3D Banff Lesion Scoring for Kidney Transplant Pathology: Feasibility and Utility for Volumetric Quantification

Yanfan Zhu[1] YANFAN.ZHU@VANDERBILT.EDU
Juming Xiong[1] JUMING.XIONG@VANDERBILT.EDU
Junchao Zhu[1] JUNCHAO.ZHU@Vanderbilt.Edu
Ruining Deng[2] RUD4004@MED.CORNELL.EDU
Shilin Zhao[3] SHILIN.ZHAO.1@VUMC.ORG
Yu Wang[3] YU.WANG.2@VUMC.ORG
Yaohong Wang[4] YAOHONGWANG@MDANDERSON.ORG
Chongyu Qu[1] CHONGYU.QU@VANDETBILT.EDU
Zhengyi Lu[1] ZHENGYI.LU@VANDETBILT.EDU
Yuechen Yang[1] YUECHEN.YANG@VANDETBILT.EDU
Mengment Yin[3] MENGMENG.YIN.1@VUMC.ORG
Yuqing Liu[5] LIUYUQING0307@TONGJI.EDU.CN
Yihan Wang[3] YIHAN.WANG@VUMC.ORG
Andrew J Rauchr[3] ANDREW.RAUCH@VUMC.ORG
Haichun Yang[3] HAICHUN.YANG@VUMC.ORG
Yuankai Huo[1] YUANKAI.HUO@VANDERBILT.EDU

[1] *Vanderbilt University, Nashville TN 37235, USA*
[2] *Weill Cornell Medicine, New York, NY 10021, USA*
[3] *Vanderbilt University Medical Center, Nashville TN 37232, USA*
[4] *UT MD Anderson Cancer Center,TX 77030, USA*
[5] *Tongji University School of Medicine, Shanghai, 200092, China*

Editors: Under Review for MIDL 2026

Abstract

Quantifying glomerular inflammation in kidney-transplant biopsies is traditionally performed on single whole-slide sections using the Banff "most-severe section" rule, despite the inherently three-dimensional nature of lesion distribution. This 2D based assessment might lead to instability when inflammation varies across slices. This paper reports a pilot study examining the technical feasibility and clinical relevance of extending Banff scoring into three dimensions. To this end, we propose a 3D Banff lesion-scoring framework that reconstructs glomeruli from serial sections, aligns structural counterparts, tracks glomerular identities, and integrates inflammatory-cell counts in 3D. In the experiments, glomerulitis (g-scores) was used as an example Banff metric and applied to multi-section renal allograft biopsies. Our findings indicate that 3D Banff g-scores are more consistent across slices and , under the semi-automatic setting, correlate more strongly with clinical biomarkers than traditional 2D scores. These results show that 3D volumetric quantification offers promising added value, underscoring the potential benefit of 3D-aware Banff scoring for kidney transplant pathology.

Keywords: Banff Classification, Kidney Transplant Pathology, Deep Learning, Digital Pathology

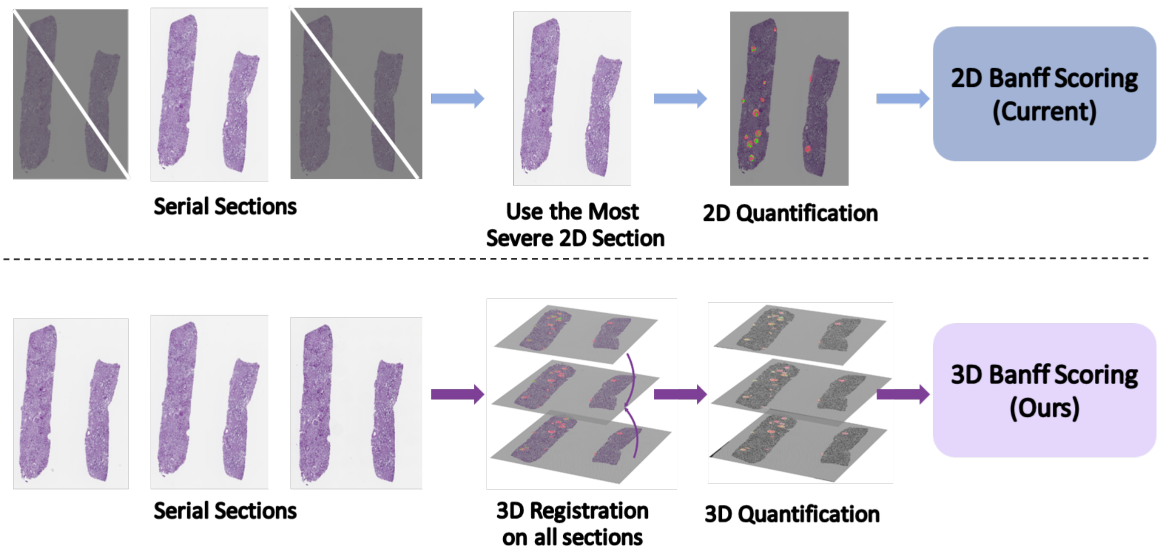
Serial Sections
Use the Most Severe 2D Section
2D Quantification
2D Banff Scoring (Current)
Serial Sections
3D Registration on all sections
3D Quantification
3D Banff Scoring (Ours)

Figure 1: **2D Banff scoring vs. 3D Banff scoring. Upper panel** shows the 2D Banff pipelines, where each section is analyzed independently and the final Banff g-score is assigned based on the most severe single section. This "worst-section" rule might ignore cross-section correspondence and be sensitive to sampling variability. **Lower panel** presents the proposed 3D Banff framework, which registers serial sections into a 3D volumetric quantification.

1. Introduction

Renal allograft biopsy is fundamental to diagnosing rejection and determining post-transplant management. The Banff Classification of Allograft Pathology provides the international standard for evaluating renal transplant biopsies, using a set of semiquantitative lesion scores—including interstitial inflammation (i), tubulitis (t), peritubular capillaritis (ptc), intimal arteritis (v), and glomerulitis (g)—to diagnose and grade rejection (Farris et al., 2025; Banff Foundation for Allograft Pathology, 2023; Zhu et al., 2025). Although a single biopsy block is routinely cut into 20–30 serial sections, each representing a slightly different tissue depth, current clinical practice assigns the g-score using only one two-dimensional (2D) histology section—the slice showing the most severe inflammation. This worst-section approach aims to avoid underestimating rejection but is inherently sensitive to sampling variation, sectioning depth, and local tissue distortion. Clinically, glomerular inflammation is a three-dimensional (3D) biological phenomenon, yet its scoring relies on a single 2D snapshot. This mismatch between disease geometry and the evaluation protocol leads to variability in diagnosis and reduces interpretability for both pathologists and clinicians(Loupy et al., 2022).

Recent advances in computational pathology have produced accurate models for segmenting glomeruli(Deng et al., 2023, 2025; Jiang et al., 2021; Yu et al., 2024) and detecting inflammatory cells(Deotale et al., 2025; Lynn, 2023) on whole slide images(WSIs)(Labriffe et al., 2022; Yi et al., 2024). However, nearly all existing systems operate strictly in 2D: each WSI is processed independently; predictions are not aligned or reconciled across serial sections;and case-level scoring still selects the single most severe slice, replicating the clinical worst-section paradigm. These approaches fail to reconstruct the 3D topology of injury, do not assign consistent glomerulus identity across depth, and cannot quantify volumetric inflammatory distribution. From a clinical standpoint, this limits robustness, explainability, and trustworthiness, all essential for adoption of AI tools in routine transplant workflows.

We propose a shift from 2D slice-based evaluation to biopsy-level 3D volumetric scoring. Our framework aligns serial histology sections, reconstructs glomeruli in 3D, aggregates inflammatory-cell detections, and produces a biopsy-level Banff g-score derived from the full depth of the tissue block. This requires addressing several technical challenges: nonlinear distortions and missing sections that hinder registration; substantial variation in staining and section quality; identity tracking of glomeruli across levels; and designing an interpretable and clinically grounded method for compatible and volumetric aggregation. To overcome these challenges, we developed a unified computational pipeline that integrates robust 2D feature extraction, deformable 3D alignment, glomerulus identity tracking, and 3D g-score computation (Figure 1). Our contributions are as follows:

- This paper presents a pilot study examining the technical feasibility and clinical relevance of extending Banff scoring from 2D to 3D.
- We propose a computational framework for the quantitative measurement of the widely used Banff g-score, further extending it to 3D through the integration of AI-based section detection, glomerular segmentation, inflammatory-cell detection, and structural alignment across serial histological sections.
- We formulate two biopsy-level *3D Banff g-scoring* measurement strategies (compatible and volumetric), that summarizes the volumetric distribution of glomerular inflammation and benchmark their performance against 2D Banff strategy.

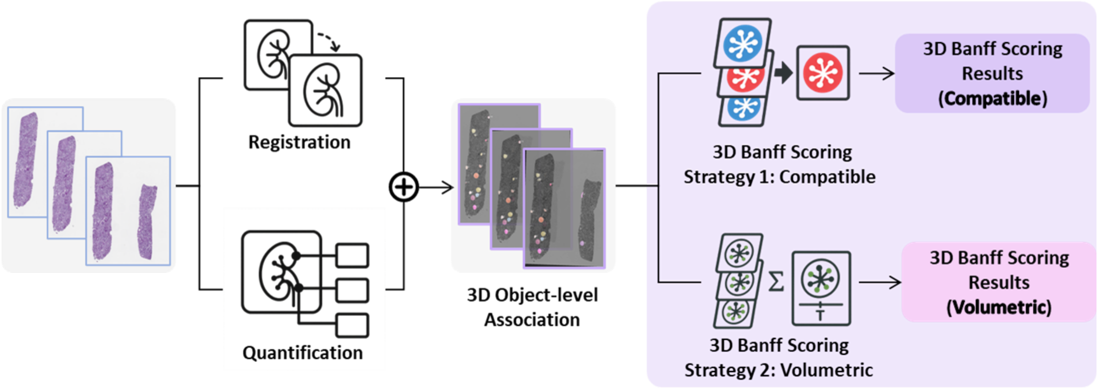
Registration
Quantification
3D Object-level Association
3D Banff Scoring Strategy 1: Compatible
3D Banff Scoring Results (Compatible)
Σ
3D Banff Scoring Strategy 2: Volumetric
3D Banff Scoring Results (Volumetric)

Figure 2: **Overview of the proposed 3D Banff lesion scoring pipeline.** Left: serial sections undergo *registration* and (*quantification*); Middle: quantification features are combined with registration outputs, yielding consistent 3D object associations for depth-wise tracking; Right: two 3D scoring paradigms are computed: *Strategy 1: Compatible*, and *Strategy 2: Volumetric*. Details of both strategies are provided in Figure 3.

2. Methods

2.1. Overall Pipeline and Data Processing

Figure 2 illustrates the complete workflow of the proposed 3D Banff lesion scoring pipeline. Raw serial-section WSIs first enter two parallel modules: **Registration**, which aligns all sections from a biopsy into a common spatial frame, and **Quantification**, which uses deep-learning models to extract per-section glomerular and inflammatory-cell features. The outputs of these modules are then combined to establish consistent 3D object associations for depth-wise tracking, whereby corresponding glomeruli across sections are linked based on spatial overlap, geometric continuity, and instance-level matching. These associations produce coherent 3D glomerular entities with aggregated inflammatory-cell detections along depth.

Finally, two 3D Banff scoring strategies are derived from these reconstructed volumes. **Compatible scoring** applies the traditional Banff worst-section rule in 3D by selecting the most inflamed slice for each glomerulus, while **Volumetric scoring** integrates inflammatory burden across the entire 3D glomerular volume to produce a depth-aware, biopsy-level g-score.

2.2. Per-section (2D) Banff Glomerulitis Scoring

As a baseline consistent with conventional pathology workflows, we implement an automated per-section g-score following the classical Banff definition of glomerulitis as the presence of leukocytes within one or more glomerular capillaries. After glomerulus segmentation and inflammatory-cell detection on each section, let $\{R_1, \dots, R_N\}$ denote the segmented glomeruli and $\mathcal{P}$ the detected cells. For each glomerulus R_k,

$$n_k = \sum_{p_i \in \mathcal{P}} \mathbb{I}[p_i \in R_k], \qquad \delta_k = \mathbb{I}[\, n_k > \tau \,],$$

with threshold $\tau = 3$. The section-level inflamed proportion is

$$\rho_g = \frac{1}{N} \sum_{k=1}^{N} \delta_k,$$

which is mapped to the Banff-style ordinal score

$$g = \begin{cases} 0, & \rho_g = 0, \\ 1, & 0 < \rho_g < 0.25, \\ 2, & 0.25 \le \rho_g \le 0.75, \\ 3, & \rho_g > 0.75. \end{cases} \tag{1}$$

This module provides an automated analogue of the traditional per-section Banff assessment. Although simplified and not intended to replicate full morphological interpretation, it captures the core signal of glomerular inflammation and serves as a reproducible reference for comparison with our 3D, cross-section–aggregated scoring (Section 2.3).

2.3. Serial-Section Alignment and 3D Banff Scoring Pipeline

For each biopsy case, serial histologic sections are scanned into WSIs and first processed to detect glomeruli and inflammatory cells on a per-section basis. All sections belonging to the same biopsy are then aligned into a common spatial frame using a hybrid registration

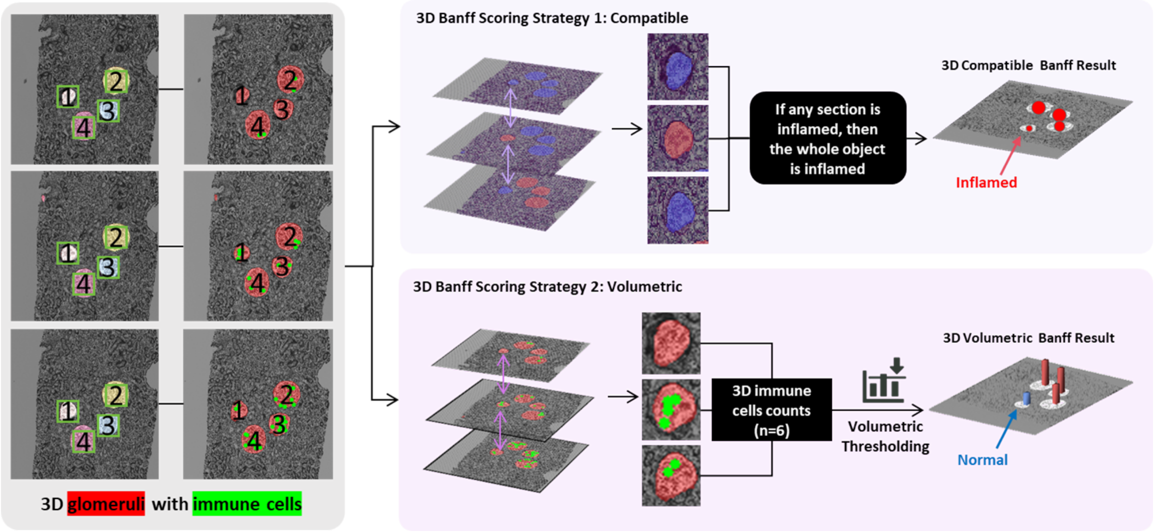
3D glomeruli with immune cells
1
2
3
4
3D Banff Scoring Strategy 1: Compatible
If any section is inflamed, then the whole object is inflamed
3D Compatible Banff Result
Inflamed
3D Banff Scoring Strategy 2: Volumetric
3D immune cells counts (n=6)
Volumetric Thresholding
3D Volumetric Banff Result
Normal

Figure 3: **3D Banff scoring strategies enabled by cross-section glomerular alignment.** Based on 3D glomeruli with immune cells, we introduce two alternative 3D scoring strategies: **Compatible strategy** applies a direct 3D extension of the Banff score, in which a glomerulus is labeled inflamed if *any* of its sectional instances is inflamed; and **Volumetric strategy** aggregates immune-cell burden across sections and applies a volumetric threshold. 3D Banff result visualization with compatible restuls pictures inflamed glomeruli as red and non-inflamed glomeruli as blue, with the volumetric strategy further represented using red and blue bar heights proportional to inflammatory load.

pipeline that combines feature matching, affine chaining, and deformable refinement to correct for tissue distortion and slice-to-slice variability. After alignment, glomerular masks and features from neighboring sections are compared to evaluate cross-sectional consistency, and corresponding glomerular instances exhibiting sufficient spatial overlap, centroid proximity, and morphological continuity are linked across depth. These matched instances are merged through a union–find–based matching procedure to establish stable 3D glomerular identities, which subsequently aggregate inflammatory-cell detections throughout the volume for downstream scoring.

As illustrated in Figure 3, once cross-sectional glomeruli are aligned and assigned consistent 3D identities, the pipeline supports two distinct 3D Banff scoring strategies. The first, the *Compatible* strategy, extends the Banff worst-section rule into 3D: if any sectional instance of a glomerulus is inflamed, the entire 3D glomerulus is labeled inflamed . The second, the *Volumetric* strategy, leverages the reconstructed 3D identity to aggregate immune-cell detections across depth, applying a volumetric threshold to distinguish inflamed from non-inflamed glomeruli. These complementary perspectives respectively capture the maximal local severity and the cumulative inflammatory burden of each glomerulus.

Strategy 1: Compatible. For each 3D glomerulus G_i with sections S_i and per-section inflammatory-cell counts $c_{i,s}$, the compatible strategy labels the glomerulus as inflamed if *any* supporting section exceeds the Banff threshold τ_{cell}:

$$\exists\, s \in S_i \quad \text{such that} \quad c_{i,s} \geq \tau_{\text{cell}}. \tag{2}$$

The biopsy-level inflamed proportion is then

$$\rho_g^{\text{comp}} = \frac{\#\{i : \exists s \in S_i,\ c_{i,s} \geq \tau_{\text{cell}}\}}{N}. \tag{3}$$

Strategy 2: Volumetric. Here, burden is aggregated across all sections belonging to the same 3D glomerulus:

$$C_i^{\text{vol}} = \sum_{s \in S_i} c_{i,s}. \tag{4}$$

A glomerulus is considered inflamed only if the total burden exceeds the section-normalized Banff threshold:

$$C_i^{\text{vol}} \geq \tau_{\text{cell}} \times |S_i|, \tag{5}$$

yielding the biopsy-level proportion

$$\rho_g^{\text{vol}} = \frac{\#\{i : C_i^{\text{vol}} \geq \tau_{\text{cell}}|S_i|\}}{N}. \tag{6}$$

Both proportions are finally mapped to Banff categories using the standard ordinal thresholds (Eq. 1). These definitions distinguish a worst-section 3D interpretation from a true depth-integrated volumetric formulation.

3. Data and Experiments

3.1. Dataset

We evaluated our framework on a retrospective cohort of renal allograft biopsies collected at Vanderbilt University Medical Center. The dataset comprises 50 patients and 410 WSIs, with each biopsy block typically sectioned into 20–30 serial histological levels stained with PAS, H&E, or Jones silver. Multiple WSIs from the same patient may correspond to different serial sections from a single biopsy block.

3.2. Object Extraction

Section ROI Detection. Section were detected by a Yolov8-based detector(Jocher and Ultralytics, 2023) trained on several WSIs with renal pathologist annotated. The model outputs section-level images for the further section-based feature extraction utilization.
Glomerulus Segmentation. Glomeruli were segmented using the WSI glomerulus segmentation pipeline proposed as 1st in the KPIs challenge (Cap, 2024). The model outputs polygon-level glomerular boundaries, which we used as spatial units for quantifying inflammation.
Inflammatory Cell Detection. The Monkey Challenge inflammatory-cell detector (Lynn, 2023) was used as the initial model and further finetuned on two renal-transplant biopsy cases that were fully QA-corrected by a renal pathologist, using a YOLOv8-based architecture (Jocher and Ultralytics, 2023). The finetuned detector provides morphology-compatible outputs suitable for downstream QA and for use in our 3D scoring pipeline.
Registration and 3D Tracking. Serial sections were aligned to a common reference using a zero-shot dense feature matching pipeline (Langlois et al., 2024) with RANSAC-based affine estimation, scale-conjugate transform recovery, and affine chaining. To handle

whole-slide geometry, masks and point coordinates were warped via a tiled affine strategy. Glomerular instances were extracted per section using connected-components segmentation with morphological refinement. Correspondences across sections were established by a multi-criteria matcher leveraging dilated IoU, centroid proximity, and area-ratio consistency, and merged via a Union–Find structure to yield 3D-consistent glomerulus identifiers, enabling depth-wise aggregation of inflammatory-cell detections.

All experiments were performed on NVIDIA A6000 GPUs under Python 3.9 and PyTorch 1.12.

3.3. 2D vs. 3D Banff Scoring

Since the scoring procedure is fully described in Section 2, here we briefly summarize how the two evaluation settings are compared.

2D setting. Each biopsy section is scored independently, and a case-level score is obtained using the conventional worst-section strategy.

3D setting. With cross-section registration and glomerulus identities tracking, inflammation evidence is aggregated across serial sections to produce a single 3D score per WSIs.

We evaluate the two scoring strategies in two ways: (1) across 50 patients with 410 WSIs, we assess the variance and stability differences between 2D and 3D scoring; and (2) for a subset of 9 patients with available longitudinal clinical follow-up, we compare their clinical validity (Section 3.4), under both semi-automatic and fully-automatic settings.Because expert Banff g scores are known to exhibit interobserver inconsistency, clinical outcome measures serve as the reference signal for evaluating the utility of the proposed scoring methods.

3.4. Evaluation Metrics

We assess the AI-derived 2D and 3D Banff g-scores using both semi-automatic and fully automatic pipelines and evaluate their clinical relevance against longitudinal renal functional outcomes in nine patients. For patient-level correlation analysis, when multiple WSIs are available from the same patient, the corresponding case-level g-score is obtained by averaging the WSI-level g-scores under each scoring strategy. Temporal changes in serum creatinine (ΔsCr), blood urea nitrogen (ΔBUN), and creatinine clearance ($-\Delta$Ccr) over 0–6 months serve as quantitative indicators of allograft status. Clinical validity is measured using Kendall τ to capture monotonic correspondence between predicted inflammation severity and renal trajectories. To complement clinical validity, we quantify scoring consistency through intra-patient WSI agreement, defined as the proportion of section pairs assigned identical g-scores under each strategy (2D, 3D-Compatible, 3D-Volumetric). High agreement reflects robustness to section-to-section variability.

Table 1: Correlation analysis of semi-automatic vs. fully-automatic pipelines

	Semi-Automatic			**Fully-Automatic**		
Outcome	2D	Compat. 3D	Vol. 3D	2D	Compat. 3D	Vol. 3D
$\Delta \mathrm{sCr}_{0\text{–}6\mathrm{m}} \uparrow$	0.187	0.227	**0.511**	0.227	**0.233**	**0.233**
$\Delta \mathrm{BUN}_{0\text{–}6\mathrm{m}} \uparrow$	-0.093	-0.045	**0.274**	0.136	**0.140**	**0.140**
$-\Delta \mathrm{Ccr}_{0\text{–}6\mathrm{m}} \uparrow$	0.000	0.045	**0.353**	0.045	**0.047**	**0.047**

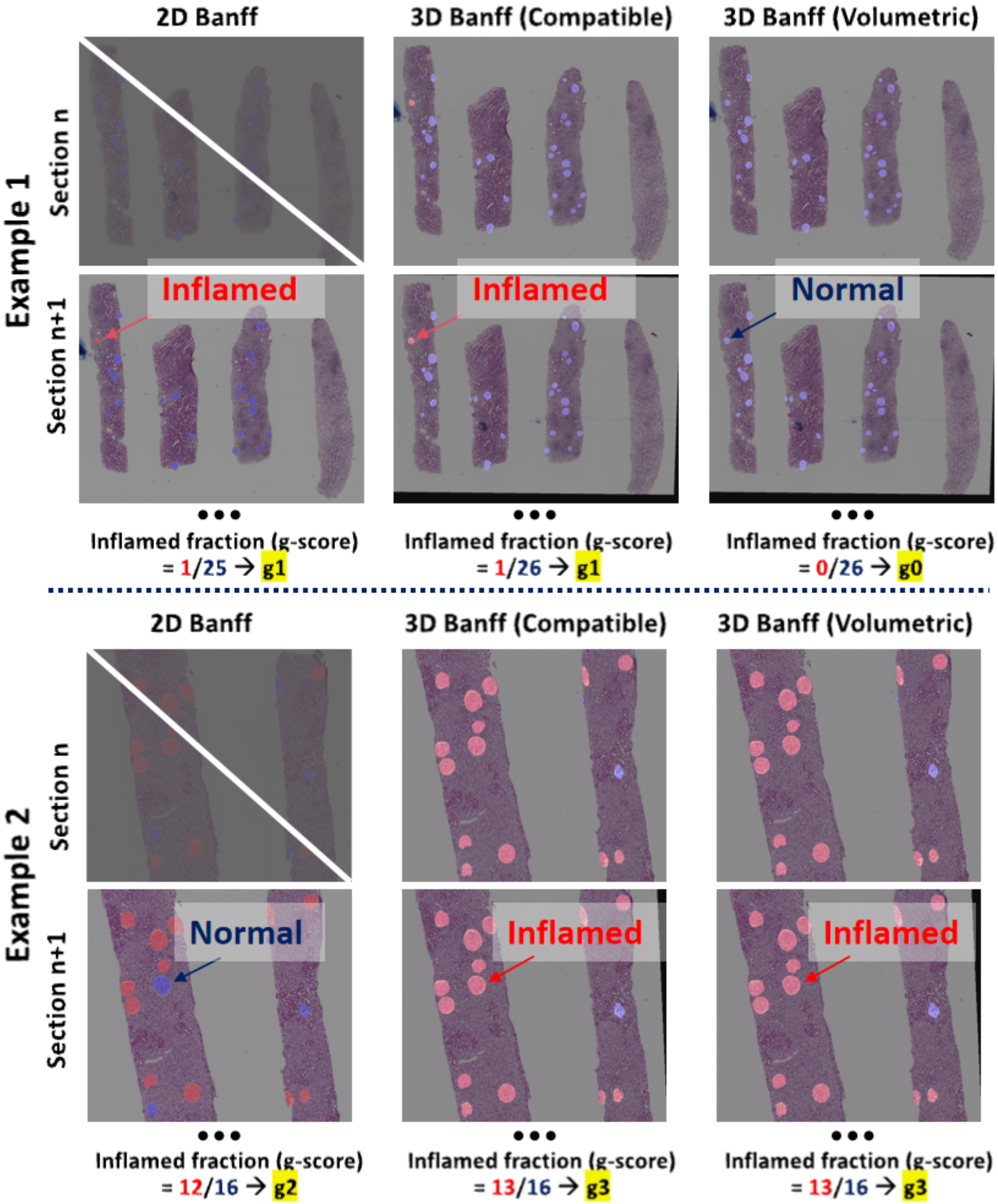
2D Banff
3D Banff (Compatible)
3D Banff (Volumetric)
Example 1
Section n
Section n+1
Inflamed
Inflamed
Normal
Inflamed fraction (g-score) = 1/25 → g1
Inflamed fraction (g-score) = 1/26 → g1
Inflamed fraction (g-score) = 0/26 → g0
2D Banff
3D Banff (Compatible)
3D Banff (Volumetric)
Example 2
Section n
Section n+1
Normal
Inflamed
Inflamed
Inflamed fraction (g-score) = 12/16 → g2
Inflamed fraction (g-score) = 13/16 → g3
Inflamed fraction (g-score) = 13/16 → g3

Figure 4: **Qualitative comparison of glomerular inflammation scoring across four analysis modes. Upper panel (Case 1):** 2D worst-section scoring detects 1/25 inflamed glomeruli (4%, g1); 3D-Compatible identifies 1/26 (3.8%, g1); 3D-Volumetric detects 0/26 (0%, g0). **Lower panel (Case 2):** 2D scoring finds 12/16 inflamed glomeruli (75%, g2); 3D-Compatible and 3D-Volumetric both detect 13/16 (81.3%, g3).

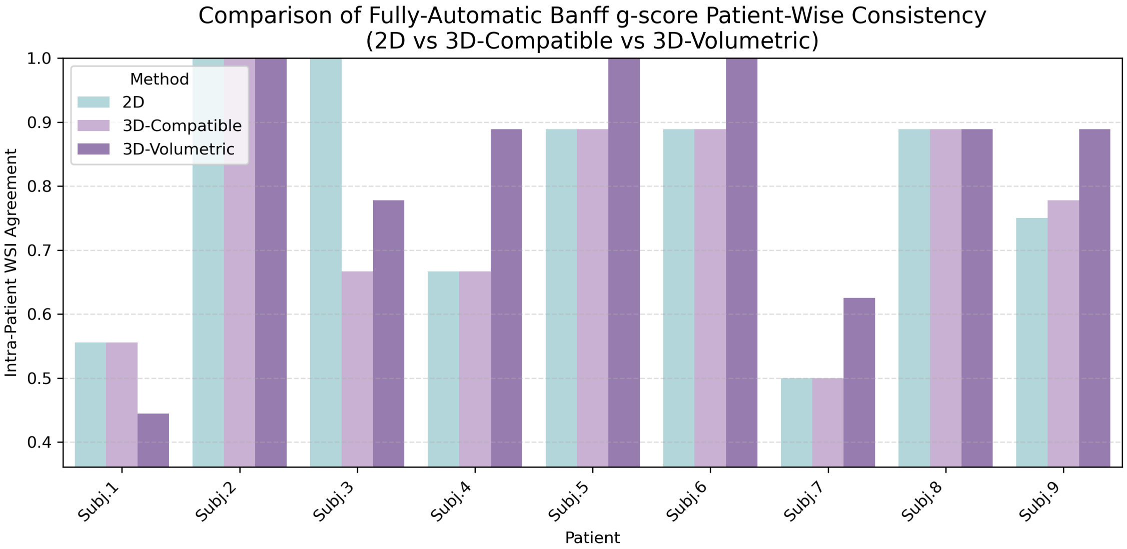

Figure 5: **Intra-patient WSI agreement of Banff *g*-scores under Semi-Automatic pipelines.** For each patient, all WSIs belonging to the same biopsy are compared pairwise, and agreement is defined as the fraction of WSI pairs assigned the same *g*-score. Three scoring strategies are evaluated: 2D worst-section scoring, 3D-Compatible scoring, and 3D-Volumetric scoring. Higher bars represent better consistency.

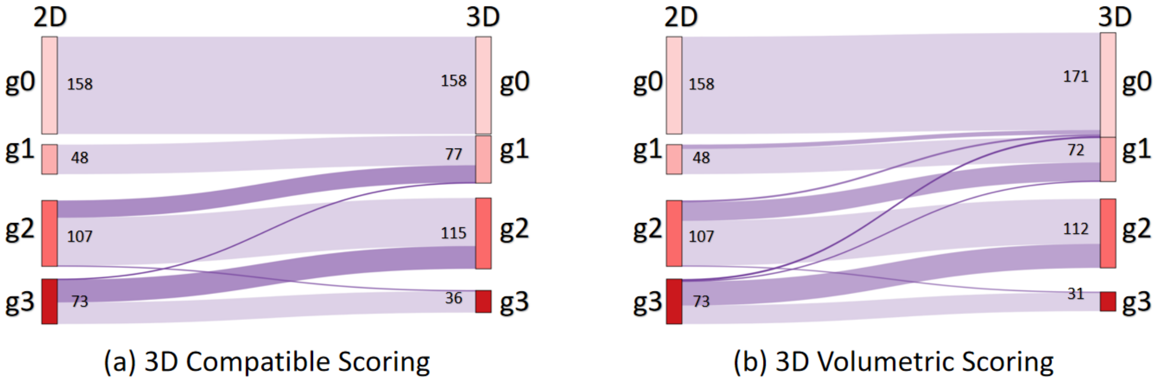

Figure 6: Alluvial comparison of 2D and 3D Banff *g*-scores across 50 patients comprising 410 WSIs. The diagram illustrates how glomerulitis grades shift when transitioning from slice-based 2D scoring to two different 3D scoring paradigms: (a) **Compatible** scoring; and (b) **Volumetric** scoring, which aggregates inflammatory burden across the reconstructed 3D glomerular volume. Thicker flows indicate more frequent grade transitions.

4. Results

Table 1 reports the Kendall's Tau correlations between Banff g-scores and short-term clinical changes (ΔsCr, ΔBUN, and $-\Delta$Ccr) under both the Semi-Automatic and Fully-Automatic pipelines. For each outcome, three scoring strategies—2D, 3D Compatible, and 3D Volumetric—are compared side by side. Boldface entries denote the highest correlation among the three scoring modes for a given outcome. Qualitative examples comparing 2D worst-section scores with the proposed 3D Banff scores are provided in Figure 4. Figure 5 presents the intra-patient WSI agreement of Banff g-scores across all WSIs belonging to the same biopsy. Agreement is computed by comparing every pair of WSIs from a given patient and measuring the proportion that receive identical g-scores. The three scoring strategies—2D, 3D-Compatible, and 3D-Volumetric—are evaluated under the Semi-Automatic pipelines. Figure 6 summarizes the distributional changes in Banff g-scores when transitioning from conventional 2D scoring to the two proposed 3D scoring strategies across 50 patients (410 WSIs) under the fully automatic pipeline. In both subfigures, the left axis shows the original 2D g-score and the right axis shows the corresponding 3D g-score.

5. Discussion

Across a cohort of 50 patients with 410 WSIs, the proposed 3D scoring strategies exhibited improved intra-patient consistency compared with conventional 2D scoring, indicating enhanced robustness to section-to-section variability. In a limited subset of 9 patients with longitudinal clinical follow-up, the Volumetric 3D scores demonstrated higher rank correlation with short-term renal function changes under the semi-automatic pipeline, suggesting that aggregating inflammatory burden across depth may better reflect the biological extent of glomerular injury. However, under the fully automatic setting, the performance differences between 2D and 3D strategies largely collapsed, highlighting that the potential clinical benefit of 3D scoring is currently constrained by the accuracy of upstream segmentation, detection, and registration modules.

Taken together, our results support the technical feasibility of reconstructing and scoring glomerular inflammation in three dimensions from routine serial histology, while also clarifying its present limitations. Rather than serving as a replacement for expert Banff assessment, the proposed framework provides a proof-of-concept for depth-aware quantification and for systematically examining how volumetric information alters conventional lesion grading. Future work will focus on validation in larger multi-center cohorts, improving fully automatic cell- and structure-level characterization, incorporating explicit 3D cell registration to enable fully quantitative volumetric inflammation metrics, integrating statistical inference for clinical association testing, and extending the 3D formulation to additional Banff lesion categories beyond glomerulitis.

6. Conclusion

In this work, we presented a computational framework for extending Banff g-scoring from conventional 2D, slice-based assessment to biopsy-level 3D volumetric quantification using serial histology sections. By integrating deep-learning–based glomeruli segmentation, inflammatory-cell detection, serial-section registration, and cross-section glomerular identity tracking, our pipeline enables both a 3D extension of the Compatible strategy and a Volumetric scoring formulation.

Acknowledgments

This research was supported by NIH R01DK135597 (Huo), DoD HT9425-23-1-0003 (HCY), and KPMP Glue Grant. This work was also supported by Vanderbilt Seed Success Grant, Vanderbilt Discovery Grant, and VISE Seed Grant. This project was supported by The Leona M. and Harry B. Helmsley Charitable Trust grant G-1903-03793 and G-2103-05128. This research was also supported by NIH grants R01EB033385, R01DK132338, REB017230, R01MH125931, and NSF 2040462. We extend gratitude to NVIDIA for their support by means of the NVIDIA hardware grant. This work was also supported by NSF NAIRR Pilot Award NAIRR240055.

References

Banff Foundation for Allograft Pathology. Central repository for the banff classification, 2023. Available at https://banfffoundation.org/central-repository-for-banff-classification-resources-3/.

Quan Huu Cap. An effective pipeline for whole-slide image glomerulus segmentation. *arXiv preprint arXiv:2411.04782*, 2024.

Ruining Deng, Quan Liu, Can Cui, Tianyuan Yao, Jun Long, Zuhayr Asad, R. Michael Womick, Zheyu Zhu, Agnes B. Fogo, Shilin Zhao, Haichun Yang, and Yuankai Huo. Omni-seg: A scale-aware dynamic network for renal pathological image segmentation, 2023. URL https://arxiv.org/abs/2206.13632.

Ruining Deng, Junchao Zhu, Juming Xiong, Can Cui, Tianyuan Yao, Junlin Guo, Siqi Lu, Marilyn Lionts, Mengmeng Yin, Yu Wang, Shilin Zhao, Yucheng Tang, Yihe Yang, Paul Dennis Simonson, Mert R. Sabuncu, Haichun Yang, and Yuankai Huo. Irs: Incremental relationship-guided segmentation for digital pathology, 2025. URL https://arxiv.org/abs/2505.22855.

Gunjan Deotale, Abhishek Ambast, Lavish Ramchandani, Dev Kumar Das, and Tijo Thomas. Ensemble object detection methodology for automated detection of inflammatory cells in kidney biopsies. In *MIDL 2025 Short Papers*, 2025. doi: pending. URL https://openreview.net/forum?id=i9Ray4JAEn. short-paper; submission number 23; 3rd place in MONKEY challenge.

Alton B. Farris, Jeroen van der Laak, and Dominique van Midden. Artificial intelligence-enhanced interpretation of kidney transplant biopsy: focus on rejection. *Current Opinion in Organ Transplantation*, 30(3):201–207, 2025. doi: 10.1097/MOT.0000000000001213. URL https://pubmed.ncbi.nlm.nih.gov/40171636/.

Lei Jiang, Wenkai Chen, Bao Dong, Ke Mei, Chuang Zhu, Jun Liu, Meishun Cai, Yu Yan, Gongwei Wang, Li Zuo, and Hongxia Shi. A deep learning–based approach for glomeruli instance segmentation from multistained renal biopsy pathologic images. *The American Journal of Pathology*, 191(8):1431–1441, 2021. doi: 10.1016/j.ajpath.2021.05.004. URL https://pubmed.ncbi.nlm.nih.gov/34294192/.

Glenn Jocher and Ultralytics. Yolov8: Newest version of yolo by ultralytics. https://github.com/ultralytics/ultralytics, 2023. Accessed: 2025-01-01.

Marc Labriffe, Jean-Baptiste Woillard, Wilfried Gwinner, Jan-Hinrich Braesen, Dany Anglicheau, Marion Rabant, Priyanka Koshy, Maarten Naesens, and Pierre Marquet. Machine learning-supported interpretation of kidney graft elementary lesions in combination with clinical data. *American Journal of Transplantation*, 22(12):2821–2833, 2022. ISSN 1600-6135. doi: https:

//doi.org/10.1111/ajt.17192. URL https://www.sciencedirect.com/science/article/pii/S1600613523000345.

Pierre-Alain Langlois, Ignacio Rocco, and Jakob Verbeek. Xfeat: Accelerated features for lightweight image matching. *arXiv preprint arXiv:2404.17407*, 2024. URL https://arxiv.org/abs/2404.17407.

Alexandre Loupy, Michael Mengel, and Mark Haas. Thirty years of the international banff classification for allograft pathology: the past, present, and future of kidney transplant diagnostics. *Kidney International*, 101(4):678–691, 2022. doi: 10.1016/j.kint.2021.11.028. URL https://doi.org/10.1016/j.kint.2021.11.028.

Meng et al. Lynn. Monkey challenge: Inflammatory cell detection in kidney allograft biopsies. https://github.com/lynn0304/MONKEY_CHALLENGE, 2023. Leaderboard and details at https://monkey.grand-challenge.org/.

Zhengzi Yi, Caixia Xi, Madhav C. Menon, Paolo Cravedi, Fasika Tedla, Alan Soto, Zeguo Sun, Keyu Liu, Jason Zhang, Chengguo Wei, Man Chen, Wenlin Wang, Brandon Veremis, Monica Garcia-Barros, Abhishek Kumar, Danielle Haakinson, Rachel Brody, Evren U. Azeloglu, Lorenzo Gallon, Philip J. O'Connell, Maarten Naesens, Ron Shapiro, Robert B. Colvin, Stephen Ward, Fadi Salem, and Weijia Zhang. A large-scale retrospective study enabled deep-learning based pathological assessment of frozen procurement kidney biopsies to predict graft loss and guide organ utilization. *Kidney International*, 105(2):281–292, 2024. doi: 10.1016/j.kint.2023.09.031. URL https://pubmed.ncbi.nlm.nih.gov/37923131/.

Lining Yu, Mengmeng Yin, Ruining Deng, Quan Liu, Tianyuan Yao, Can Cui, Junlin Guo, Yu Wang, Yaohong Wang, Shilin Zhao, Haichun Yang, and Yuankai Huo. Glo-in-one-v2: Holistic identification of glomerular cells, tissues, and lesions in human and mouse histopathology, 2024. URL https://arxiv.org/abs/2411.16961.

Yanfan Zhu, Juming Xiong, Ruining Deng, Yu Wang, Yaohong Wang, Shilin Zhao, Mengmeng Yin, Yuqing Liu, Haichun Yang, and Yuankai Huo. How close are we? limitations and progress of ai models in banff lesion scoring, 2025. URL https://arxiv.org/abs/2510.27158.