaningful GBM and PFP boundaries can already be identified at moderate downsampling levels, with progressively improved delineation as resolution increases.

Quantitative analysis further confirms this observation (Supplementary Fig. 2C), showing that Dice scores for both GBM and PFP increase rapidly with higher resolution and reach a stable plateau beyond a downsampling ratio of approximately 0.1. Notably, segmentation performance remains consistently high across a broad range of resolution scales, indicating that the proposed method is robust to substantial resolution degradation and generalizes well across varying imaging conditions.

Concordance of TEAMKidney Predictions with Ground Truth Across Species

To evaluate how segmentation accuracy affects ultrastructure measurement, we generated ground truth labels for a validation dataset comprising 48 images from 8 mouse glomeruli, 107 images from 12 rat glomeruli, and 91 images from 5 human glomeruli. All validation images were excluded from model training. GBM and PFP annotations were manually labeled by three human experts to generate precise ultrastructure masks. Then, the post-segmentation quantification algorithm was applied to both the TEAMKidney predictions and the expert-generated ground truth masks in this validation dataset. Quantified GBM and PFP width values obtained from both ground truth labels and TEAMKidney predictions were compared across mouse, rat, and human samples at the individual image level. No statistically significant differences were observed between the ground truth and TEAMKidney measurements for either GBM or PFP width across all three species (Fig. 2F), demonstrating strong concordance and reliability of the proposed framework. These results indicate that our method provides highly reliable GBM and PFP width measurements across diverse datasets.

Distinguishing Ultrastructural Changes in Multiple Animal Models of Kidney Disease

Animals, especially rodents, are widely used to model a variety of kidney diseases. Ultrastructural changes reflect disease progression and serve as indicators of treatment efficacy during drug development[47]. To test the robustness of our framework for 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, B). In the mouse models, we used Col4a3 KO, an animal model of Alport syndrome[48] and ILK cKO mice, an animal model of focal segmental glomerulosclerosis (FSGS), to represent two distinct mechanisms of kidney glomerular injury[35, 36, 48]. Col4a3 KO mice exhibit severe GBM disorganization due to the lack of type IV α3 collagen, which is a crucial component of the mature GBM, whereas ILK cKO mice develop podocyte detachment and cytoskeletal abnormalities with the absence of ILK protein, which is required for podocyte adhesion to the GBM (Fig. 3A, C). Consistent with these pathological features, our automated analysis revealed significantly increased GBM widths in Col4a3 KO mice compared with wild-type (WT) controls. Similarly, PFP widths were markedly widened for the ILK cKO mouse model. These quantitative differences align with the expected ultrastructural alterations associated with abnormal GBM of the Alport mouse model and podocyte injury in the ILK cKO FSGS mouse model (Fig. 3C, D).

We next examined the rat PHN model of immune-mediated podocyte injury[37]. In this model, rats were injected with anti-Fx1a antibody to induce the formation of immune complexes that deposit beneath podocytes and trigger complement-mediated podocyte injury, leading to progressive PFP effacement and GBM thickening (Fig. 3B, E). Automated measurements faithfully captured these ultrastructural changes: compared with WT animals, PHN rats showed a striking increase in GBM width (Fig. 3F). These differences are consistent with classical TEM-based characterizations of PHN, an animal model of human membranous nephropathy, confirming

239 that our approach is sensitive to detecting pathological remodeling of the glomerular filtration
240 barrier.
241
242 **Identifying Ultrastructural Remodeling in Human Kidney Diseases**
243 TEM-based ultrastructural quantification is commonly used as diagnostic evidence in clinical renal
244 pathology practice. To investigate the translational utility of our framework in human kidney
245 disease states, we analyzed TEM images from patients with DKD and Fabry disease (Fig. 4A).
246 DKD is characterized by hyperglycemia-induced hyperfiltration, GBM thickening, diffuse
247 mesangial matrix expansion, leading to podocyte stress and progressive glomerular injury, PFP
248 effacement, glomerular hypertrophy, and eventually podocyte loss and severe nodular
249 glomerulosclerosis[49-51]. Fabry disease results from *GLA* mutations that cause accumulation of
250 globotriaosylceramide (Gb3) in podocytes and subsequent structural damage, PFP effacement,
251 podocyte detachment, and global or segmental glomerulosclerosis[52, 53] (Fig. 4A). Representative
252 TEM images predicted by our TEAMKidney model demonstrate marked distinct pathological
253 features in both disease cohorts compared with normal controls (Fig. 4B).
254 Quantitative measurements revealed disease-specific ultrastructural alterations (Fig. 4C).
255 In DKD patients, GBM thickness was significantly increased. Notably, the mean PFP width in
256 DKD showed only a slight increase compared with normal samples. The relatively preserved PFP
257 width in our DKD cohort warrants further investigation, as it may reflect the disease stage of the
258 included patients or capture an early phase where GBM pathology precedes overt podocyte
259 architectural disruption. Longitudinal studies correlating PFP measurements with clinical disease
260 staging would help clarify the temporal evolution of these ultrastructural changes[39, 54]. In contrast,

261 patients with Fabry disease exhibited significant increases in both GBM thickness and PFP width,
262 reflecting the more severe podocyte injury characteristic of lysosomal storage disorders[55, 56].
263 Furthermore, sex-stratified analysis of Fabry patients revealed that both male and female
264 patients exhibited substantial GBM and PFP widening (Fig. 4D, E), with more pronounced changes
265 in Fabry males. These findings are consistent with the known pathology of Fabry disease, in which
266 hemizygous males typically exhibit more severe phenotypes due to the X-linked inheritance
267 pattern[57, 58].
268 These results highlight the ability of our automated framework to detect clinically relevant
269 and disease-specific ultrastructural alterations in human glomerulopathies. By accurately capturing
270 the differential patterns of GBM thickening and PFP effacement in DKD and Fabry disease, the
271 method demonstrates strong potential for application in both diagnostic pathology and clinical
272 research, where reproducible and scalable measurements are critical for disease characterization,
273 staging, patient stratification and management.
274 To assess robustness across imaging scales, we evaluated our method on TEM images
275 acquired at multiple magnifications in the Fabry dataset. As shown in Fig. 4F, the predicted GBM
276 and PFP mean widths remain consistent across different magnifications, despite substantial
277 differences in image resolution. These results indicate that our method is stable and effective across
278 imaging scales and disease conditions. Similar results for the DKD and normal datasets are shown
279 in Supplementary Fig. 3.
280
281 **Comparison of PFP Quantification with a Benchmark Model**
282 To better evaluate the performance, we benchmarked TEAMKidney against existing automated
283 TEM quantification approaches using human, mouse, and rat kidney samples. As FPW-DL[31] is the

284 only method with publicly available code or model, and is designed exclusively to measure PFP
285 width, the comparison was limited to PFP width quantification.

286 Considering FPW-DL was trained only on human samples, our comparison focused on
287 human diseases. Even though all images were obtained from the FPW-DL study, TEAMKidney
288 consistently outperformed FPW-DL across both normal and disease cohorts (Fig. 5A, B,
289 Supplementary Fig. 4). Scatter plots show strong agreements between TEAMKidney and the
290 ground truth for both PFP width and counts (Fig. 5A, B). In contrast, FPW-DL tended to
291 overestimate PFP width and underestimate PFP counts, with this bias being more pronounced in
292 DKD than in the normal cohort. Mean error of PFP width was reduced from 24.80 pixels (FPW-
293 DL) to 4.46 pixels using our framework (Supplementary Fig. 4).

294 Further investigation revealed that PFP miscounting in FPW-DL arose from the use of slit
295 diaphragm detection (yellow dots in Fig. 5C) rather than direct PFP detection. Under disease
296 conditions or variable imaging quality, PFP effacement and overlap can obscure slit diaphragms,
297 leading to overestimation of PFP width when slit diaphragms are missed, or preferential detection
298 of relatively normal-appearing PFPs while pathologically altered regions are overlooked. Accurate
299 ultrastructural quantification, therefore, critically depends on high-quality segmentation, which is
300 the primary design goal of TEAMKidney. Representative TEM images show that our approach
301 yields more continuous and anatomically faithful segmentation of podocyte FPs and the GBM
302 compared with alternative methods (Fig. 5C), resulting in robust performance across both human
303 disease states and animal models.

304 The performance gap between the two methods became more pronounced in animal sample
305 (Fig. 5A, B), where FPW-DL, which was trained exclusively on human kidney data, showed
306 substantially larger deviations from ground truth (Fig. 5A). This limitation stems from two primary

307 factors: first, morphological differences between species, particularly the thinner GBM in rodents
308 compared to humans, poses additional challenges for accurate structure identification; second,
309 differences in imaging resolution and scale between our rodent datasets and FPW-DL's training
310 data further compromised its generalizability. In contrast, TEAMKidney maintained robust
311 performance across species, demonstrating superior adaptability to diverse morphological features
312 and imaging conditions.
313
314 **Agreement of TEAMKidney and Manual Measurements Across Species**
315 Manually tracing and labeling each ultrastructural component is not feasible in practice. Instead,
316 pathologists and biomedical scientists typically quantify ultrastructure by measuring multiple
317 regions within an image to obtain representative values (see Methods). To assess discrepancies,
318 we compared TEAMKidney measurements with manual measurements across normal, DKD, and
319 Fabry disease (male and female) from human cohorts (Fig. 6) and animal models (Supplementary
320 Fig. 5). For GBM widths, TEAMKidney quantification showed no significant differences from
321 manual measurements in all groups (Fig. 6A, Supplementary Fig. 5), confirming that the
322 framework preserves accuracy for this ultrastructural feature. Similarly, no significant differences
323 were detected in PFP counts between TEAMKidney and manual counting across all groups (Fig.
324 6B, Supplementary Fig. 5), indicating that the framework maintains biological fidelity in
325 enumerating PFP.
326 We observed a significant discrepancy in PFP width measurements between the two
327 methods, with mean values from TEAMKidney consistently lower than those obtained by manual
328 measurement. This difference likely arises from two factors. First, due to the irregular shape of
329 PFPs, in the manual approach, PFP width is not measured directly but instead estimated by dividing

330 the GBM length by the PFP count. GBM length is often overestimated, particularly in disease
331 models with pronounced morphological alterations. Second, manual measurements include the slit
332 diaphragm as part of the PFP, which inflates the estimated PFP width. In contrast, the
333 TEAMKidney measurement strategy segments individual PFPs and slit diaphragm separately;
334 therefore, the slit diaphragm is excluded from the PFP measurement, resulting in higher accuracy
335 compared with the manual method (Fig. 6C, Supplementary Fig. 5).

336 Together, these results demonstrate strong concordance between TEAMKidney and
337 manual measurements, establishing the framework as a reliable substitute for labor-intensive
338 tracing. By accurately replicating expert-level quantification across both healthy and diseased
339 human samples, the method enhances reproducibility and offers a scalable solution for
340 ultrastructural kidney assessment.

341

342 **Comparison of Clinical Protocol and TEAMKidney for GBM and PFP Quantification**

343 Although manual measurement is widely used in basic and clinical research, area selection is
344 subject to individual judgment, leading to substantial inter-operator variability. In clinical settings,
345 more stringent area-selection protocols are applied to improve reproducibility (see Methods). To
346 contextualize the advantages of our automated approach, we compared it with the conventional
347 diagnostic manual workflow commonly used for ultrastructural measurements of GBM and PFP
348 widths (Fig. 7). Pathologists rely on ImageJ (or FIJI) based protocols, which require calibration
349 with scale bars, grid overlays, and labor-intensive tracing of GBM thickness at multiple grid
350 intersections, followed by foot process identification and marking of lateral edges (Fig. 7A). This
351 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 PFP 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

In recent years, a global shortage of pathologists has been widely reported[59, 60], affecting even large medical centers and resource-rich regions. The preparation and interpretation of tissue samples for TEM imaging require additional specialized training, particularly for renal pathologists, further exacerbating this workforce constraint. To address the urgent needs in nephrology research and clinical pathology, our team of experienced renal pathologists, nephrologists, biologists, and computational scientists designed and implemented TEAMKidney. Precise measurement of the GBM and PFP provides a much-needed solution to address labor-intensive and often subjective assessments of glomerular diseases. GBM and PFP are not only essential ultrastructure for diagnosing and monitoring chronic kidney diseases, but they also represent two distinct classes of TEM features. GBM is characterized by its continuous structure with a relatively regular shape and ambiguous boundaries throughout the image. In contrast, PFP consists of small, isolated structures with clear boundaries and irregular shapes. Accurate segmentation and quantification of these structures demonstrate the capability of the framework to extend beyond nephrology and renal pathology to general TEM image analysis across other fields.

High-quality labeled data is critical for training deep learning models, but manual annotation of TEM images is extremely time-consuming and requires specialized pathological expertise. To overcome this challenge, we adopted semi-automated pre-labeling strategies to accelerate the annotation workflow and implemented a multi-stage quality control process, ensuring that each label was independently reviewed and refined. These steps significantly reduced inter-operator variability and improved the overall reliability of the training data. Combined with a self-training strategy, our framework achieves high accuracy with only a few hundred finely annotated images, making it feasible to adapt to other tasks efficiently.

Previous TEM analysis methods have primarily focused on segmenting a single ultrastructural feature, often failing to generalize to other ultrastructures. Most studies have relied on U-Net based or semi-automated approaches (Table 2), which reduce the need for large amounts of training data but are well known for their limited generalizability to diverse morphologies. Because these methods typically use data from a single source and a single species, their practical applicability in animal disease modeling and clinical settings is further constrained. In addition to collecting and generating diverse datasets across multiple sources, species, and disease states, our specifically designed panoptic segmentation model offers two key advantages over directly applying generic computer vision models to this task. First, semantic segmentation provides spatially consistent, region-level guidance, reducing noisy or fragmented instance predictions during pseudo-labeling. Second, the trained HRNet model serves as the backbone for the panoptic segmentation network, enabling direct transfer and reuse of feature representations learned during semantic segmentation. This strategy not only accelerates model convergence but also improves feature consistency between semantic and panoptic tasks, thereby enhancing boundary precision and instance recognition. Overall, this hierarchical training strategy achieves greater 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, yielding consistent segmentation quality and reliable quantitative measurements across a wide range of magnification differences.

Our proposed framework demonstrates high accuracy and reproducibility in segmenting and quantifying GBM and PFP, and generalizes effectively across species and disease states while accommodating substantial morphological and magnification variability. Nonetheless, there remains room for further improvement. Because TEM sections are two-dimensional representations of inherently three-dimensional structures, the spatial orientation of PFP can influence quantitative measurements, although this effect is unlikely to substantially alter overall conclusions when analyzed across whole images. To further enhance measurement accuracy, we are developing new methods that explicitly incorporate PFP orientation analysis. In parallel, we are extending the framework to include additional ultrastructural features relevant to other kidney diseases, such as immune deposits, slit diaphragms, and endothelial cells. As the first framework to achieve both segmentation and quantitative measurement with expert-level precision, TEAMKidney enables accurate characterization of glomerular morphological changes across diverse clinical conditions, ranging from normal kidneys and common diseases such as DKD to

rare disorders such as Fabry disease. This work establishes a solid foundation for expanding the framework to quantify the full spectrum of glomerular diseases and to support future applications, including disease detection, classification, and automated report generation.

While open-source software is standard practice in computer science, it remains uncommon in the field of TEM image analysis. With the exception of FPW-DL, all other existing studies have not released their code or trained models, making independent evaluation impossible and significantly limiting the broader adoption and impact of these methods. We greatly respect the FPW-DL authors for openly sharing their code, models, and data, and for their prompt and helpful responses to our inquiries. We strongly believe that making our own code and models publicly available will foster community-wide experimentation, accelerate methodological advances, and ultimately benefit patient care. This openness also reflects our commitment to transparency and confidence in the robustness of our approach.

449 **References**

450

451 1. Koster, A.J. & Klumperman, J. Electron microscopy in cell biology: integrating structure
452 and function. *Nat Rev Mol Cell Biol* **Suppl**, SS6-10 (2003).
453 2. Curry, A., Appleton, H. & Dowsett, B. Application of transmission electron microscopy to
454 the clinical study of viral and bacterial infections: present and future. *Micron* **37**, 91-106
455 (2006).
456 3. Roingeard, P., Raynal, P.I., Eymieux, S. & Blanchard, E. Virus detection by transmission
457 electron microscopy: Still useful for diagnosis and a plus for biosafety. *Rev Med Virol* **29**,
458 e2019 (2019).
459 4. Graham, L. & Orenstein, J.M. Processing tissue and cells for transmission electron
460 microscopy in diagnostic pathology and research. *Nat Protoc* **2**, 2439-2450 (2007).
461 5. L'Imperio, V. et al. Digital pathology for the routine diagnosis of renal diseases: a standard
462 model. *J Nephrol* **34**, 681-688 (2021).
463 6. Murray, R. et al. Kidney-Related Research in the United States: A Position Statement From
464 the National Kidney Foundation and the American Society of Nephrology. *Am J Kidney*
465 *Dis* **78**, 161-167 (2021).
466 7. Assady, S., Benzing, T., Kretzler, M. & Skorecki, K.L. Glomerular podocytes in kidney
467 health and disease. *Lancet* **393**, 856-858 (2019).
468 8. Faucon, A.L. CKD Beyond the Clinic: A Holistic Projection of its Global Burden. *Kidney*
469 *Int Rep* **10**, 2524-2526 (2025).

470 9. Chadban, S. et al. Projecting the economic burden of chronic kidney disease at the patient
471 level (Inside CKD): a microsimulation modelling study. *EClinicalMedicine* **72**, 102615
472 (2024).
473 10. Ramachandran, R. et al. Challenges in Diagnosis and Management of Glomerular Disease
474 in Resource-Limited Settings. *Kidney Int Rep* **7**, 2141-2149 (2022).
475 11. George, C., Mogueo, A., Okpechi, I., Echouffo-Tcheugui, J.B. & Kengne, A.P. Chronic
476 kidney disease in low-income to middle-income countries: the case for increased screening.
477 *BMJ Glob Health* **2**, e000256 (2017).
478 12. Rajapurkar, M.M. et al. What do we know about chronic kidney disease in India: first report
479 of the Indian CKD registry. *BMC Nephrol* **13**, 10 (2012).
480 13. Sharma, M., Doley, P. & Das, H.J. Etiological profile of chronic kidney disease: A single-
481 center retrospective hospital-based study. *Saudi J Kidney Dis Transpl* **29**, 409-413 (2018).
482 14. Yamashita, M., Lin, M.Y., Hou, J., Ren, K.Y.M. & Haas, M. The Continuing Need for
483 Electron Microscopy in Examination of Medical Renal Biopsies: Examples in Practice.
484 *Glomerular Dis* **1**, 145-159 (2021).
485 15. May, R.M. et al. A multi-center retrospective cohort study defines the spectrum of kidney
486 pathology in Coronavirus 2019 Disease (COVID-19). *Kidney Int* **100**, 1303-1315 (2021).
487 16. Pisarek-Horowitz, A. et al. Loss of Roundabout Guidance Receptor 2 (Robo2) in Podocytes
488 Protects Adult Mice from Glomerular Injury by Maintaining Podocyte Foot Process
489 Structure. *Am J Pathol* **190**, 799-816 (2020).
490 17. Veron, D. et al. Overexpression of VEGF-A in podocytes of adult mice causes glomerular
491 disease. *Kidney Int* **77**, 989-999 (2010).

492 18. Sethi, S. et al. Mayo Clinic/Renal Pathology Society Consensus Report on Pathologic
493 Classification, Diagnosis, and Reporting of GN. *J Am Soc Nephrol* **27**, 1278-1287 (2016).
494 19. Levin, A. et al. Global kidney health 2017 and beyond: a roadmap for closing gaps in care,
495 research, and policy. *Lancet* **390**, 1888-1917 (2017).
496 20. Rangayyan, R.M., Kamenetsky, I. & Benediktsson, H. Segmentation and analysis of the
497 glomerular basement membrane in renal biopsy samples using active contours: a pilot
498 study. *J Digit Imaging* **23**, 323-331 (2010).
499 21. Unnersjo-Jess, D. et al. Deep learning-based segmentation and quantification of podocyte
500 foot process morphology suggests differential patterns of foot process effacement across
501 kidney pathologies. *Kidney Int* **103**, 1120-1130 (2023).
502 22. Geraghty, M. et al. Sources of Variability in Podocyte Foot Process Width Measurements
503 and Approaches to Mitigation: PO2004. *Journal of the American Society of Nephrology*
504 **31**, 617 (2020).
505 23. Kamenetsky, I., Rangayyan, R.M. & Benediktsson, H. Analysis of the glomerular basement
506 membrane in images of renal biopsies using the split-and-merge method: a pilot study. *J*
507 *Digit Imaging* **23**, 463-474 (2010).
508 24. Wu, H.S., Dikman, S. & Gil, J. A semi-automatic algorithm for measurement of basement
509 membrane thickness in kidneys in electron microscopy images. *Comput Methods*
510 *Programs Biomed* **97**, 223-231 (2010).
511 25. Wang, Y. et al. Segmentation and thickness calculation of glomerular basement membrane
512 using RADS-Net in glomerular microscopic images. *Biomedical Signal Processing and*
513 *Control* **88**, 105557 (2024).

26. Laudon, A. et al. Digital pathology assessment of kidney glomerular filtration barrier ultrastructure in an animal model of podocytopathy. *Biol Methods Protoc* **10**, bpaf024 (2025).

27. Lin, G. et al. GCLR: A self-supervised representation learning pretext task for glomerular filtration barrier segmentation in TEM images. *Artif Intell Med* **146**, 102720 (2023).

28. 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. *Comput Methods Programs Biomed* **268**, 108842 (2025).

29. Zhang, Z. et al. Glo-UMF: A Unified Multi-model Framework for Automated Morphometry of Glomerular Ultrastructural Characterization. *arXiv preprint arXiv:2508.10351* (2025).

30. 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).

31. 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).

32. Kirillov, A. et al. in Proceedings of the IEEE/CVF international conference on computer vision 4015-4026 (2023).

33. Ma, J. et al. Segment anything in medical images. *Nat Commun* **15**, 654 (2024).

34. Ma, P. et al. AI-Based System for Analysis of Electron Microscope Images in Glomerular Disease. *JAMA Netw Open* **8**, e2534985 (2025).

536 35. Huang, M. et al. Integrin-Linked Kinase Deficiency in Collecting Duct Principal Cell
537 Promotes Necroptosis of Principal Cell and Contributes to Kidney Inflammation and
538 Fibrosis. *J Am Soc Nephrol* **30**, 2073-2090 (2019).
539 36. El-Aouni, C. et al. Podocyte-specific deletion of integrin-linked kinase results in severe
540 glomerular basement membrane alterations and progressive glomerulosclerosis. *J Am Soc*
541 *Nephrol* **17**, 1334-1344 (2006).
542 37. Jefferson, J.A., Pippin, J.W. & Shankland, S.J. Experimental Models of Membranous
543 Nephropathy. *Drug Discov Today Dis Models* **7**, 27-33 (2010).
544 38. Yang, H. et al. Podocyte programmed cell death in diabetic kidney disease: Molecular
545 mechanisms and therapeutic prospects. *Biomed Pharmacother* **177**, 117140 (2024).
546 39. Marshall, C.B. Rethinking glomerular basement membrane thickening in diabetic
547 nephropathy: adaptive or pathogenic? *Am J Physiol Renal Physiol* **311**, F831-F843 (2016).
548 40. Bokhari, S.R.A. & Safdar, A. in StatPearls (Treasure Island (FL); 2026).
549 41. Levstek, T. et al. Identification of genetic variants associated with Fabry nephropathy
550 progression using whole-exome sequencing. *Mol Genet Metab* **145**, 109191 (2025).
551 42. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for
552 semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).
553 43. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A.L. Deeplab: Semantic
554 image segmentation with deep convolutional nets, atrous convolution, and fully connected
555 crfs. *IEEE transactions on pattern analysis and machine intelligence* **40**, 834-848 (2017).
556 44. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. in Proceedings of the
557 European conference on computer vision (ECCV) 801-818 (2018).

558 45. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. in Proceedings of the IEEE conference on
559 computer vision and pattern recognition 2881-2890 (2017).
560 46. Sun, K., Xiao, B., Liu, D. & Wang, J. in Proceedings of the IEEE/CVF conference on
561 computer vision and pattern recognition 5693-5703 (2019).
562 47. Beck, L.H., Jr. et al. PODO: Trial Design: Phase 2 Study of PF-06730512 in Focal
563 Segmental Glomerulosclerosis. *Kidney Int Rep* **6**, 1629-1633 (2021).
564 48. Cosgrove, D. et al. Collagen COL4A3 knockout: a mouse model for autosomal Alport
565 syndrome. *Genes Dev* **10**, 2981-2992 (1996).
566 49. Lin, Y.C., Chang, Y.H., Yang, S.Y., Wu, K.D. & Chu, T.S. Update of pathophysiology and
567 management of diabetic kidney disease. *J Formos Med Assoc* **117**, 662-675 (2018).
568 50. Sugahara, M., Pak, W.L.W., Tanaka, T., Tang, S.C.W. & Nangaku, M. Update on diagnosis,
569 pathophysiology, and management of diabetic kidney disease. *Nephrology (Carlton)* **26**,
570 491-500 (2021).
571 51. Mora-Fernandez, C. et al. Diabetic kidney disease: from physiology to therapeutics. *J*
572 *Physiol* **592**, 3997-4012 (2014).
573 52. Alroy, J., Sabnis, S. & Kopp, J.B. Renal pathology in Fabry disease. *J Am Soc Nephrol* **13**
574 **Suppl 2**, S134-138 (2002).
575 53. Eng, C.M. et al. Fabry disease: guidelines for the evaluation and management of multi-
576 organ system involvement. *Genet Med* **8**, 539-548 (2006).
577 54. Najafian, B. et al. Progressive podocyte injury and globotriaosylceramide (GL-3)
578 accumulation in young patients with Fabry disease. *Kidney Int* **79**, 663-670 (2011).
579 55. Wise, A.F. et al. Fabry Disease Podocytes Reveal Ferroptosis as a Potential Regulator of
580 Cell Pathology. *Kidney Int Rep* **10**, 535-548 (2025).

581 56. Monte Neto, J.T.D. & Kirsztajn, G.M. The role of podocyte injury in the pathogenesis of
582 Fabry disease nephropathy. *J Bras Nefrol* **46**, e20240035 (2024).
583 57. Mehta, A. & Hughes, D.A. in GeneReviews((R)). (eds. M.P. Adam et al.) (Seattle (WA);
584 1993).
585 58. Germain, D.P. Fabry disease. *Orphanet J Rare Dis* **5**, 30 (2010).
586 59. Osman, M.A. et al. Global nephrology workforce: gaps and opportunities toward a
587 sustainable kidney care system. *Kidney Int Suppl (2011)* **8**, 52-63 (2018).
588 60. Collaboration, G.B.D.C.K.D. Global, regional, and national burden of chronic kidney
589 disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017.
590 *Lancet* **395**, 709-733 (2020).
591 61. Salant, D.J. & Cybulsky, A.V. Experimental glomerulonephritis. *Methods Enzymol* **162**,
592 421-461 (1988).
593 62. Russell, B.C., Torralba, A., Murphy, K.P. & Freeman, W.T. LabelMe: A Database and Web-
594 Based Tool for Image Annotation. *International Journal of Computer Vision* **77**, 157-173
595 (2008).
596 63. Lahdenkari, A.T. et al. Podocytes are firmly attached to glomerular basement membrane in
597 kidneys with heavy proteinuria. *J Am Soc Nephrol* **15**, 2611-2618 (2004).
598 64. Shea, S.M. & Morrison, A.B. A stereological study of the glomerular filter in the rat.
599 Morphometry of the slit diaphragm and basement membrane. *J Cell Biol* **67**, 436-443
600 (1975).

601

602

603 **Methods**

604 **Sample Preparation and Image Acquisition**

605 *ILK podocyte-specific knockout (ILK cKO) and Col4a3 knockout (Col4a3 KO) models.* The ILK
606 cKO mice were generated by crossing $ILK^{flox/flox}$ mice with Podocin-Cre^{+} mice[36]. Col4a3 KO mice
607 were purchased from the Jackson Laboratory (RRID:IMSR_JAX:002908). The littermate WT
608 mice were used as healthy controls. Mice were sacrificed at 16 weeks of age, and the kidneys were
609 dissected. The fresh kidney samples were fixed in Karnovsky's fixative and processed according
610 to the published protocol[4]. In brief, small pieces of kidney samples (<1 mm thick) were fixed in
611 buffered 2.5% glutaraldehyde. Kidney specimens were then post-fixed in 1% osmium tetroxide
612 (Electron Microscopy Sciences), dehydrated, embedded in propylene oxide/Epon resin, and
613 sectioned (60 to 90 nm thick) using the ultramicrotome. Collected sections were post-stained with
614 uranyl acetate and lead citrate (Electron Microscopy Sciences), according to the published
615 protocol[4]. Images were obtained using the Jeol JEM-1011 transmission electron microscope (JEOL
616 USA Inc.) equipped with an Erlangshen ES100W digital camera (Gatan, Pleasanton, CA). The
617 operator was blinded as to the genotype and the injury status of the animals at the time of image
618 collection and analyses[16].

619 *Rat PHN injury model.* A modified PHN injury model was induced in Lewis rats (strain code 004;
620 Charles River Laboratories, Wilmington, MA) by one dose of i.v. injection [1.5 to 2 mL/kg sheep
621 anti-rat Fx1A serum (PTX-002S; Probetex, San Antonio, TX)], according to Probetex and
622 published protocol[61]. Rats were euthanized at day 16 after anti-rat Fx1A serum injection, and
623 kidneys were collected and fixed in Karnovsky's fixative for TEM analysis. Sterile 0.9% sodium
624 chloride (normal saline) was used in control rats. A total of six animals in each experimental group

625 were studied. Kidney samples were prepared, and TEM imaging was performed as described
626 above.
627 *Human diseases.* TEM images used in this study were obtained from an open-source dataset
628 described in a recent publication[31]. The dataset includes kidney biopsy samples from patients with
629 Fabry disease, type 2 diabetes, and minimal change disease, as well as healthy controls from living
630 kidney transplant donors. All biopsies were processed at eight EM laboratories following
631 standardized protocols. Ultrathin glutaraldehyde fixed sections were stained with toluidine blue
632 for glomerular identification, and high magnification transmission EM images (original
633 magnification ×30,000) were captured using a systematic uniform random sampling strategy.
634
635 **Label Generation, Manual Measurement, and Image Prepressing**
636 *Ground Truth Label.* All TEM images were divided into training, validation, and testing datasets
637 as summarized in Table 1. The training/validation and testing sets were intentionally sourced from
638 different cohorts, encompassing distinct species and disease types, to ensure that the testing dataset
639 could robustly evaluate the generalization capability of our algorithm. For the training and
640 validation sets, regions corresponding to the GBM and PFP were manually annotated using
641 LabelMe software[62]. The annotations were performed by trained researchers using high-resolution
642 TEM images and subsequently reviewed by renal pathologists to ensure consistency and structural
643 accuracy. This manual labeling process served as the foundation for model training and validation,
644 while the independently collected testing data were reserved for assessing model performance
645 across unseen imaging and disease conditions.
646 *Manual measurement for GBM and FP widths.* Manual podocyte foot process width was traced
647 and measured on ×12,000 high magnification TEM images using ImageJ software version 1.47v,

648 according to published methods[16]. Briefly, the total length of the GBM spanning the area of the
649 slit diaphragms was measured in each image using ImageJ, according to published methods[17, 63,
650 64]. The results were then pasted on an Excel sheet and adjusted for the scale bar. After that, all slit
651 diaphragms covering the GBM were counted, and the PFP width was calculated using the total
652 length of GBM divided by the total number of slit diaphragms in each image. The GBM width was
653 measured as the distance from the podocyte membrane to the endothelial membrane. Each image
654 was randomly measured in 20 areas around the GBM, and the harmonic mean was calculated for
655 each image.

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

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

Ultrastructure Quantification Based on Clinical Protocols. GBM thickness and PFP 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 due to sectioning angle, folding, or detachment. The Orthogonal Intercept Method (OIM) was applied to calculate mean GBM thickness. For PFP width measurement, the lateral edges of individual PFPs were manually identified, and width was measured as the linear distance between adjacent borders. Only clearly defined and intact PFPs were included in the analysis.

Glom2Mask: 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 PFP in TEM images. Given the distinct structural characteristics of these components, we categorize GBM as a "stuff" class for semantic segmentation and PFP as a "things" class for instance segmentation. This formulation ensures precise differentiation between the continuous structure of the GBM and the discrete PFP instances within the glomerular filtration 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 preserve spatial details throughout the network, which is essential for capturing fine structural differences in TEM images. Unlike traditional backbones that downsample feature maps at early stages, HRNet

maintains high-resolution representations, allowing for better delineation of GBM boundaries and accurate identification of individual PFP 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 a PFP 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 PFP 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 of a modified Mask2Former model for panoptic segmentation, thereby improving both training efficiency and segmentation accuracy.

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

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 capabilities.

The model performance was evaluated using the mean Dice score. For 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^{-5} 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 to step 1. The model was optimized using the AdamW optimizer with a learning rate of 1×10^{-5} 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 PFP. 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, Intel Xeon Gold 6338 CPU (40 cores). Deterministic behavior was enforced using a fixed seed (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 PFP segmentation tiles predicted by the deep learning model to measure the mean GBM width and PFP width. We will describe the 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 the following equation

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

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 for calculating the length of the GBM is as follows: First, we obtain the skeleton of the GBM using the skeletonize method with the 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 (Fig. 1J). 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.

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

Next, we use morphological operations in OpenCV to dilate the GBM, enabling us to obtain its contour. 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 PFP by multiplying the PFP instance matrix by the matrix of the dilated GBM contours. This operation yields the width by computing the pixel length of PFPs. Similarly, the specimen PFP 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.

Author contributions

C.Z., W.L., A.Z., and X.F. conceived the study. A.Z. developed the deep learning model. W.L., X.F., S.R.V., S.P.B., H.Y., 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 and revised the final manuscript. All authors reviewed and approved the final manuscript.

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.

Disclosures

SRV, SPB, and HY are current and former Pfizer employees who helped generate the PHN rat kidney TEM images. WL and JMP received a research grant from Pfizer Centers for Therapeutic Innovation. All other authors have nothing to disclose.

Acknowledgments

We thank Casey Ritenour, Cheryl Tyszkiewicz, Seo-Kyoung Hwang, Chang-Ning Liu, and Lindsay Tomlinson of Pfizer Global Pathology and Comparative Medicine and Drug Safety R&D for help with rat PHN models. We thank Richa Sharma, Sudhir Kumar, and Boston University's Animal Science Center for their help in generating Col4a3 and ILK knockout mice. This work is supported by research grants from NIH/NIDDK grants R21DK143406 (CZ, WL), R01DK133940 (WL), R01DK078226 (WL), and a DOD grant E01 HT9425-23-1-1058 (WL). This work is also supported by BIT-ARC and B4D-ARC grants from Boston University Evans Center (WL, CZ, JMH), a grant from Pfizer Centers for Therapeutic Innovation (WL), a BU CTSI Integrated Pilot Grant funded by the Department of Medicine (CZ, WL), and the Boston University Undergraduate Research Opportunities Program (JJ). This publication is also supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health, through BU-CTSI Grant Number 1UL1TR001430. Its contents are solely the authors' responsibility and do not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript

843 **Table 1. Data Collection and Selection Criteria.**

Group	Total	Manually labeled	Manually measured	Training & Validation	Evaluation
Mouse models					
WT	107	48	20	48	61
ILK cKO	241	67	20	67	174
Col4A3	93	0	20	0	93
Rat models					
WT	69	0	20	0	69
PHN	584	540	20	540	44
Human models					
Normal	1361	71	20	0	1361
Fabry males	4301	76	20	468	3833
Fabry females	2487	40	20	0	2487
DKD	2463	33	20	0	2463

844

845

846 **Table 2. Comparison Between Proposed Model and Other Published Methods.**

Method	Measurement of FP	Measurement of GBM	Label Dependency	Segmentation Target	DL model	Open Source	Year
Wu et al.[22]	No	Yes	N/A	N/A	N/A	No	2010
Yang et al.[28]	No	No	Supervised	GBMDD	U-Net	No	2022
Smerkous et al.[29]	Yes	No	Supervised	PGBMI, SD	U-Net	Yes	2023
Unnersjo-Jess et al.[19]	Yes	No	Supervised	PFP, SD	U-Net	No	2023
Lin et al.[25]	No	No	Self-supervised	GBM, PFP	GCLR	No	2023
Wang et al.[23]	No	Yes	Supervised	GBM	RADS-Net	No	2024
Laudon et al.[24]	Yes	Yes	Supervised	GBM	U-Net	Yes	2025
Yamashita et al.[26]	Only Effacement	Yes	Supervised	GBM, PFP	DeepLab V3	No	2025
Zhang et al.[27]	No	Yes	Supervised	GBM, PFP	U-Net	No	2025
Ma et al.[32]	No	Yes	Supervised	GBM	YOLO+SAM	No	2025
Our Method	**Yes**	**Yes**	**Supervised**	**GBM, PFP**	**Glom2Mask**	**Yes**	**2025**

847

848

849

850 **Figure legends**

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

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

871 **Fig. 3. Application of TEAMKidney to three animal models of kidney disease.** (A) Schematic
872 illustrating two mouse models of glomerular disease: knocking out Col4a3 or specific podocyte

ILK. (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 PFP 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 PFP 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 were analyzed using the TEAMKidney model. (C) Violin plots showing the average GBM and PFP widths per glomerulus across normal, Fabry disease, and DKD samples. Each point represents the mean width of a single glomerulus. (D) Representative TEAMKidney predicted images from male and female Fabry disease patients. (E) Violin plots comparing GBM and PFP widths between male and female Fabry disease groups. (F) Violin plots illustrating the distributions of GBM and PFP widths across different magnifications in human kidney TEM images.

Fig. 5. Comparison of PFP 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 PFP width and (B) PFP 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) GBM mean width. (B) PFP mean width. (C)

PFP count. From left to right: Normal kidney samples, DKD samples, Fabry disease samples from male patients. (D) Fabry disease samples from female patients.

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

Table 1. Data Collection and Selection Criteria.

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

Fig. 1

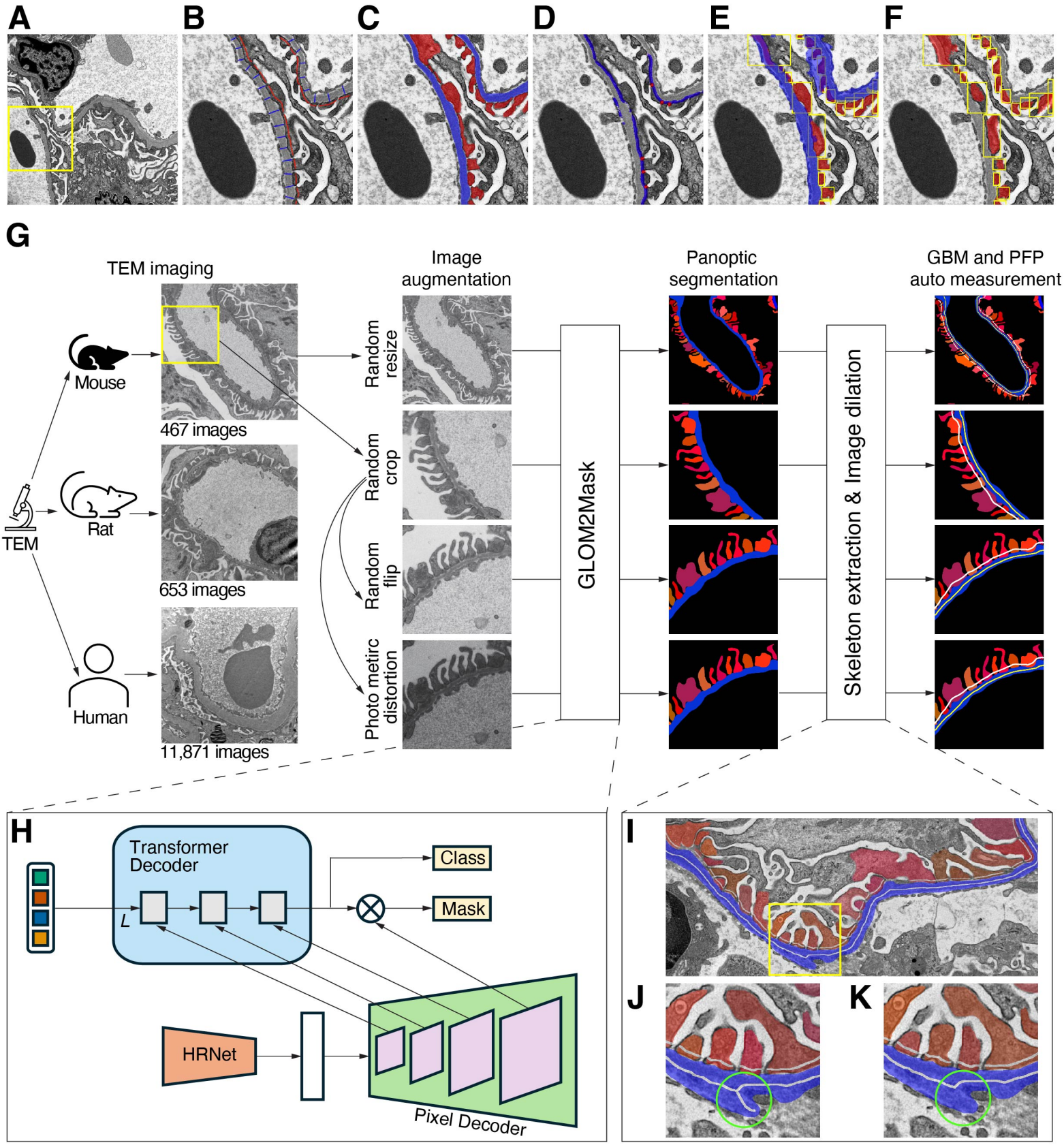
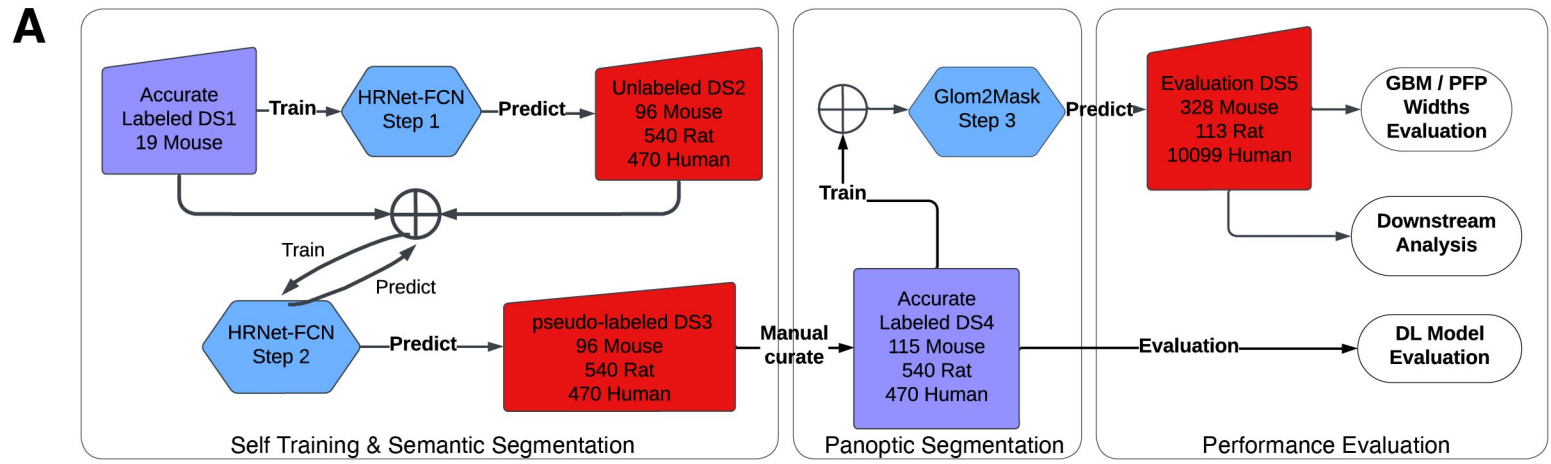
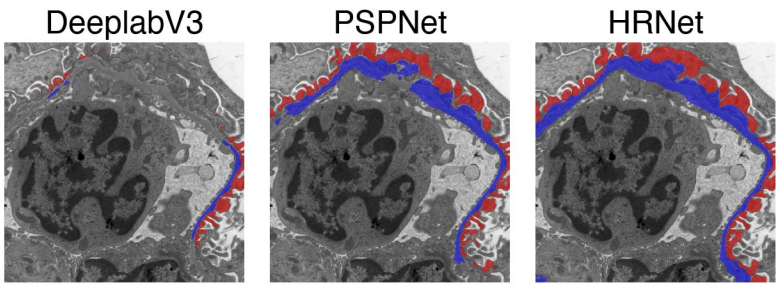

A
B
C
D
E
F
G
TEM imaging
Image augmentation
Panoptic segmentation
GBM and PFP auto measurement
Mouse
Rat
Human
TEM
467 images
653 images
11,871 images
Random resize
Random crop
Random flip
Photo metirc distortion
GLOM2Mask
Skeleton extraction & Image dilation
H
Transformer Decoder
L
Class
Mask
HRNet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

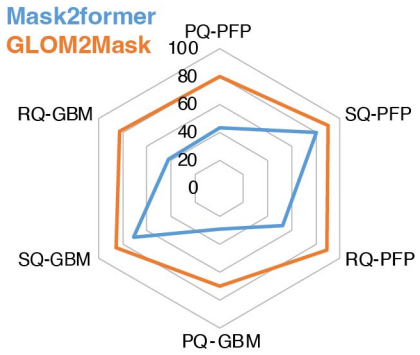
Fig. 2

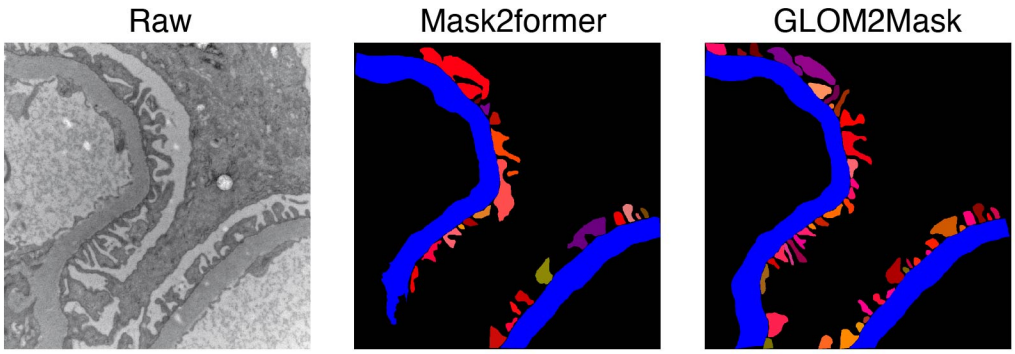


B

Model	Mouse		Rat		Human		Time (h)
	GBM	PFP	GBM	PFP	GBM	PFP	
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







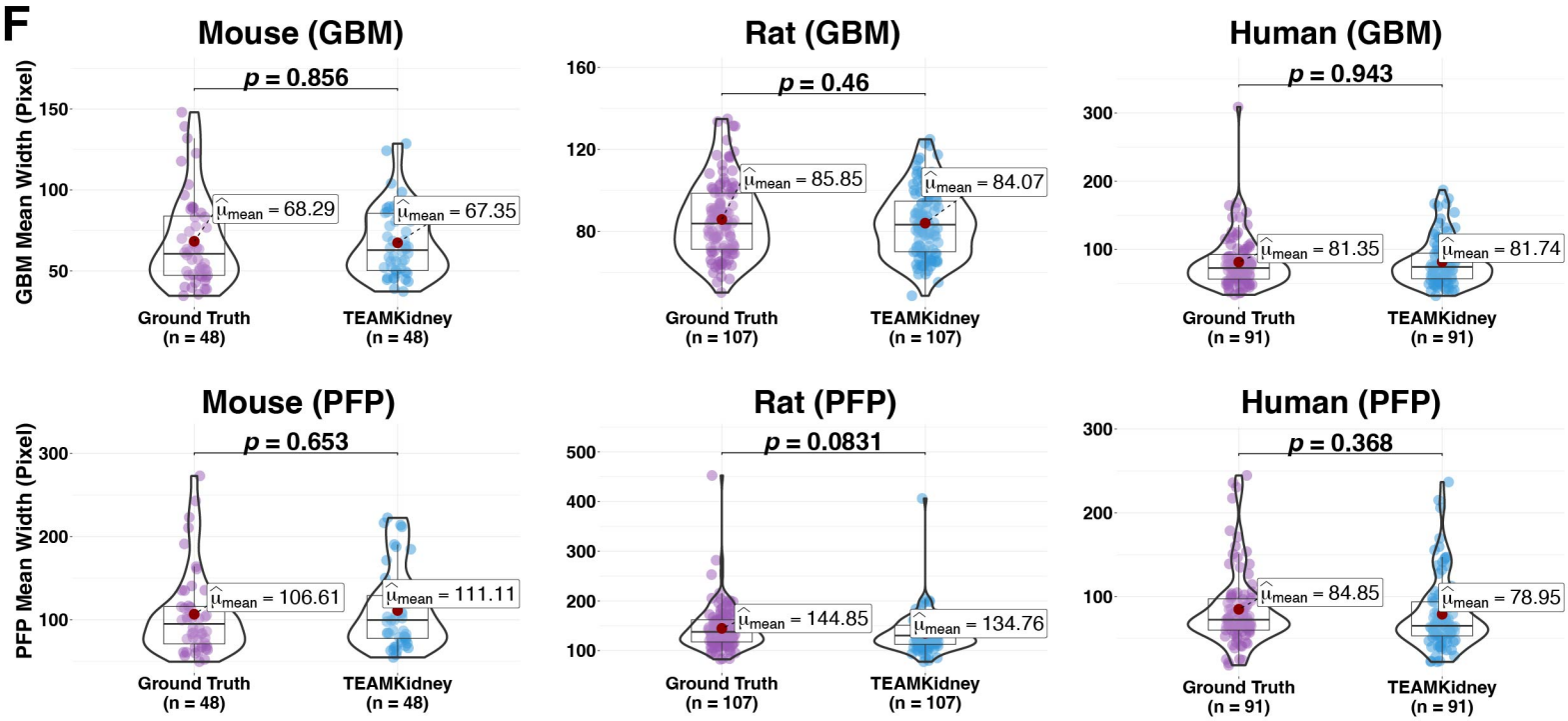


Fig. 3

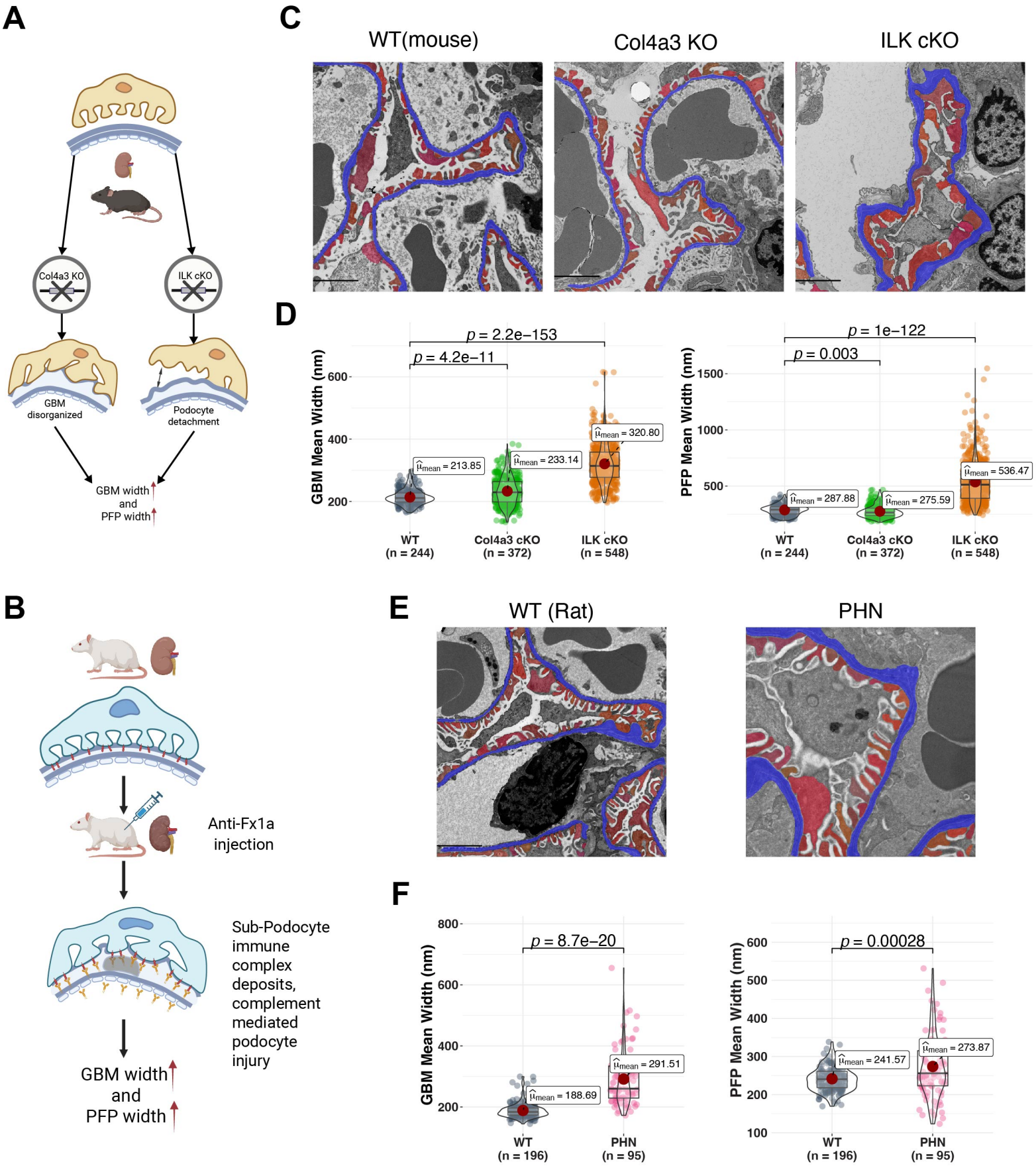

A
Col4a3 KO
ILK cKO
GBM disorganized
Podocyte detachment
GBM width and PFP width
B
Anti-Fx1a injection
Sub-Podocyte immune complex deposits, complement mediated podocyte injury
GBM width and PFP width
C
WT(mouse)
Col4a3 KO
ILK cKO
D
GBM Mean Width (nm)
p = 2.2e−153
p = 4.2e−11
μ̂mean = 213.85
μ̂mean = 233.14
μ̂mean = 320.80
WT (n = 244)
Col4a3 cKO (n = 372)
ILK cKO (n = 548)
PFP Mean Width (nm)
p = 1e−122
p = 0.003
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)
PFP Mean Width (nm)
p = 0.00028
μ̂mean = 241.57
μ̂mean = 273.87

Fig. 4

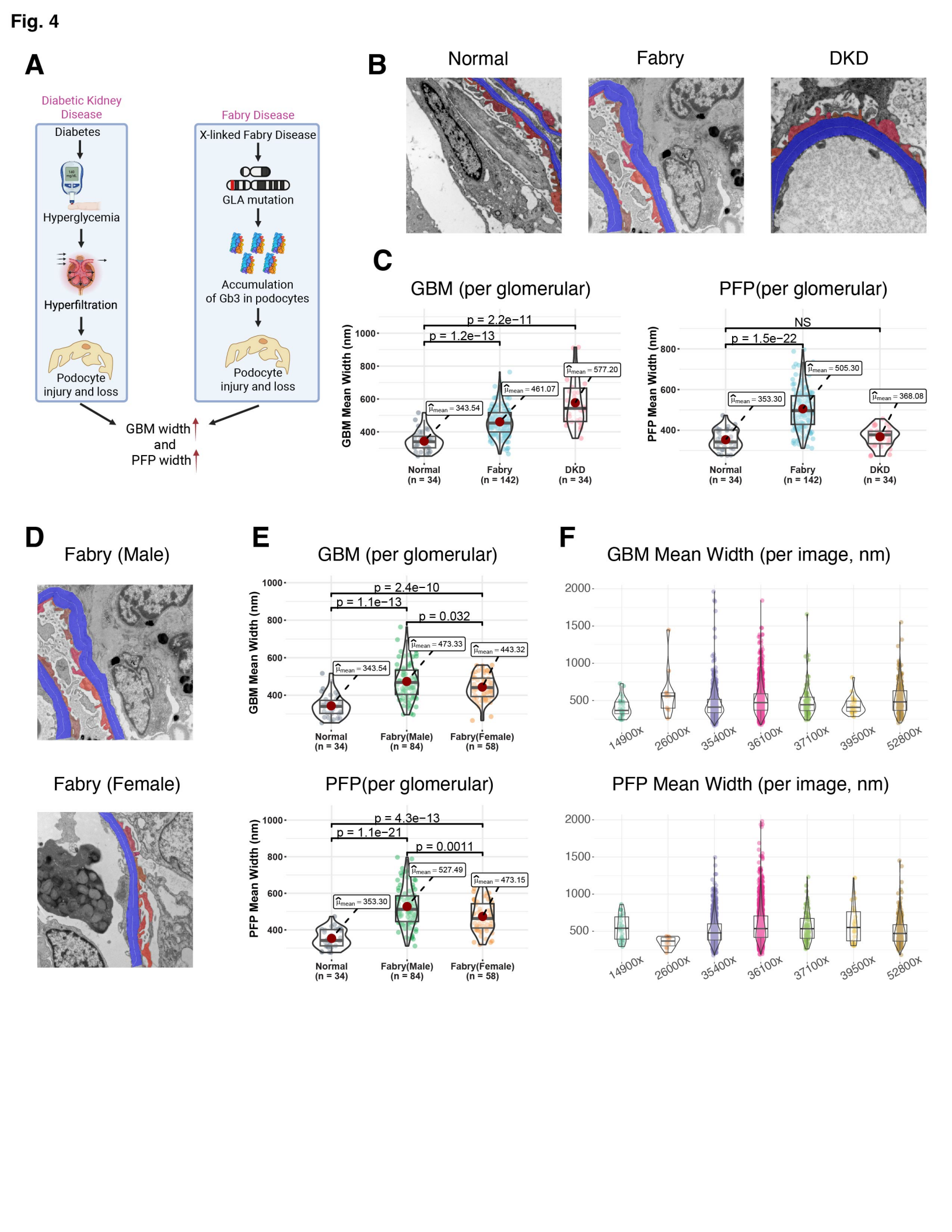

A
Diabetic Kidney Disease
Diabetes
Hyperglycemia
Hyperfiltration
Podocyte injury and loss
Fabry Disease
X-linked Fabry Disease
GLA mutation
Accumulation of Gb3 in podocytes
Podocyte injury and loss
GBM width ↑ and PFP width ↑
B
Normal
Fabry
DKD
C
GBM (per glomerular)
GBM Mean Width (nm)
p = 2.2e−11
p = 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)
PFP(per glomerular)
PFP 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)
PFP(per glomerular)
PFP 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 Mean Width (per image, nm)
PFP Mean Width (per image, nm)
14900x
26000x
35400x
36100x
37100x
39500x
52800x

Fig. 5

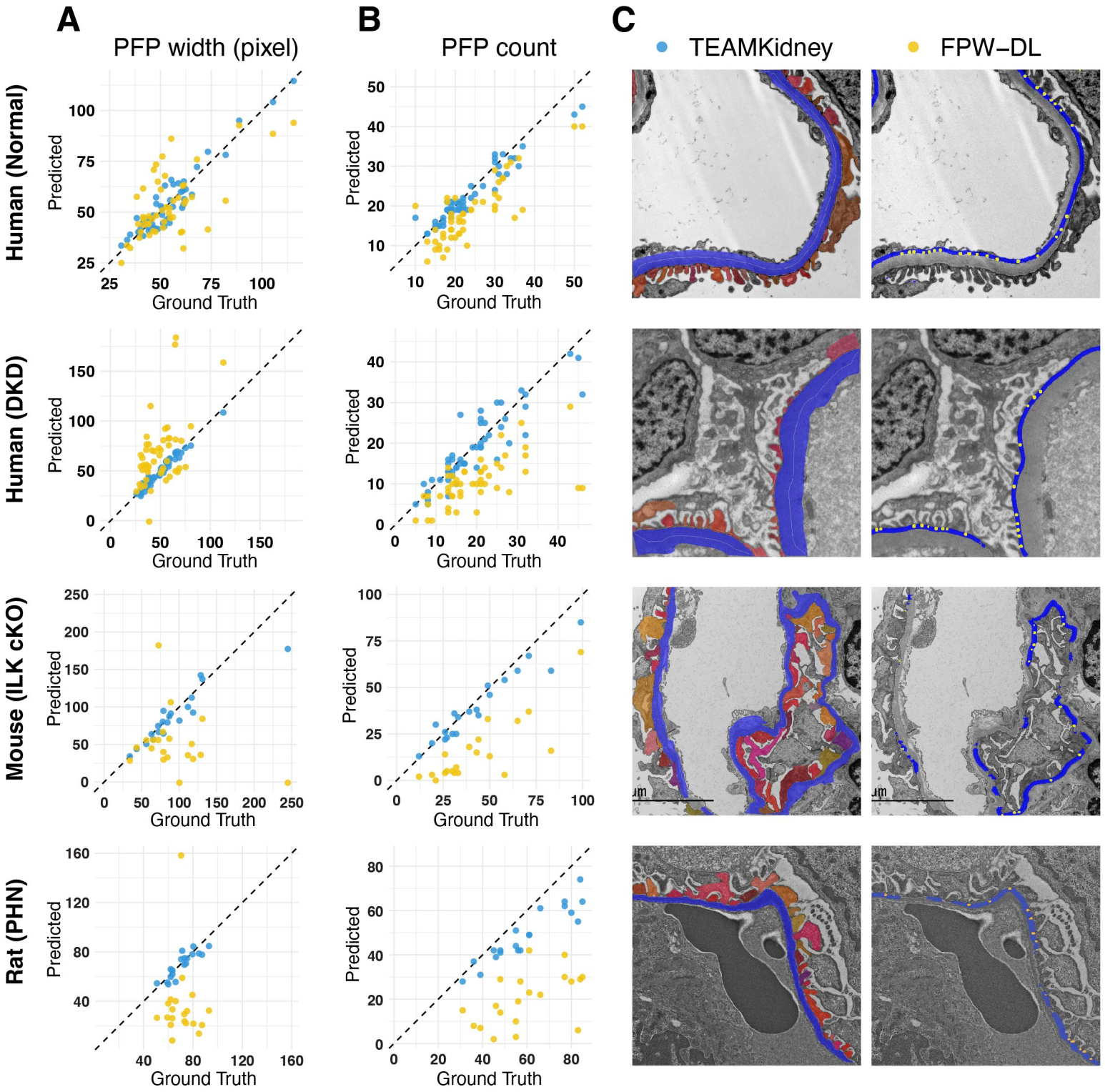

A
PFP width (pixel)
B
PFP count
C
TEAMKidney
FPW−DL
Human (Normal)
Human (DKD)
Mouse (ILK cKO)
Rat (PHN)
Predicted
Ground Truth

Fig. 6

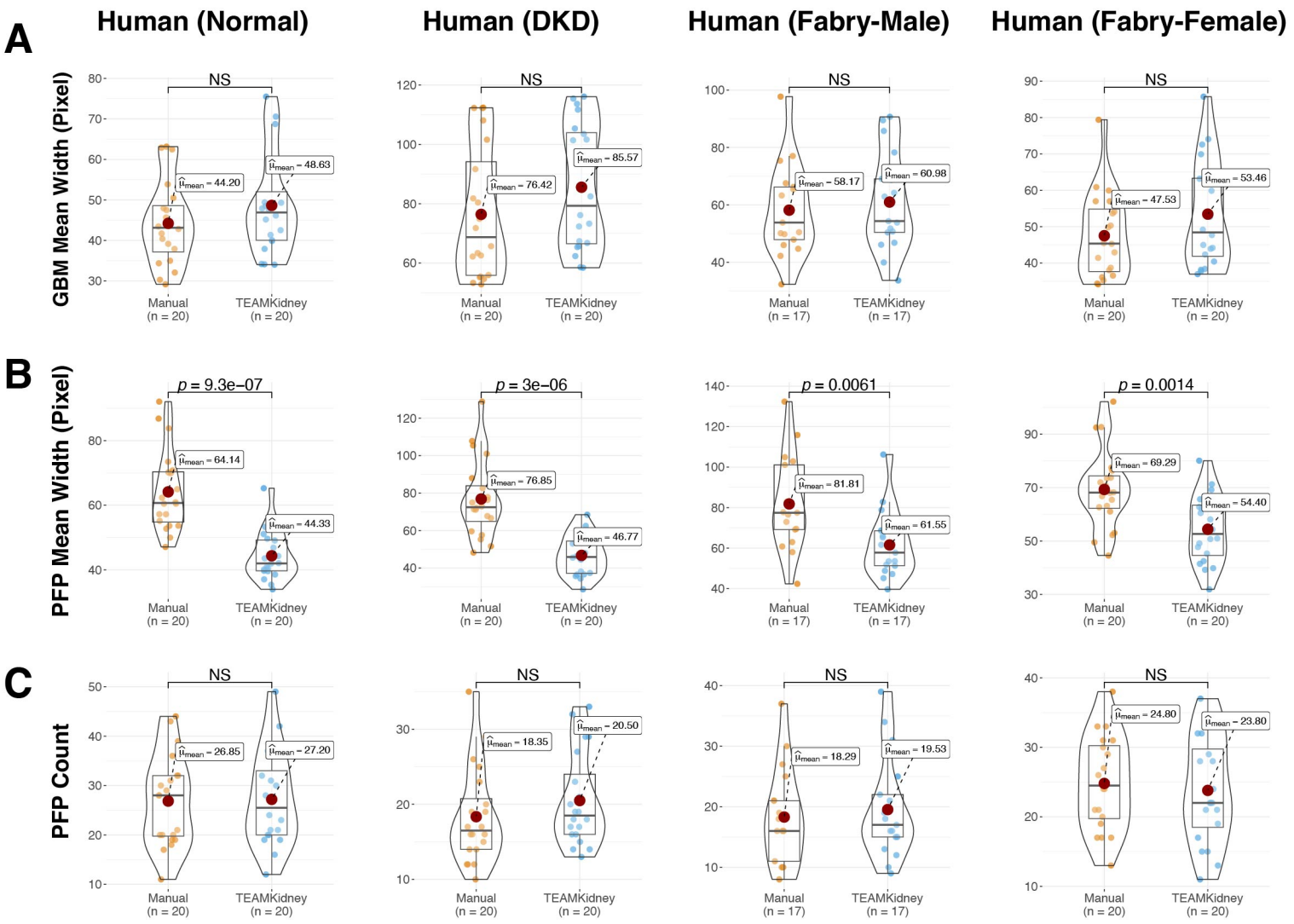
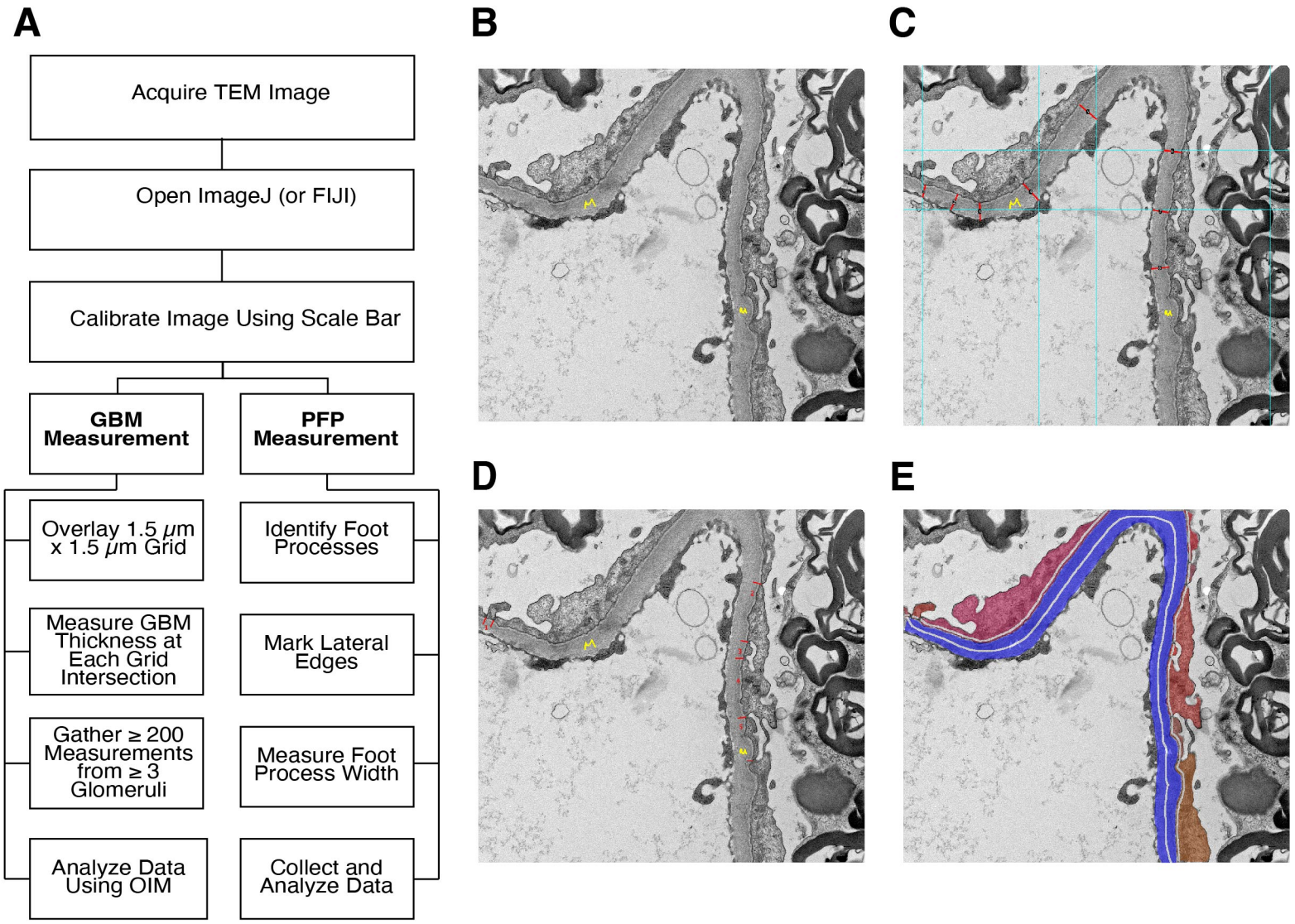

A
Human (Normal)
Human (DKD)
Human (Fabry-Male)
Human (Fabry-Female)
GBM Mean Width (Pixel)
NS
$\hat{\mu}_{mean}$ = 44.20
$\hat{\mu}_{mean}$ = 48.63
$\hat{\mu}_{mean}$ = 76.42
$\hat{\mu}_{mean}$ = 85.57
$\hat{\mu}_{mean}$ = 58.17
$\hat{\mu}_{mean}$ = 60.98
$\hat{\mu}_{mean}$ = 47.53
$\hat{\mu}_{mean}$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

Fig. 7


A
Acquire TEM Image
Open ImageJ (or FIJI)
Calibrate Image Using Scale Bar
GBM Measurement
PFP 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