Deep-Learning based segmentation and quantification in experimental kidney histopathology

Citation for published version:
Bouteldja, N, Klinkhammer, BM, Bulow, RD, Droste, P, Otten, SW, Stillfried, SV, Moellmann, J, Sheehan, SM, Korstanje, R, Menzel, S, Bankhead, P, Mietsch, M, Drummer, C, Lehrke, M, Kramann, R, Floege, J, Boor, P & Merhof, D 2020, 'Deep-Learning based segmentation and quantification in experimental kidney histopathology', *Journal of the American Society of Nephrology*. https://doi.org/10.1681/ASN.2020050597

Digital Object Identifier (DOI):
10.1681/ASN.2020050597

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of the American Society of Nephrology

Deep Learning based segmentation and quantification in experimental kidney histopathology

Journal:	*Journal of the American Society of Nephrology*
Manuscript ID	JASN-2020-05-0597.R1
Manuscript Type:	Original Article - Basic Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Bouteldja, Nassim; RWTH Aachen University, Institute of Imaging & Computer Vision Klinkhammer, Barbara; RWTH Aachen University, Institute for Pathology Buelow, Roman; RWTH Aachen University, Institute for Pathology Droste, Patrick; RWTH Aachen University, Institute for Pathology Otten, Simon; RWTH Aachen University, Institute for Pathology Freifrau von Stillfried, Saskia; RWTH Aachen University, Institute for Pathology Moellmann, Julia; RWTH Aachen University, Department of Cardiology and Vascular Medicine Sheehan, Susan; The Jackson Laboratory Korstanje, Ron; The Jackson Laboratory Menzel, Sylvia; RWTH Aachen University, Department of Nephrology and Immunology Bankhead, Peter; University of Edinburgh, Edinburgh Pathology; University of Edinburgh, Institute of Genetics and Molecular Medicine Mietsch, Matthias; German Primate Centre Leibniz Institute for Primate Research, Laboratory Animal Science Unit Drummer, Charis; German Primate Centre Leibniz Institute for Primate Research, Platform Degenerative Diseases Lehrke, Michael; RWTH Aachen University, Department of Cardiology and Vascular Medicine Kramann, Rafael; RWTH Aachen University, Department of Nephrology and Immunology; Erasmus Medical Center, Department of Internal Medicine, Nephrology and Transplantation Flöge, Jürgen; RWTH Aachen University, Department of Nephrology and Immunology Boor, Peter; RWTH Aachen University, Institute of Pathology; RWTH Aachen University, Department of Nephrology and Immunology Merhof, Dorit; RWTH Aachen University, Institute of Imaging and Computer Vision
Keywords:	digital pathology, segmentation, histopathology, animal model

Significance Statement (109/120 Words)

Preclinical animal experiments are of high importance in nephrology research, with histology as a major readout. Here, the authors provide a multiclass histology segmentation tool to evaluate animal kidney disease models using deep learning. A convolutional neural network (CNN) enabled a rapid, automated, high-performance whole slide segmentation of renal histology, allowing high-throughput analyses in various species and multiple murine disease models. The CNN also showed high performance in patient samples, providing a translational bridge between preclinical and clinical research. Extracted quantitative morphological features closely correlated with standard morphometric measurements. In conclusion, deep learning-based segmentation in experimental renal pathology opens new dimensions of reproducible, unbiased and high-throughput quantitative digital nephropathology.

Deep Learning based segmentation and quantification in experimental kidney histopathology

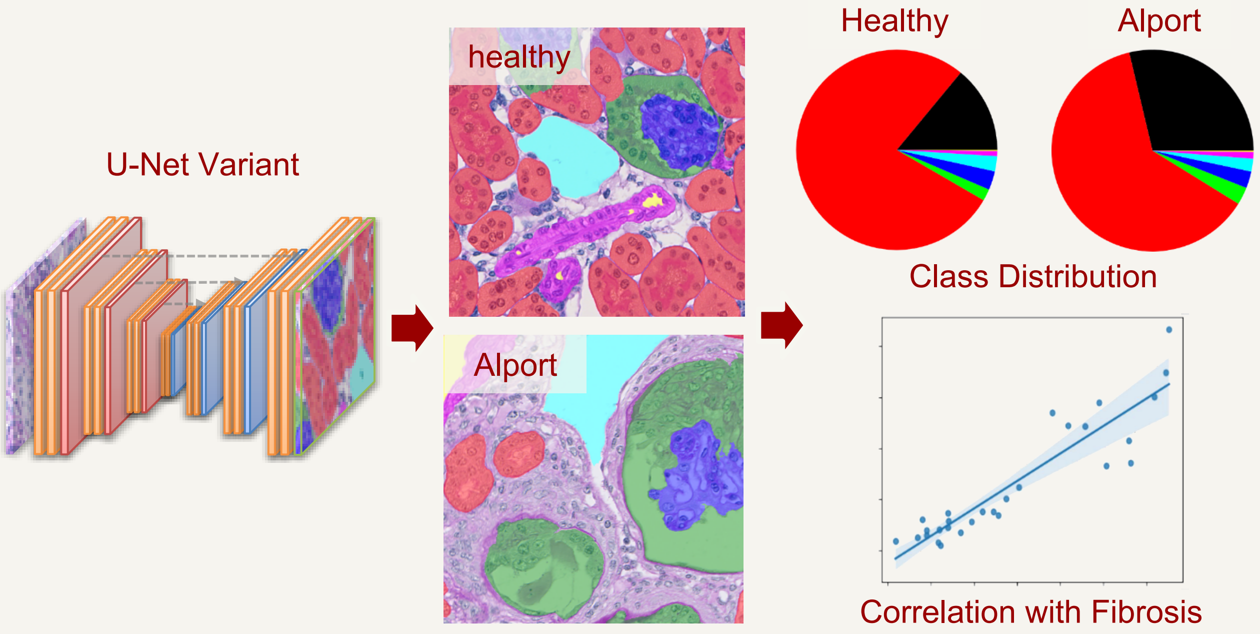
METHODS

- 5 murine disease models, 6 species
- 72722 Annotations in 2930 patches (2100 Training / 160 Val. / 670 Test)
- Classes:
  - Tubule
  - Glomerular tuft
  - Full glomerulus
  - Artery
  - Arterial lumen
  - Vein
  - Remaining tissue

RESULTS

- Instance Dice Scores:
  91.9% Tubule, 96.5% Glom., 94.7% Tuft, 84.1% Artery, 78.2% Lumen, 94.2% Vein
- Strong IHC/fibrosis correlations with remaining tissue area coverage

Kidney Whole Slide Segmentation – Quantitative Analysis



U-Net Variant

healthy

Alport

Healthy

Alport

Class Distribution

Correlation with Fibrosis

CONCLUSION

Accurate multispecies-, multimodel- Whole Slide Segmentation enabling automated quantitative analysis of renal histopathology and facilitating high-throughput experimental nephropathology.

Deep-Learning based ~~multi-disease, multi-species, multi-class~~ segmentation and quantification in experimental kidney histopathology

Running Title: DL in experimental nephropathology

Nassim Bouteldja[2,*], Barbara M. Klinkhammer[1,3,*], Roman D. Bülow[1,*], Patrick Droste[1], Simon W. Otten[1], Saskia von Stillfried[1], Julia Moellmann[4], Susan M. Sheehan[5], Ron Korstanje[5], Sylvia Menzel[3], Peter Bankhead[6,7], Matthias Mietsch[8], Charis Drummer[9], Michael Lehrke[4], Rafael Kramann[3,10], Jürgen Floege[3], Peter Boor[1,3,*,#], Dorit Merhof[2,11,*]

1 Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany
2 Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany
3 Department of Nephrology and Immunology, RWTH Aachen University Hospital, Aachen, Germany
4 Department of Cardiology and Vascular Medicine, RWTH Aachen University Hospital, Aachen, Germany
5 The Jackson Laboratory, Bar Harbor, Maine
6 Edinburgh Pathology, University of Edinburgh, Edinburgh, UK
7 Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
8 Laboratory Animal Science Unit, German Primate Center, Goettingen, Germany
9 Platform Degenerative Diseases, German Primate Center, Goettingen, Germany
10 Department of Internal Medicine, Nephrology and Transplantation, Erasmus Medical Center, Rotterdam, The Netherlands
11 Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

* Authors contributed equally
# Address correspondence to:

Peter Boor, MD, PhD
Institute of Pathology
RWTH Aachen University Hospital
Pauwelsstrasse 30
52074 Aachen, Germany
Phone: +49 241 80 85227
Fax: +49 241 80 82446
E-mail: pboor@ukaachen.de

Abstract: 246
Main text (excl. Method section): 3452/3500

Key Words: Digital pathology, Segmentation, Animal models, Histopathology

Significance Statement (109/120 Words)

Preclinical animal experiments are of high importance in nephrology research, with histology as a major readout. Here, the authors provide a multiclass histology segmentation tool to evaluate animal kidney disease models using deep learning. A convolutional neural network (CNN) enabled a rapid, automated, high-performance ~~multiclass-, multispecies- and multi-disease~~ whole slide segmentation of renal histology, allowing high-throughput analyses in various species and multiple murine disease models. The CNN also showed high performance in patient samples, providing a translational bridge between preclinical and clinical research. Extracted quantitative morphological features closely correlated with ~~gold-~~standard morphometric measurements. In conclusion, deep learning-based segmentation in experimental renal pathology opens new dimensions of reproducible, unbiased and high-throughput quantitative digital nephropathology.

Abstract (246/250 Words)

Background: Preclinical animal models are essential for understanding kidney disease pathophysiology and for identifying novel diagnostic and therapeutic approaches. Nephropathological analyses represent major outcome parameters of such models. With increasing demands on precision medicine, novel high-throughput tools for quantitative, unbiased, reproducible and efficient histopathological analyses are required.

Methods: We propose a convolutional neural network (CNN) architecture for accurate segmentation of PAS stained kidney tissue of healthy mice and five commonly used murine disease models and other species used in preclinical research. The CNN was trained to segment six major renal structures, i.e. glomerular tuft, glomerulus including Bowman's capsule, tubules, arteries, arterial lumina, and veins. To achieve high accuracy, we performed a large number of expert-based annotations (~~68,523~~72,722 in total).

Results: Multiclass segmentation performance was very high in all disease models. The CNN allowed high-throughput and large-scale, quantitative and comparative analyses of various models. Computational feature extraction in disease models revealed interstitial expansion, tubular dilation and atrophy, and glomerular size variability. Validation showed a high correlation with the current ~~gold-~~standard morphometric analysis. The CNN also showed high performance in other species used in research, including rats, pigs, bears, and marmosets as well as in humans, providing a translational bridge between preclinical and clinical studies.

Conclusions: We have developed a deep learning algorithm for accurate multiclass segmentation of digital whole-slide images of PAS stained kidneys from various species and renal disease models. This enables highly reproducible quantitative

histopathology analyses in preclinical models, potentially also applicable to clinical studies.

Introduction

Many basic science and preclinical studies require experiments in animals with histopathological assessment representing a major readout. The demands on robust but at the same time objective, precise and quantitative data steadily increase. In both clinical practice and research, histopathological evaluations are often performed manually. This is both time-consuming and not seldom poorly reproducible, particularly if not performed by experts. The projected decrease in pathologist workforce, which is particularly noticeable in highly specialized fields like nephropathology, and heavy engagement in clinical duties further complicate the situation[1].

High-throughput digitization of histological slides, generating so-called whole slide images (WSIs), enables the effective use of computer-assisted histopathological analysis. Deep Learning (DL) is a subset of artificial intelligence (AI) that applies computer algorithms to find meaningful representations of raw data through multiple layers of abstraction[2]. DL's most popular technique, the Convolutional Neural Network (CNN), is increasingly applied in pathology[3] due to its high performance in tasks like detection of nuclei[4], histology segmentation[5] or prediction of molecular alterations from hematoxylin- and eosin-stained (H&E) sections[6]. We have previously shown that ML- and DL-based techniques can facilitate glomerulus detection and segmentation in WSIs[7-10]. Recently, two other groups reported the feasibility of the DL-based segmentation of human kidney WSIs[11, 12] and glomerulus segmentation was already successfully used for subsequent analysis of glomerulosclerosis in PAS[13, 14] or Trichrome-stained biopsies[15]. The usefulness of DL in animal models with broad histopathological injury patterns was not yet analyzed.

Our main aim was to develop a CNN for multiclass segmentation of mouse kidney Periodic Acid Schiff (PAS)-stained histology, focusing on five commonly used models of kidney diseases. We demonstrate the applicability of our CNN for large-scale histopathological segmentation followed by quantitative data extraction and confirm the performance by correlation with traditional image analysis tools. We also show the ~~cost effective~~ applicability for other species used in research, as well as for patient kidney samples.

Methods

Histology samples

We used paraffin-embedded kidney tissue fixed in formalin or methyl Carnoy's solution. 1-2 µm thick sections were stained with periodic acid–Schiff (PAS) and counterstained with hematoxylin. Slides were digitalized using the whole-slide scanners NanoZoomer HT2 with 20x objective (Hamamatsu Photonics, Hamamatsu, Japan) or Aperio AT2 with 20x or 40x objective (Leica Biosystems, Wetzlar, Germany). All samples from mice, rats, and pigs came from already published studies and were retrospectively analyzed[16-21]. All animal experiments were approved by the local government authorities: mouse, rats, pigs: Landesamt für Umwelt und Verbraucherschutz Nordrhein Westfalen; marmosets: Institutional animal welfare committee and subsequently by the Lower Saxony State Office for Consumer Protection and Food Safety (LAVES) (reference number 33.19-42502-04-17/2496); bears: bear samples were obtained by hunters during the hunting seasons in Maine. Hunters were asked to participate on a voluntary base and no bears were killed for the specific purpose of this study. All methods were carried out in accordance with relevant guidelines and regulations).

Mouse models

We re-analyzed healthy male 10-12 week old C57BL/6N mice (n=41) and five widely used murine models of kidney diseases with different etiologies, i.e. unilateral ureteral obstruction (UUO, n=15)[16, 17], adenine-induced nephropathy (adenine, n=15)[18], *Col4a3* knock out (Alport, n=15)[16], unilateral ischemia-reperfusion injury (IRI, n=15)[16, 17], and nephrotoxic serum nephritis (NTN, n=15)[19] as well as an additional sixth model used only for testing, the diabetic/metabolic nephropathy (db/db, n=3)[20]. The surgical UUO and IRI models were conducted in male 10-12 week old C57BL/6N mice as previously described[16, 17]. An additional UUO day 10 cohort of three male C57BL/6J mice was contributed by R. Kramann ~~aK~~ and S. Menzel used as an external control cohort. For the adenine model, male 10-12 weeks old mice on C57BL/6N background were fed with 0.2% adenine-enriched diet as previously described[18]. For the NTN model, kidneys from male 12-14 weeks old 129X1/SvJ mice were harvested 10 days after i.v. injection of a sheep-anti-mouse glomerulus antiserum[19]. *Col4a3* knockout mice were bred on a 129X1/SvJ genetic background and sacrificed at eight weeks of age. The db/db mice (BKS.Cg-*Dock7*^{m}+/+Lepr^{db}/J) were fed a high-fat Western diet for 9 weeks and a normal diet for another 5 weeks before sacrifice[20].

In the UUO (sham, day 5, day 10 samples), IRI (sham, day 14, day 21 samples) and adenine model (day 1, day 14, day 21 samples), additional immunostainings and quantifications were performed as previously described[17, 18] for comparison with network-based automated segmentation results from PAS stainings. In short, sections were deparaffinized and endogenous peroxidase was blocked with 3% H_2O_2. Slides were incubated with a primary antibody against α-SMA (α-smooth muscle actin; Dako/Agilent, M085101-2, Santa Clara, CA) followed by colorimetric detection using DAB and nuclear counterstain with methyl green. The stainings were digitalized and further processed using the viewing software NDP.view (Hamamatsu Photonics, Hamamatsu, Japan). The percentage of positively stained area was analyzed in whole cortices at 20x magnification using ImageJ software by measuring DAB positive pixels in 8-Bit images (National Institutes of Health, Bethesda, MD) as previously described[16, 18]. All analyses were performed in a blinded manner.

Patient samples

~~Twelve~~ Sixteen PAS stained sections from formalin-fixed and paraffin-embedded human kidney specimens (~~eight~~ nine tumor-nephrectomies and ~~four~~ seven biopsies (two minimal change disease, one pauci-immune glomerulonephritis, four acute tubular injury)) were anonymously obtained from the archive of Institute for Pathology of the RWTH Aachen University. In the case of tumor nephrectomies, healthy tissue far away from the tumors was used. Patient characteristics were: M:F = 7:9, age = 63.13±11.86 years. The study was approved by the local ethical committee of the RWTH University (No. EK315/19).

Further species

For an extended analysis across different species, we used healthy kidney tissue from rats, pigs, common marmosets and black bears. We used renal tissue from male Wistar rats (n=8) and German landrace pigs (n=6). Renal tissue from male (n=2) and female (n=6) common marmosets was provided by the German Primate Center, Goettingen. Kidney tissue from black bears (n=8) was provided by ~~SMS and Ro~~the Jackson Laboratory and collected by local huntsmen from male animals at different ages all across Maine, US. Hunters were provided with detailed collection directions and provided datasheets voluntarily about deviations to requested timing in sample collection and fixation as well as metadata about the bears.

Data set and ground truth

All technical terms used in the following sections are described in a glossary in Supp. Table 1. The Whole slide images (WSI, n = 1684 in total) were split into training, validation and test sets as follows: the 41 healthy mouse WSI - 30 training, three validation, eight test, the 15 WSI from each mouse model - 11 training, one validation, three test, the three db/db and three external UUO were only used for the test, the six pig WSI - five training, one test, and the eight marmosets, bears and rats WSI – each split to five training, three test, the 16 human WSI - ten training, six test slides: two test

WSI for performance quantifications and all four slides of acute tubular injury to visually show transferability to human disease.

Ground truth annotations were generated for patches of size 174 x 174 μm^2 (resampled into 516 x 516 pixels integer label images) by eight qualified annotators as outlined in Section “Data quality and quantity” using QuPath[22]. All annotations were corrected by a nephropathologist and researcher with long experience in nephrological basic research. Six predefined classes (i.e. renal structures) were annotated: 1) full glomerulus, 2) glomerular tuft, 3) tubule, 4) artery, 5) arterial lumen, 6) vein including renal pelvis and large non-tissue areas. Classes and annotation procedure are defined in detail in Supp. Table 2 and Supp. Fig. 1A-G. The remaining tissue comprising capillaries, adventitia of arteries, interstitial cells and matrix, and urothelium, was defined as the “interstitium”. For annotations, we mostly selected 20 random patches per slide. An overview of our annotations is provided in Supp. Table 3.~~per slide for mice and humans and ten for the remaining species, overall~~ In total, we performed 2,930 annotated patches and 72,722 annotated structures and split the annotated patches into 2,100 training (600 murine healthy, 220 each murine model, 200 human, 50 each remaining species), 160 validation (60 murine healthy, 20 each murine model) and 670 test patches (160 murine healthy, 60 each murine model, 30 murine db/db, 30 external murine UUO, 30 each remaining species including human) for the development of our CNN (Supp. Table 4, Fig 1).~~resulting in 2,7202,930 annotated patches and 68,52372,722 annotated structures (Supp. Table 2, Fig. 1).~~

Data quality and quantity

The most crucial prerequisite for high-performance of a deep learning system is the optimization of data quality and quantity. We performed the following optimization techniques: 1) the expert annotators were instructed and coached to precisely comply with the developed structure definitions (Supp. Table 2~~1~~ and Supp. Fig. 1) to reduce inter-annotator variability, thus yielding consistent annotations. 2) After manual annotation of about 20% of all annotations, we used these to train an initial segmentation network. We then used its predictions as pre-annotations facilitating the annotation effort for the annotators. These predictions were loaded into QuPath, converting the manual annotation task into a prediction correction task, reducing the annotation effort (Supp. Fig. 1H). ~~This effectively reduced annotation effort from approximately 30 minutes for manual patch annotation to about three to five minutes for patch prediction correction, i.e. a six- to ten-fold increase in effectivity.~~ 3) We applied the concept of active learning[23] to optimize the selection of image patches for annotation. We used the initial segmentation network to compute whole-slide segmentation results and visually selected patches with the highest prediction errors most often showing complex or rare structures. We have repeated step 2) and 3) when about 60% of all annotations have been performed. This concept yields an extremely high degree of sample efficiency to ensure that the network will learn and improve in an optimal way.

CNN development

CNN-Model

Our employed deep learning model was based on the U-Net architecture[24] (for details see Supp. Table 4). The U-Net was initially developed for biomedical image segmentation and represents one of the most popular and powerful segmentation techniques nowadays. We applied the following changes to the original architecture: 1) we increased its depth by one to increase its receptive field, 2) we then used half channel numbers on each architectural level to reduce the risk of overfitting, 3) we did not half feature channel numbers when upsampling *via* transposed convolutions to effectively increase its capacity, and 4) we empirically applied instance normalization as well as leaky ReLU activation due to its empirically shown superiority over batch normalization and ReLU activation[25], overall resulting in about 37 million learnable parameters in our CNN. As network inputs, ~~W~~we extracted bigger image slide patches of 216 x 216 µm^2, resampled into 640 x 640 pixels RGB images, around the annotated patches of 174 x 174 µm^2, to improve prediction accuracy close at borders due to the resulting context-awareness[26].

Border class

To ensure the separation of different, touching instances of the same class, we introduced a new border class following[27] by performing dilation on all tubules using a ball-shaped structuring element of radius three pixels. Considering arteries and glomeruli, only the overlap between their dilated versions, employing a radius of seven pixels, was also assigned to the border class. This way, the network was able to maintain a continuous label transition prediction from afferent and efferent arteriole to the glomerulus, thereby greatly improving the prediction accuracy of small afferent and efferent arterioles. The border class mainly represented the tubular basement membranes.

Training routines

We trained our CNN using the optimizer RAdam[28] on random mini-batches of size six and applied weight decay with a factor of 1E-5 for regularization. We further scheduled the learning rate in a reduce-on-plateau fashion to reduce overfitting as follows: it was initially set to 0.001 and was divided by three when the validation loss had not fallen for 15 epochs. When the learning rate fell below 4E-6, training terminated and the network configuration providing the lowest validation error was chosen as the final model. Also, our data augmentation pipeline consisted of spatial, i.e. affine, piecewise affine, elastic, flipping, 90-degree rotation, and color transformations, i.e. hue and saturation shifting, gamma contrast, normalization, to improve the CNN's generalizability by simulating variance in tissue morphology and staining. The weighted categorical cross-entropy (WCE) and the Dice-loss[29] were applied as equally weighted loss functions measuring the dissimilarity between prediction and ground truth for network optimization. Using WCE, we gave the border class a ten times greater weight than other classes to strongly enforce the separation of different instances from the same class. ~~We chose hyperparameters based on the lowest validation loss.~~ Overall, 3-channel input (RGB) of spatial resolution 640 x 640 pixels were being forwarded through the network producing eight class probability maps, i.e. full glomerulus, glomerular tuft, tubule, vein including non-tissue background and renal pelvis, artery, arterial lumen, tubular border, remaining tissue representing our interstitium class, of spatial size 516 x 516 pixels. For each pixel, the class with the highest probability was assigned as the predicted label. To account for reproducibility, our code is publicly available at (https://github.com/NBouteldja/KidneySegmentation_Histology).

Postprocessing

In contrast to network ensembling, we applied the regularization technique test-time augmentation (TTA) to improve the CNN's robustness at low cost. During inference, TTA forwards flipped versions of the input and averages their respectively back-flipped predictions to reduce prediction variance by considering multiple estimations. We also performed the following postprocessing techniques to all classes except the interstitium: 1) we removed too small instance predictions and assigned them to the remaining interstitium class, except for respective glomerular tuft and arterial lumen predictions that were assigned to their superior classes glomerulus and artery, 2) we performed hole filling, and 3) dilated tubular instance predictions due to their thicker border predictions.

Evaluation

Quantitative evaluation

We quantitatively evaluated network performance using instance-level Dice scores, i.e. in all image/ground truth pairs, we computed regular Dice scores between each ground truth instance and its maximally overlapping prediction (0 for false negatives), and *vice versa* for each prediction instance to also account for false positives. These Dice scores were averaged over all instances in all images, resulting in the instance-level Dice score. This metric accurately denoted the mean detected area coverage per instance. We also employed the commonly used average precision (AP) as a detection metric. After counting and summing all true positives (TP), false positives (FP) and false negatives (FN) across all images, the AP was calculated as follows:

$$AP = \frac{TP}{TP + FP + FN}$$

A prediction was considered a TP when it overlapped with at least 50% of a ground-truth instance. Both metrics range from 0 (maximal discordance: no overlap / TP) to 1 (maximal agreement: perfect overlap / detections).

~~*Semi-quantitative and qualitative evaluation*~~

~~Performance on species other than mice and the external (held-out) dataset (db/db) was assessed as expert agreement. For this purpose, two experts in nephropathology independently assessed the predictions from the network on 30 patches of size 174 x 174 µm^2 per species equally distributed on respective test slides. Segmentations with more than approximately 10% divergence from the original structure were considered false. Incorrectly classified instances were considered false as well. Correctly classified~~

~~predictions with 90% or more overlap with the respective structure were counted as true positives. Finally, mean values from both experts were calculated and normalized to the total number of annotations per class.~~

~~We further evaluated our network's capabilities to generalize using an external UUO cohort from a different laboratory by providing visual segmentation results.~~

Performance vs. amount of training data

A key unresolved issue regarding deep-learning systems is the specification of the minimum amount of training data necessary to reach satisfactory performances for a given task. Therefore, we performed an ablation study on performance differences when training on different training set sizes. In total, we trained another 13 CNNs from scratch using the following training sets: From all 2,100 training patches (representing our full CNN), we removed human patches or other species patches, or using murine patches only and in a stepwise manner removing randomly 9.1% of the patches (i.e. using only 90.9%, 81%...9.1% of the murine patches, but always including patches from healthy and each model).~~, i.e. using all data, removing human data, removing other species data, or using only 90.9% to 9.1% murine data of each model.~~ The validation and test sets as employed for our full CNN always remained the same.

~~All-rounder~~Full CNN vs. specialized single models

We examined the impact on network performance when jointly training on data from different domains, i.e. different species and murine disease models. We compared our full CNN trained on all training data (including murine models and species) with 6 networks, each solely trained and tested on a particular single murine model~~s~~, i.e. healthy, UUO, adenine, Alport, IRI, NTN, to analyze whether the network a) benefits from shared multi-domain information by potentially learning more specialized class features or b) can learn the same domain-specific features maintaining equal segmentation performance or c) whether the heterogeneity of multi-domain information might ~~irritate~~perturb the network resulting in lower prediction accuracies.

State-of-the-art model comparison

We compared our model with its unmodified variant, the vanilla U-Net[26], to explore whether our technical modifications to the standard network architecture had an impact on performance. We also compared our network with the context-encoder network[30], another novel state-of-the-art segmentation network particularly suitable for the segmentation of structures with different sizes that was shown to outperform the vanilla U-Net. For all comparisons, the same train and test-sets were used.

Comparative feature extraction

Based on the CNN segmentation results, we extracted the following histological features from cortical areas: 1) relative proportions of tissue area covered by each class, 2) single class instance sizes (including sizes of Bowman's space by subtracting the glomerular tuft area from each full glomerulus) and 3) tubular diameters. We included all instances independent of the plane they were cut. We used data from four individual mice at each of the following model time points: UUO day 10, adenine day 14, Alport mice at eight weeks of age, IRI day 14, NTN day 10 and randomly chosen healthy mice. In each WSI, we extracted ten cortical patches of size 700 x 700 µm^2 for feature computation. We defined the maximum tubular diameter as the diameter of the largest circle fully fitting inside the tubules, a feature that can represent both tubular dilation and atrophy. Tubular diameter computation was performed by employing the *distance transform* function and extracting its maximum value. For class instance size and tubular diameter computation, only instances fully inside our selected patches were considered.

Correlation with immunohistochemical analysis

Next to qualitative and quantitative performance evaluation, we correlated our results with ~~gold~~ standard morphometric analyses, to assess the capabilities of facilitating relevant histopathological applications. We employed data from the three different murine models UUO, adenine, and IRI. We extracted five cortical patches of size 700 x 700 µm^2 in each WSI and correlated the remaining interstitial area coverage predicted by our automated approach with results from a computer-assisted morphometric analysis of immunohistochemical stainings for α-SMA from the same kidneys, in which big vessels were always excluded [16, 18].

Statistics

To measure the strength of the (linear) correlation between immunohistochemical fibrosis quantifications and network-based interstitial area estimations, we employed the Pearson correlation coefficient (PCC) and the Spearman correlation coefficient (SCC) and computed respective p-values based on the t-distribution. We used ~~T~~t-tests for comparison between CNN, the vanilla U-Net and the context-encoder by comparing respective Dice score distributions of each class across all models, and to ~~pairwisely~~ compare pairwise class instance sizes from healthy and all disease models ($p<0.05$ was considered statistically significant).

Results

Ground truth

For the training and evaluation of our full CNN, we performed ~~68,523~~72,722 annotations of six classes, i.e. renal structures, selected based on the most commonly performed compartment-specific quantifications in animal models: tubule, full glomerulus, glomerular tuft, artery (including intima and media but excluding adventitia), arterial lumen, and vein (including renal pelvis and non-tissue slide background). We used kidneys from murine disease models, different species and humans (Supp. Table 3, Supp. Fig 1, Fig. 1). Inclusion of renal pelvis and large non-tissue areas in the "vein" instead of our "interstitium" class improved predictions of such large white structures due to their great local similarities and was an important prerequisite for more precise quantitative analyses, particularly of the interstitium. We have not distinguished different tubular segments, particularly due to the difficult distinction of injured tubules in the disease models. The tubular class did not include tubular basement membranes, to allow a very specific analysis of tubular cells. Both cortex and medulla were annotated, whereas perirenal tissues were not included. We recognized some obstacles in generating annotations, outlined in detail in Supp. Fig. 2. All annotations were ultimately corrected by two experts in nephropathology and structures that were not feasible to assign to a class based on our class definitions with sufficient certainty and consensus were not included in annotations (altogether representing only very few instances).

Accurate multiclass segmentation of murine kidney sections

While network training took about 8.5 hours on the graphics processing unit (GPU) RTX2080Ti and required approximately 10 GB of GPU memory, automated segmentation of a whole murine kidney longitudinal cross-section was performed in less than five minutes on the ~~graphics processing unit RTX2080Ti~~same GPU. Qualitative segmentation results of representative WSIs from healthy and diseased kidneys showed high accuracy for all six classes (Fig. 2A-C and Supp. Fig. 3A-C). In a healthy kidney, an accidental scratch was correctly assigned to the vein class including non-tissue areas (Fig. 2A, arrow). In healthy murine kidneys, our CNN was able to detect almost 95% of all tubular structures with an instance segmentation accuracy of 93.2%. Almost all glomeruli were correctly detected and segmented, while detection and segmentation accuracy were lowest for arteries and arterial lumina (Fig. 3A-A'). Segmentation performance in UUO (Fig. 3B-B') and IRI (Fig. 3C-C') were similar to healthy kidneys for tubules, glomeruli and vein classes (all >90%). Alport mice represented the most complex model, with correct segmentation of 91% of all tubules and 95% of all glomeruli, including those with severe and global pathological alterations such as extracapillary proliferates (cellular crescents) or focal segmental glomerulosclerosis (FSGS) (Fig. 3D-D'). Detection and segmentation results for arteries and their lumina were the lowest ranging from 79.1% (segmentation artery in IRI) to 88.1% (segmentation artery in healthy) and from 73.5% (segmentation arterial lumen in IRI) to 81.1% (segmentation arterial lumen in Alport), respectively. The CNN was able to correctly detect and segment disease-specific pathologies, e.g. dilated tubules in UUO (Fig. 3B), atrophic tubules in IRI (Fig. 3C), glomerular crescents and FSGS in Alport mice and NTN (Fig. 3D; Supp. Fig. 4A, arrows), and tubules with renal crystals in the adenine model (Supp. Fig. 4B, arrows). Medullary structures were also accurately segmented in all models (Supp. Fig. 5A-F''). Almost every segmented item, e.g. one tubular cross-section, was recognized as an individual instance despite potentially touching other class instances and could be therefore further analyzed separately on instance level (Supp. Fig. 5A''-F'').

A very small fraction of structures was not correctly detected or not precisely segmented (Supp. Fig. 6). These included glomeruli with a direct connection to the proximal tubule, in which either a part of the glomerulus was identified as tubule or tubular cells are marked as part of the glomerulus (Supp. Fig. 6A-A’, arrow). Those examples also included special instances, e.g. fibrin within crescents (Supp. Fig. 6B-B’, arrow), which was missing in the training data set. We also observed some incorrectly detected tubules, mostly if severely injured, present as denuded basement membrane (Supp. Fig. 6C-C’ arrow), massively dilated (Supp. Fig. 6D-D’, arrowhead) or atrophic (Supp. Fig. 6D-D’ arrow).

Detection rates were improved in all models by providing more training data (Supp. Fig. 7). In all models and almost all classes (except arteries and arterial lumina), approximately 35% of ground truth data was already sufficient to obtain 90% or higher detection rates. Especially for more complex structures such as arteries or very small structures like arterial lumina, detection performance could be substantially improved by integrating more training data, indicating that further improvement of segmentation accuracy for some classes is feasible (Supp. Fig. 7). For other classes, especially tubules, the performance was high and stable even in case of only about 9% training data.

We compared our CNN with its variants, that have been solely trained and tested on single murine models (healthy, UUO, adenine, Alport, IRI, NTN). In almost all models and classes, especially arteries and lumina, our ~~universal~~ full CNN trained on all domains, provided higher segmentation performances compared to the variants (Supp. Fig. 8A-F).

We next compared our CNN with its unmodified variant, the vanilla U-Net, and with a context-encoder, a novel state-of-the-art segmentation framework which was shown to outperform the U-Net[30]. Our modified CNN significantly outperformed the unmodified vanilla U-Net (Supp. Table 5) and the context-encoder (Supp. Table 5) in the majority of classes and models, including arterial structures. Thus, our modified architecture was suitable for the specific task of kidney histology segmentation.

Multiclass segmentation in external UUO test set and held-out db/db model

We next examined performance of our full CNN on PAS slides from an external UUO cohort and also in a completely different disease model, i.e. the db/db mice on a high-fat diet[20], both not included in the training. ~~Semiquantitative~~ Quantitative evaluation ~~according to our expert agreement~~ confirmed very high segmentation accuracies of at least 95% area coverage with the ground truth for glomeruli, tufts, and tubules in both experiments (Table 2, Supp. Fig. 9A-D''). As in other models, the segmentation of arteries and their lumina were less accurate (both approximately 80%, respectively~~59.4% and 84.2%, respectively~~). Overall, these results are comparable to the other models included in training indicating strong generalization capabilities of our CNN across different laboratories and models.

~~We also used PAS slides from an external UUO cohort to estimate the CNN's generalization capabilities. Our CNN correctly segmented the classes in both the cortex and the medulla (Supp. Figure 9C-D''), supporting the applicability on datasets from different laboratories that were not included in the training.~~

Multiclass segmentation of murine kidney sections enables feature extraction and analysis

The CNN based segmentation made it possible to extract quantitative histological features on a large scale. We analyzed each of the six classes in all disease models (Fig. 4A-F), overall analyzing 70,311 cortical instances. We compared healthy kidneys, UUO day 10, adenine day 14, Alport at eight weeks of age, IRI day 14 and NTN day 10. The glomerular area significantly increased in all models, particularly in those with primary glomerular damage, i.e. Alport and NTN. This expansion of glomeruli reached areas of above 14,000 µm^2 in NTN, compared to 6,000 µm^2 as the largest measured glomerular area in healthy mice. We observed similar findings for glomerular tufts, except for Alport mice, in which the tuft size was significantly reduced due to sclerosis (Fig. 4B). Specific analyses of the area of Bowman's space confirmed its expansion in the two models with known glomerular damage, i.e. NTN and Alport. In addition, the Bowman's space was also significantly increased in the Adenine model but decreased in the IRI model (Fig. 4H). Healthy tubules exhibited two major groups with peak areas of 900 µm^2 and 400 µm^2, likely representing different tubular segments. In all disease models, tubular area distributions converged to a single peak at about 400-500 µm^2, in line with tubular damage and simplification. Tubular dilation was found in several disease models, and prominently increased tubular sizes were detected in NTN (maximum tubular size: 20,000 µm^2), Alport (17,000 µm^2) and UUO (15,000 µm^2), compared to healthy (11,000 µm^2) (Fig. 4C).

The maximum cross-sectional area of arteries was not changed while the arterial lumen was slightly reduced in disease models compared to healthy kidneys and significantly decreased in the IRI model (Fig. 4D,E).

The segmentation also allowed us to analyze changes in the relative proportions of tissue area coverage of all classes in all models (Fig. 5A-F). Compared to the interstitial area in healthy kidneys (mean 14%), it increased in all disease models by two- to threefold ((UUO: 38.6%; adenine 26.3%; Alport: 28,7%; IRI: 36.5%; NTN: 23.9%). Conversely, the tubular area decreased in all models by 15-30% (from 78% in healthy to 55.3% - 66.3% in disease). We found no differences in the area occupied by arteries or their lumina.

To analyze tubular changes in more detail, we measured the maximum tubular diameter in cortical tubular cross-sections. This was defined as the diameter of the largest circle completely fitting into a segmentation of a single tubular cross-section (Fig. 6A-A'). In line with ~~the single instance area of tubules~~tubular size (Fig. 4C), diameter distribution in healthy kidneys showed two major groups with approx. 15 and 30 µm diameter, likely representing proximal and distal tubules versus collecting ducts (Fig. 6A). In all disease models, the maximum diameter of tubules was higher than in healthy kidneys (means of healthy: 49 µm, UUO: 56 µm, adenine: 63 µm, Alport: 83 µm, IRI: 56 µm, NTN: 67 µm) (Fig. 6B-G). However, in UUO, IRI, and Alport, the number of small tubules also increased, representing tubular atrophy and being in line with the results of significantly decreased tubular instance sizes (Fig 4C). In the adenine model, the number of medium-sized tubules increased due to intratubular adherent or obstructing crystals. The NTN model contained the most tubules with a maximum diameter of 20 µm.

Segmentation-based feature correlates with ~~gold-~~standard morphometric analyses

Our interstitium class includes several histological compartments, namely the true interstitium, capillaries, and adventitia of arteries. To understand whether this class can still provide useful quantitative information, we compared the interstitial area of the cortex with computer-assisted morphometric analyses of the same kidneys of three selected models. We used immunohistochemical stainings for α-SMA, a widely used marker for the expansion of interstitial myofibroblasts, which is highly upregulated in the UUO, IRI, and adenine model $^{16, 18}$. Representative segmentation showed that compared to healthy kidneys (Fig. 2), the non-classified interstitial areas increased in all renal disease models (Fig. 7A-C). Interstitial area estimated by our CNN strongly correlated with the expression of the myofibroblast marker α-SMA in all models (Fig. 7A',B',C').

Translation of multiclass segmentation to kidneys from different species and humans

To show the broader applicability of our CNN, we applied it to kidneys of other species, including rats, pigs, black bears, and marmosets. With only a few additional training sets per species, i.e. 50 annotated patches each, the CNN was able to detect and segment all classes in the cortex (Fig. 8A-D'') and medulla (Supp. Fig. 10A-D'') in all species, overall providing very high detection and segmentation accuracies of all classes (Table 2). ~~Tubules and glomeruli were detected most often in all species (Table 2). Considering all classes, detection accuracy was similarly high in rats, pigs, and bears but was lower in marmoset kidneys (Table 2).~~

Finally, we tested the CNN on normal human renal biopsies and nephrectomy samples.~~from both human biopsies and nephrectomies from normal and diseased kidneys.~~ Our full CNN segmented all classes in both cortex and medulla and was applicable to large tissue specimens from nephrectomies and on renal biopsies (Fig. 8E-F'', Supp. Fig. 10E-F''). ~~Semi-q~~Quantitative validation confirmed high detection segmentation accuracies of all classes. However, as compared to other species, ~~with exception of arteries, for which the~~ performance was lower for glomerular tuft, arteries and their lumina (Table 2). As a proof-of-concept we additionally provided visual segmentation results in human biopsies showing acute tubular damage, a feature that is also common in many animal models, yielding promising segmentation results (Supp. Fig. 11).

Discussion

We developed a CNN for automated multiclass segmentation of renal histology of different mammalian species and different experimental disease models with broad pathological alterations. In comparison, the currently available multiclass segmentation model was developed on patients' samples only and focused on transplant specimens [10]. Compared to the previous work [10], we also technically extended the segmentation pipeline by employing suitable task-specific modifications to network architecture, novel approaches for data quality and quantity improvement, modern network training and regularization routines, and network performance quantification based on novel and precise evaluation metrics. As a proof of concept, we used the segmentation results to provide quantitative metrics for efficient, comparative, high-throughput histopathological analyses.

To standardize the annotation procedure, we first developed precise class definitions and performed several training sessions with all expert annotators. This step was also used in difficult radiological segmentation tasks, in which experts underwent a period of training of up to several months, until they had reached a defined reproducibility ensuring sufficient quality of manual annotations[31]. These definitions can also guide future training for further model improvement. The annotation process is highly time-consuming, which is a major limiting factor. In order to facilitate the process, we loaded predictions into QuPath, which served as pre-annotations and reduced manual annotation effort by up to 90%. This made it possible to perform an exceedingly large number of expert-based annotations (~~68,523~~72,722 in total), representing the largest study to date for histopathologic structure segmentation. We also applied active learning for patch selection, i.e. we visually selected patches with the largest prediction errors and corrected them, which further strongly improved the

CNN performance while reducing the number of required annotations as described by others[32]. Besides, QuPath currently represents the most widely used open-source and freely available software for digital pathology, enabling broad, vendor-independent applicability.

We have chosen six different broadly used murine models in nephrology research. The models provide a wide variety of distinct etiologies and histopathological alterations, i.e. obstructive nephropathy, ischemia-reperfusion injury, crystal-induced nephropathy, immune-mediated glomerulonephritis, genetic glomerulopathy, and metabolic (diabetic) nephropathy. Despite the broad differences in histopathology, our CNN was able to segment all structures in all models with high accuracy. Our results suggest that a single comprehensive CNN might perform better compared to specific CNNs trained for each model, and that performance can be further improved by integrating data from different species, including humans. This follows from the partial class similarities across all models and species, effectively yielding more useful training data and thus contributing to learning more generalizable class features.

Only one-third of the training data was sufficient to reach approximately 90% accuracy in all classes, except for arteries and their lumina. For both latter classes, performance improved continuously as training data sets increased, indicating options for further improvements. Due to the amount of training data, strong color augmentations and active learning, our CNN yielded accurate segmentation of an external UUO dataset and db/db mice, a model with distinct pathology the network had never seen before. Our data also showed that it is possible to achieve promising segmentation accuracy in different species or models with rather little additional annotation effort by experts. This might allow rapid adaptation of the algorithm to samples from various laboratories and translation to additional models and pathologies. This is an important prerequisite for high-throughput and reproducible analyses and will be essential to reduce the workload while at the same time increasing the quantitative precision in experimental and potentially also clinical histopathology. As a proof of concept, we applied our model to human biopsies with acute tubular damage with promising segmentation accuracy. However, further studies will be needed to develop a model that is capable of efficiently segmenting the broad spectrum of human renal pathology.

We describe the applicability of implementing basic feature extraction on top of the segmentation results, providing compartment-specific quantifications. Using a handcrafted feature, tubular diameters on an entire slide could be analyzed within minutes, a task that would be impossible to perform manually. Such basic analyses can provide valuable quantitative information about healthy renal morphology, novel insights into experimental disease models and human kidney diseases while saving an enormous amount of time. We found that the mean instance size of glomeruli was increased all our disease models. This was expected for models with primary glomerular damage and crescent formation, i.e. Alport and NTN, which both also exhibited larger Bowman's space, but was surprising for models with primary tubulointerstitial damage. Possible explanations are compensatory glomerular hypertrophy with loss of nephrons and enlargement of Bowman's space due to obstruction of the associated tubule, e.g. in the adenine model and the IRI model. An exception was the Alport model, which exhibited significantly smaller glomerular tuft sizes due to pronounced glomerulosclerosis. For tubules, we found a significant decrease in tubular size in all disease models but at the same time an increase of the maximum tubular instances in UUO, Alport, and NTN. These data provide quantitative evidence for tubular injury and atrophy in all models and model-specific cystic tubular dilation, which was confirmed by the direct analysis of tubular dilation. Overall, these large scale precise quantitative data provide novel read-outs for interventional studies~~,~~ ~~bring new insights into pathological disease mechanisms~~ and potentially also lead to reduced numbers of animals required for research.

Our study has several limitations. First, in our current CNN, the non-segmented area comprises a collection of various histological structures, including peritubular capillaries, interstitium, arterial adventitia, tubular basement membranes, and all other non-recognized structures. Although we found a high correlation with the expression of the fibrosis marker α-SMA, our “interstitial area” does not specifically reflect fibroblasts or fibrosis. Further annotations and training of the specific subclasses, e.g. capillaries, immune cells, adventitia, and tubular basement membranes, will enable us to refine the segmentation. Second, we have not differentiated between the various tubular segments. Although automated differentiation between tubular segments would allow a more comprehensive study of tubular injury, we recognized that manual annotations of tubular segments on PAS stainings were not possible in some disease models with reasonable certainty. An automated differentiation between cortex and medulla could be the first step towards this direction. Third, our study is descriptive and does not allow to draw mechanistic implications. Fourth~~Third~~, human renal diseases show a multitude of different histopathological alterations, some of which, e.g. membranous or membranoproliferative glomerular changes, are not well reflected in our animal models. Further studies, expert annotations, consensus, and technical improvements will be required for a holistic segmentation model that comprehensively covers all (human) renal diseases. Finally, although our network showed promising results on external, held-out data from a different laboratory, multi-center studies will be required to assess the full generalization capability of the network.

In conclusion, our DL algorithm for segmentation of kidney histology for multiple murine disease models and ~~multi-~~species, ~~multi-class segmentation of kidney histology~~ provides a first~~, major~~ step towards fully automated high-throughput quantitative computational experimental nephropathology.

Author contributions

NB, BMK, RDB, DM and PB planned and oversaw the study. NB, BMK and RDB planned and conducted experiments, NB, BMK, RDB, PD, SWO and SVS performed annotations. BMK and RDB corrected annotations. NB performed statistical analyses. SS, RoK, JM, ML, SM, MM, CD, RaK and PB provided samples. NB, BMK and RDB wrote the first draft of the manuscript and arranged figures. JF, PB and DM critically reviewed the manuscript and figures. All authors read and approved the final version of the article.

Acknowledgments

The support for manual annotations from Felicitas Weiß, Timo Horstmann and the whole LaBooratory is gratefully acknowledged.

Funding

This study was funded by the German Research Foundation (DFG; SFB/TRR57, SFB/TRR219, BO3755/3-1, and BO3755/6-1), the German Federal Ministry of Education and Research (BMBF: STOP-FSGS-01GM1901A), the German Federal Ministry of Economic Affairs and Energy (BMWi: EMPAIA project) and the RWTH Aachen Exploratory Research Space (ERS Seed Fund: OPSF585).

Disclosure

The authors declare that there is nothing to disclose.

Supplementary Material

Supp. Table 1. Glossary of technical terms.

Supp. Table 2. Criteria for definition of classes.

Supp. Table 3. Quantitative information on ground truth data.

Supp. Table 4. Architecture of our full CNN.

Supp. Table 5. Performance comparison of our model, its unmodified variant vanilla u-net, and state-of-the-art context-encoder.

Supp. Fig. 1. Annotation procedure

Supp. Fig. 2. Challenging morphology for manual and automated annotations.

Supp. Fig. 3. Segmentation on whole slide images of UUO, Alport and NTN kidneys.

Supp. Fig. 4. Quantitative segmentation performance in murine NTN and adenine kidneys.

Supp. Fig. 5. Automated segmentation in the medulla of murine kidney sections.

Supp. Fig. 6. Examples of missclassifications.

Supp. Fig. 7. Relation between amount of training data and detection performance.

Supp. Fig. 8. Comparison between our full CNN and its variants independently trained on single models ~~the fully trained CNN and its variants~~.

Supp. Fig. 9. Segmentation of non-trained and external murine kidney slides.

Supp. Fig. 10. Automated segmentation of renal medulla in different species.

Supp. Fig. 11. Automated segmentation of human biopsies presenting with acute tubular damage.

References

References

1. Robboy, SJ, Weintraub, S, Horvath, AE, Jensen, BW, Alexander, CB, Fody, EP, Crawford, JM, Clark, JR, Cantor-Weinberg, J, Joshi, MG, Cohen, MB, Prystowsky, MB, Bean, SM, Gupta, S, Powell, SZ, Speights, VO, Jr., Gross, DJ, Black-Schaffer, WS: Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. *Arch Pathol Lab Med,* 137**:** 1723-1732, 2013.
2. LeCun, Y, Bengio, Y, Hinton, G: Deep learning. *Nature,* 521**:** 436-444, 2015.
3. Boor, P: Artificial intelligence in nephropathology. *Nat Rev Nephrol,* 16**:** 4-6, 2020.
4. Sirinukunwattana, K, Raza, SEA, Tsang, YW, Snead, DRJ, Cree, IA, Rajpoot, NM: Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *Ieee Transactions on Medical Imaging,* 35**:** 1196-1206, 2016.
5. Xu, Y, Jia, Z, Wang, LB, Ai, Y, Zhang, F, Lai, M, Chang, EI: Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *Bmc Bioinformatics,* 18**:** 281, 2017.
6. Kather, JN, Pearson, AT, Halama, N, Jager, D, Krause, J, Loosen, SH, Marx, A, Boor, P, Tacke, F, Neumann, UP, Grabsch, HI, Yoshikawa, T, Brenner, H, Chang-Claude, J, Hoffmeister, M, Trautwein, C, Luedde, T: Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine,* 25**:** 1054-+, 2019.
7. Gadermayr, M, Eschweiler, D, Jeevanesan, A, Klinkhammer, BM, Boor, P, Merhof, D: Segmenting renal whole slide images virtually without training data. *Computers in Biology and Medicine,* 90**:** 88-97, 2017.
8. Gadermayr, M, Gupta, L, Appel, V, Boor, P, Klinkhammer, BM, Merhof, D: Generative Adversarial Networks for Facilitating Stain-Independent Supervised and Unsupervised Segmentation: A Study on Kidney Histology. *Ieee Transactions on Medical Imaging,* 38**:** 2293-2302, 2019.
9. Gupta, L, Klinkhammer, BM, Boor, P, Merhof, D, Gadermayr, M, Ieee: STAIN INDEPENDENT SEGMENTATION OF WHOLE SLIDE IMAGES: A CASE STUDY IN RENAL HISTOLOGY. In: *2018 Ieee 15th International Symposium on Biomedical Imaging.* 2018, pp 1360-1364.
10. Sheehan, SM, Korstanje, R: Automatic glomerular identification and quantification of histological phenotypes using image analysis and machine learning. *Am J Physiol Renal Physiol,* 315**:** F1644-F1651, 2018.
11. Hermsen, M, de Bel, T, den Boer, M, Steenbergen, EJ, Kers, J, Florquin, S, Roelofs, J, Stegall, MD, Alexander, MP, Smith, BH, Smeets, B, Hilbrands, LB, van der Laak, J: Deep Learning-Based Histopathologic Assessment of Kidney Tissue. *J Am Soc Nephrol,* 30**:** 1968-1979, 2019.
12. Ginley, B, Lutnick, B, Jen, KY, Fogo, AB, Jain, S, Rosenberg, A, Walavalkar, V, Wilding, G, Tomaszewski, JE, Yacoub, R, Rossi, GM, Sarder, P: Computational Segmentation and Classification of Diabetic Glomerulosclerosis. *J Am Soc Nephrol,* 30**:** 1953-1967, 2019.
13. Bueno, G, Fernandez-Carrobles, MM, Gonzalez-Lopez, L, Deniz, O: Glomerulosclerosis identification in whole slide images using semantic segmentation. *Comput Methods Programs Biomed,* 184**:** 105273, 2020.

14. Bueno, G, Gonzalez-Lopez, L, Garcia-Rojo, M, Laurinavicius, A, Deniz, O: Data for glomeruli characterization in histopathological images. *Data Brief,* 29**:** 105314, 2020.
15. Kannan, S, Morgan, LA, Liang, B, Cheung, MG, Lin, CQ, Mun, D, Nader, RG, Belghasem, ME, Henderson, JM, Francis, JM, Chitalia, VC, Kolachalama, VB: Segmentation of Glomeruli Within Trichrome Images Using Deep Learning. *Kidney Int Rep,* 4**:** 955-962, 2019.
16. Ehling, J, Babickova, J, Gremse, F, Klinkhammer, BM, Baetke, S, Knuechel, R, Kiessling, F, Floege, J, Lammers, T, Boor, P: Quantitative Micro-Computed Tomography Imaging of Vascular Dysfunction in Progressive Kidney Diseases. *J Am Soc Nephrol*, 2015.
17. Djudjaj, S, Papasotiriou, M, Bulow, RD, Wagnerova, A, Lindenmeyer, MT, Cohen, CD, Strnad, P, Goumenos, DS, Floege, J, Boor, P: Keratins are novel markers of renal epithelial cell injury. *Kidney Int,* 89**:** 792-808, 2016.
18. Baues, M, Klinkhammer, BM, Ehling, J, Gremse, F, van Zandvoort, M, Reutelingsperger, CPM, Daniel, C, Amann, K, Babickova, J, Kiessling, F, Floege, J, Lammers, T, Boor, P: A collagen-binding protein enables molecular imaging of kidney fibrosis in vivo. *Kidney Int,* 97**:** 609-614, 2020.
19. Djudjaj, S, Lue, H, Rong, S, Papasotiriou, M, Klinkhammer, BM, Zok, S, Klaener, O, Braun, GS, Lindenmeyer, MT, Cohen, CD, Bucala, R, Tittel, AP, Kurts, C, Moeller, MJ, Floege, J, Ostendorf, T, Bernhagen, J, Boor, P: Macrophage Migration Inhibitory Factor Mediates Proliferative GN via CD74. *J Am Soc Nephrol*, 2015.
20. Moellmann, J, Klinkhammer, BM, Onstein, J, Stohr, R, Jankowski, V, Jankowski, J, Lebherz, C, Tacke, F, Marx, N, Boor, P, Lehrke, M: Glucagon-Like Peptide 1 and Its Cleavage Products Are Renoprotective in Murine Diabetic Nephropathy. *Diabetes,* 67**:** 2410-2419, 2018.
21. Mancina, E, Kalenski, J, Paschenda, P, Beckers, C, Bleilevens, C, Boor, P, Doorschodt, BM, Tolba, RH: Determination of the preferred conditions for the isolated perfusion of porcine kidneys. *Eur Surg Res,* 54**:** 44-54, 2015.
22. Bankhead, P, Loughrey, MB, Fernandez, JA, Dombrowski, Y, McArt, DG, Dunne, PD, McQuaid, S, Gray, RT, Murray, LJ, Coleman, HG, James, JA, Salto-Tellez, M, Hamilton, PW: QuPath: Open source software for digital pathology image analysis. *Sci Rep,* 7**:** 16878, 2017.
23. Settles, B: Active learning literature survey. *University of Wisconsin, Madison,* 52**:** 55-66, 2010.
24. Falk, T, Mai, D, Bensch, R, Cicek, O, Abdulkadir, A, Marrakchi, Y, Bohm, A, Deubner, J, Jackel, Z, Seiwald, K, Dovzhenko, A, Tietz, O, Dal Bosco, C, Walsh, S, Saltukoglu, D, Tay, TL, Prinz, M, Palme, K, Simons, M, Diester, I, Brox, T, Ronneberger, O: U-Net: deep learning for cell counting, detection, and morphometry. *Nat Methods,* 16**:** 67-70, 2019.
25. Isensee, F, Petersen, J, Kohl, SAA, Jäger, PF, Maier-Hein, KH: nnU-Net: Breaking the Spell on Successful Medical Image Segmentation. *arXiv e-prints.* 2019 pp arXiv:1904.08128.
26. Ronneberger, O, Fischer, P, Brox, T: U-Net: Convolutional Networks for Biomedical Image Segmentation. Cham, Springer International Publishing, 2015 pp 234-241.
27. Chen, H, Qi, X, Yu, L, Dou, Q, Qin, J, Heng, P-A: DCAN: Deep contour-aware networks for object instance segmentation from histology images. *Medical Image Analysis,* 36**:** 135-146, 2017.

28. Liu, L, Jiang, H, He, P, Chen, W, Liu, X, Gao, J, Han, J: On the Variance of the Adaptive Learning Rate and Beyond. *arXiv e-prints.* 2019 pp arXiv:1908.03265.
29. Milletari, F: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Forth International Conference on 3D Vision (3DV)***:** 565-571, 2016.
30. Gu, Z, Cheng, J, Fu, H, Zhou, K, Hao, H, Zhao, Y, Zhang, T, Gao, S, Liu, J: CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *IEEE Transactions on Medical Imaging,* 38**:** 2281-2292, 2019.
31. Kennedy, DN, Filipek, PA, Caviness, VS: Anatomic segmentation and volumetric calculations in nuclear magnetic resonance imaging. *IEEE Transactions on Medical Imaging,* 8**:** 1-7, 1989.
32. Lutnick, B, Ginley, B, Govind, D, McGarry, SD, LaViolette, PS, Yacoub, R, Jain, S, Tomaszewski, JE, Jen, KY, Sarder, P: An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nat Mach Intell,* 1**:** 112-119, 2019.

Tables

Table 1. Quantitative segmentation and detection performance of six classes in murine kidneys.

Segmentation performance was calculated by averaging all instance Dice scores from each instance in all test images denoting the mean detected area coverage per instance. We employed average precision metric to measure detection performance.

IRI = ischemia reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction

Mouse models	Detection					
	full glomerulus	glomerular tuft	tubule	artery	arterial lumen	vein
Healthy mouse	98.7	96.5	94.9	87.4	76.2	93.9
UUO	100	100	91.0	78.2	73.3	100
IRI	95.7	97.7	89.3	73.3	67.6	100
Adenine	100	100	93.0	82.4	80.3	90.3
Alport	92.5	93.4	88.6	73.2	79.2	80.0
NTN	96.2	98	93.5	86.1	74.0	89.2
Mouse models	Segmentation					
	full glomerulus	glomerular tuft	tubule	artery	arterial lumen	vein
Healthy mouse	96.5	93.7	93.2	88.1	80.3	94.3
UUO	97.5	95.6	90.9	82.3	75.0	97.6
IRI	96.0	95.4	90.2	79.1	73.5	97.7
Adenine	98.8	97.2	93.0	87.9	80.9	93.5
Alport	94.7	91.4	90.6	80.3	81.1	89.2
NTN	95.5	94.8	93.2	86.8	78.2	92.8

~~Table 2. Semi-quantitative detection performance of six classes in kidneys from different species.~~

~~Average precisions were assessed by expert agreement as described in the "Evaluation" section to measure detection performance.~~

~~Species/Model~~	~~Detection~~					
	~~full glomerulus~~	~~glomerular tuft~~	~~tubule~~	~~artery~~	~~arterial lumen~~	~~vein~~
~~db/db mice~~	~~79.3~~	~~80.8~~	~~90.7~~	~~59.4~~	~~84.2~~	~~100.0~~
~~Rat~~	~~93.8~~	~~69.6~~	~~97.0~~	~~82.8~~	~~91. 9~~	~~84.6~~
~~Pig~~	~~86.7~~	~~93.3~~	~~96.2~~	~~84.2~~	~~90.0~~	~~73.1~~
~~Black bear~~	~~95.0~~	~~87.5~~	~~95.5~~	~~75.0~~	~~85.7~~	~~93.3~~
~~Marmoset~~	~~86.7~~	~~66.7~~	~~90.7~~	~~55.4~~	~~75.0~~	~~96.2~~
~~Human~~	~~100.0~~	~~81.8~~	~~90.0~~	~~33.3~~	~~92.3~~	~~83.3~~

Table 2. Quantitative segmentation and detection performance in kidneys from different species, held-out murine disease model db/db, and external UUO.

Segmentation performance was calculated by averaging all instance Dice scores from each instance in all test images denoting the mean detected area coverage per instance. We employed an average precision metric to measure detection performance.

	Detection					
	full glomerulus	glomerular tuft	tubule	artery	arterial lumen	vein
Rat	100	82.1	94.7	85.7	81.0	92.9
Pig	93.8	100	95.6	100	95.2	84.6
Black bear	88.3	85.7	96.8	94.3	89.2	100
Marmoset	100	100	95.1	82.7	73.5	92.9
Human	88.2	72.5	91.8	66.7	68.4	72.7
db/db mice	93.1	96.3	90.5	60.6	58.3	100
External UUO	93.6	97.7	94.8	68.2	69.6	87.5

	Segmentation					
	full glomerulus	glomerular tuft	tubule	artery	arterial lumen	vein
Rat	99.5	88.9	96.5	91.6	89.5	93.9
Pig	96.5	99.0	97.9	96.9	96.3	91.6
Black bear	87.5	91.5	97.3	91.8	94.3	99.7
Marmoset	98.9	95.9	96.8	86.0	86.8	96.2
Human	93.4	76.6	95.2	79.1	77.6	85.1
db/db mice	95.9	97.5	94.9	81.0	79.1	99.0
External UUO	96.6	98.5	97.0	78.2	81.4	93.3

Figure legends

Figure 1. Overview of experimental design.

Our deep learning model (here: Full CNN) was trained with annotations from healthy and diseased murine kidneys and with annotations from five different species including humans. 72,722~~68,523~~ single instance annotations comprised six different renal structures: “tubule”, “full glomerulus”, “glomerular tuft”, “artery”, “arterial lumen” and “vein”. The model was tested on healthy and diseased murine kidneys, on five different other species, on a held-out murine disease model, and an external UUO cohort. ~~Finally, w~~We used the automatically segmented kidneys to perform quantitative feature analysis~~, e.g. instance size distributions~~ and correlations with IHC. Further experiments included an ablation study on varying training dataset sizes to analyze its impact on model performance, and we also compared the full CNN with its variants solely trained on single murine models as well as with different state-of-the-art segmentation networks including the vanilla U-net and context-encoder networks.

H = Human, IHC = immunohistochemistry, IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, P = Patch, UUO = unilateral ureteral obstruction.

Figure 2. Automated segmentation on whole slide images of murine kidneys.

The CNN generates segmentation predictions on a whole slide image (WSI) of a healthy mouse kidney (A). All six classes, i.e. tubule, glomerulus, glomerular tuft, artery, arterial lumen, and vein are precisely segmented. Even tissue damage in the form of an artificial scratch (arrow) is correctly assigned to the vein class including the background. Similar segmentation predictions are generated for WSIs of IRI (ischemia-reperfusion injury (B) and adenine (C) kidneys.

Figure 3. Quantitative segmentation performance in murine kidney disease models.

Representative PAS pictures and corresponding segmentation predictions generated by the CNN for murine healthy (A), UUO (B), IRI (C) and Alport (D) kidneys. Instance segmentation accuracy is shown by instance-Dice scores for each class in all four models (A'-D').

Data are presented in box plots with median, quartiles, and whiskers. Glom = Glomerulus, IRI = ischemia-reperfusion injury, Tuft = Glomerular tuft, UUO = unilateral ureteral obstruction.

Figure 4. ~~Single Instance class areas~~Instance sizes of each class.

Violine plots show the distribution pattern of ~~instanced areas~~cross-sectional instance sizes for each of the six automatically segmented classes: full glomerulus (A), glomerular tuft (B), tubule (C), artery (D), arterial lumen (E), vein (F) in healthy, UUO, IRI, adenine, Alport and NTN kidneys. In addition, we subtracted the glomerular tuft area from each glomerulus (G) to analyze size distribution of Bowman's space (H).

* = $p < 0.05$ vs. healthy. IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.

Figure 5. Relative area distributions of automatically segmented classes.

The relative area distributions in percent in healthy (A), UUO (B), IRI (C), adenine (D), Alport (E) and NTN (F) kidneys additionally give information on the proportion of remaining non-classified tubulointerstitial area (shown in black).

IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.

Figure 6. Quantitative analysis of tubular dilation.

An exemplary illustration of automated analysis of tubular dilation in PAS stainings of healthy (A) and UUO (A') mouse kidney (top). The maximum tubular diameter is defined as the diameter of the maximum sized circle that fits into a tubule segmentation. Violine plots show the distribution of the analyzed tubular diameter within each model, i.e. for healthy (B), UUO (C), IRI (D), adenine (E), Alport mice (F) and NTN (G).

IRI = ischemia-reperfusion injury, N=Number of analyzed tubule-instances, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.

Figure 7. Correlation between segmentation and standard computer-assisted morphometric analyses.

(A) Representative picture of the automated segmentation prediction in a murine UUO kidney section. The non-classified remaining tissue (black) correlates with α-SMA$^+$ area (A') quantified in immunostainings of the same kidneys. (B) Representative picture of the automated segmentation prediction on a murine IRI kidney section. The non-classified remaining tissue (black) correlates with α-SMA$^+$ area (B') quantified in immunostainings from the same kidneys. (C) Representative picture of the automated segmentation prediction on a murine adenine kidney section. The non-classified remaining tissue (black) correlates with α-SMA$^+$ area (C') quantified in immunostainings from the same kidneys.

IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, PCC = Pearsons correlation coefficient, SCC = Spearmans correlation coefficient, UUO = unilateral ureteral obstruction.

Figure 8. Automated segmentation of kidneys from various species.

Representative pictures illustrate the segmentation quality of the CNN in kidney tissue from rat (A-A''), pig (B-B''), black bear (C-C'') and marmoset (D-D''). Predictions (A', B', C', D') depict different classes, while A''-D'' display predictions on instance level for tubules. All classes are also correctly detected and segmented on human nephrectomy (E-E'') as well as smaller human biopsy (F-F'').

Figure 1

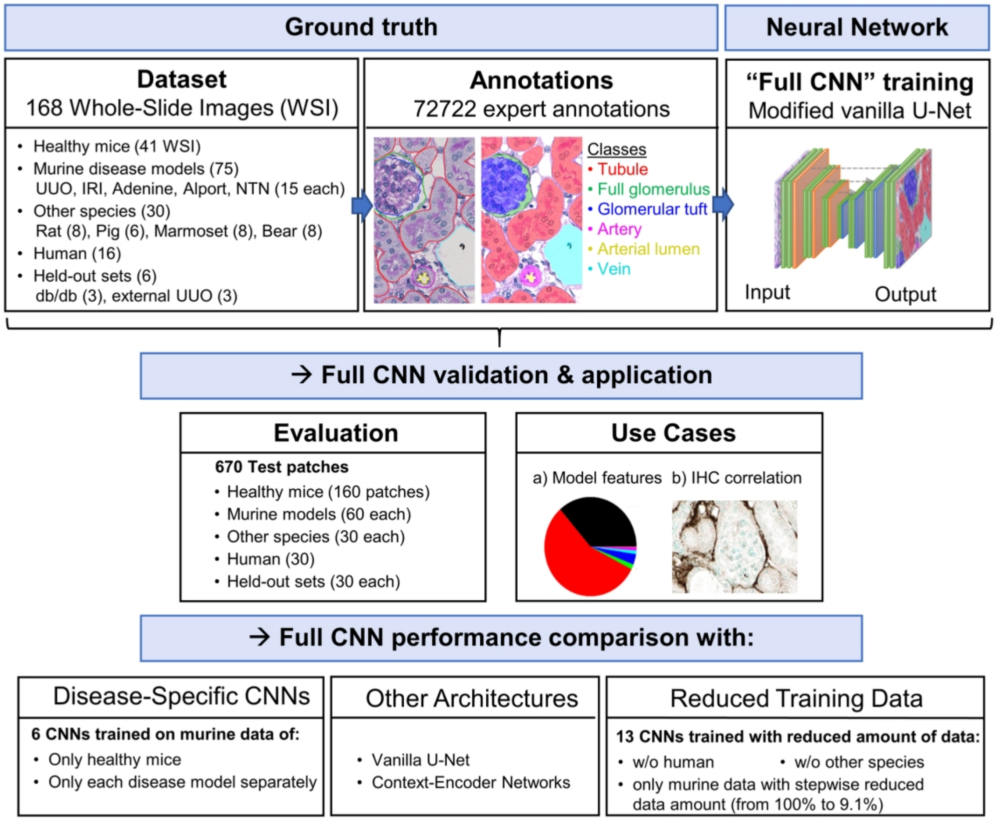

Ground truth
Neural Network
Dataset
168 Whole-Slide Images (WSI)
• Healthy mice (41 WSI)
• Murine disease models (75)
UUO, IRI, Adenine, Alport, NTN (15 each)
• Other species (30)
Rat (8), Pig (6), Marmoset (8), Bear (8)
• Human (16)
• Held-out sets (6)
db/db (3), external UUO (3)
Annotations
72722 expert annotations
Classes
• Tubule
• Full glomerulus
• Glomerular tuft
• Artery
• Arterial lumen
• Vein
"Full CNN" training
Modified vanilla U-Net
Input
Output
→ Full CNN validation & application
Evaluation
670 Test patches
• Healthy mice (160 patches)
• Murine models (60 each)
• Other species (30 each)
• Human (30)
• Held-out sets (30 each)
Use Cases
a) Model features
b) IHC correlation
→ Full CNN performance comparison with:
Disease-Specific CNNs
6 CNNs trained on murine data of:
• Only healthy mice
• Only each disease model separately
Other Architectures
• Vanilla U-Net
• Context-Encoder Networks
Reduced Training Data
13 CNNs trained with reduced amount of data:
• w/o human
• w/o other species
• only murine data with stepwise reduced data amount (from 100% to 9.1%)

Figure 1. Overview of experimental design.
Our deep learning model (here: Full CNN) was trained with annotations from healthy and diseased murine kidneys and with annotations from five different species including humans. 72,722 single instance annotations comprised six different renal structures: "tubule", "full glomerulus", "glomerular tuft", "artery", "arterial lumen" and "vein". The model was tested on healthy and diseased murine kidneys, on five different other species, on a held-out murine disease model, and an external UUO cohort. We used the automatically segmented kidneys to perform quantitative feature analysis and correlations with IHC. Further experiments included an ablation study on varying training dataset sizes to analyze its impact on model performance, and we also compared the full CNN with its variants solely trained on single murine models as well as with different state-of-the-art segmentation networks including the vanilla U-net and context-encoder networks. H = Human, IHC = immunohistochemistry, IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, P = Patch, UUO = unilateral ureteral obstruction.

169x149mm (300 x 300 DPI)

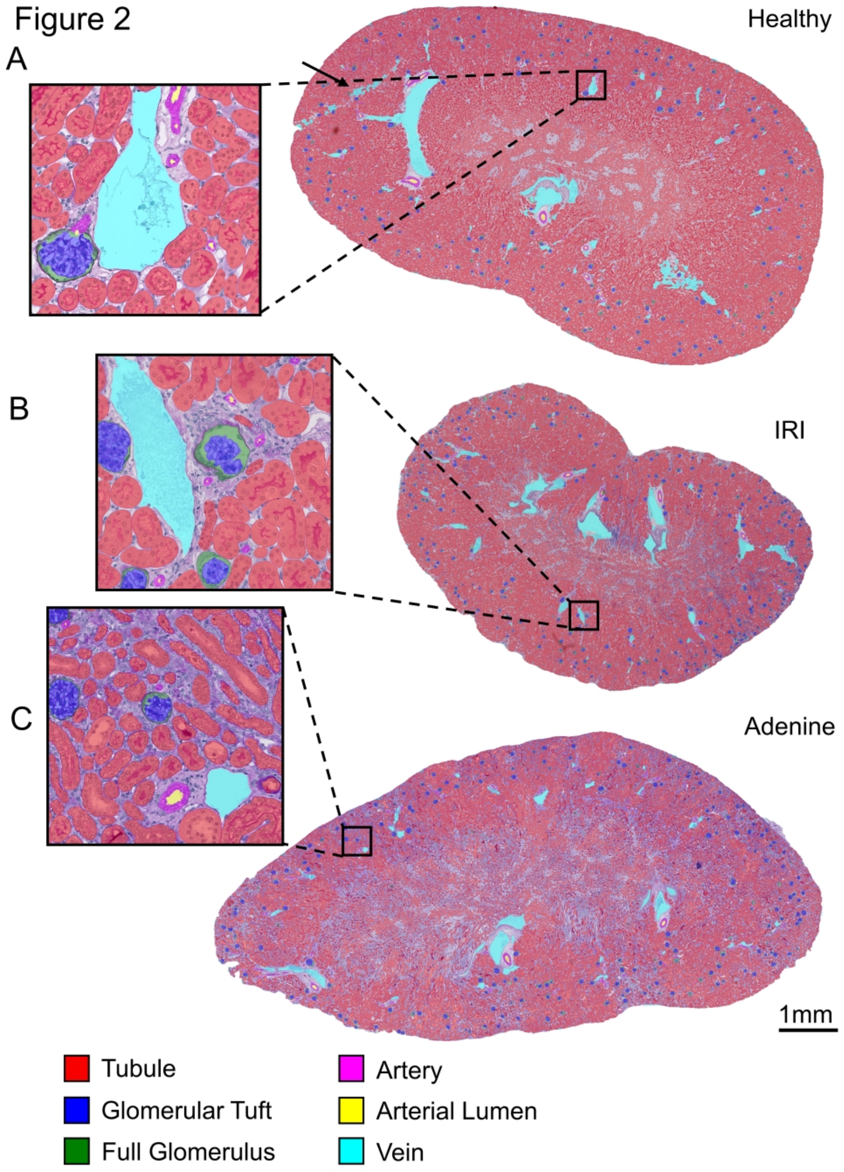


Figure 2. Automated segmentation on whole slide images of murine kidneys. The CNN generates segmentation predictions on a whole slide image (WSI) of a healthy mouse kidney (A). All six classes, i.e. tubule, glomerulus, glomerular tuft, artery, arterial lumen, and vein are precisely segmented. Even tissue damage in the form of an artificial scratch (arrow) is correctly assigned to the vein class including the background. Similar segmentation predictions are generated for WSIs of IRI (ischemia-reperfusion injury (B) and adenine (C) kidneys.

170x237mm (600 x 600 DPI)

Figure 3

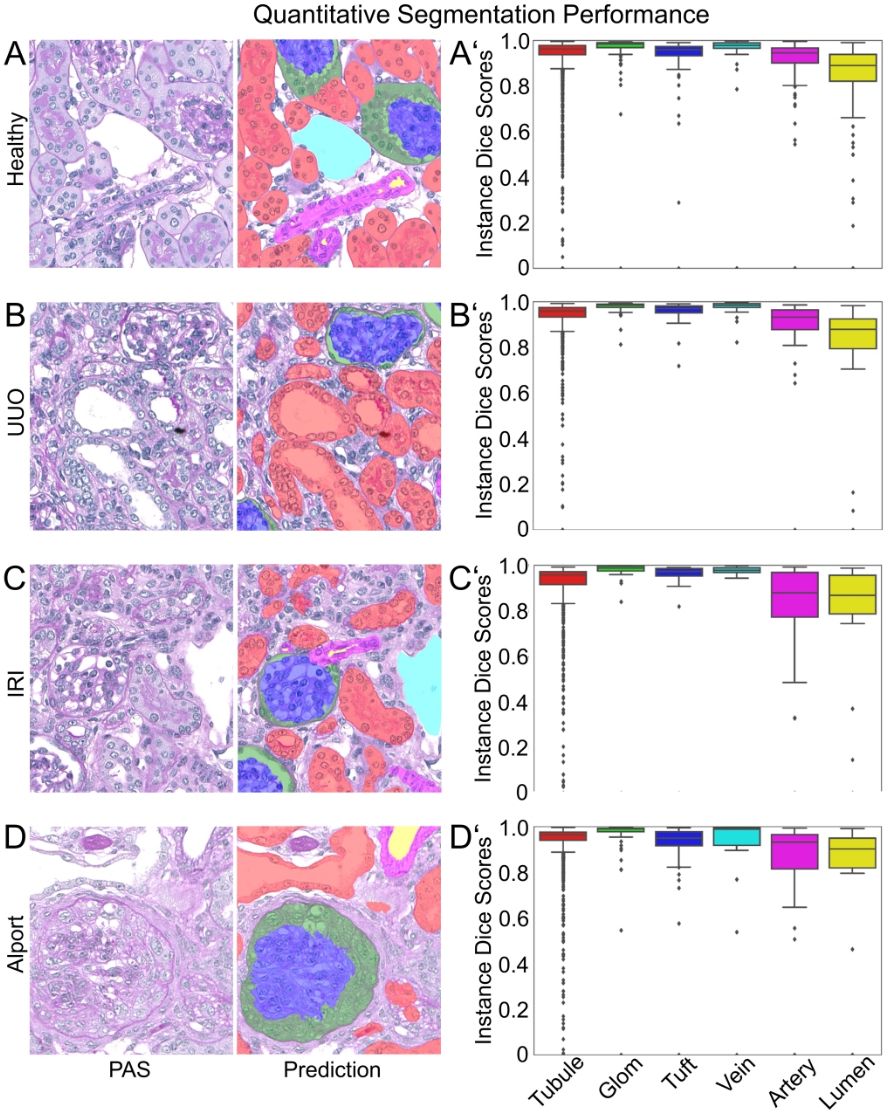


Figure 3. Quantitative segmentation performance in murine kidney disease models. Representative PAS pictures and corresponding segmentation predictions generated by the CNN for murine healthy (A), UUO (B), IRI (C) and Alport (D) kidneys. Instance segmentation accuracy is shown by instance-Dice scores for each class in all four models (A’-D’).

Data are presented in box plots with median, quartiles, and whiskers. Glom = Glomerulus, IRI = ischemia-reperfusion injury, Tuft = Glomerular tuft, UUO = unilateral ureteral obstruction.

170x229mm (600 x 600 DPI)

Figure 4

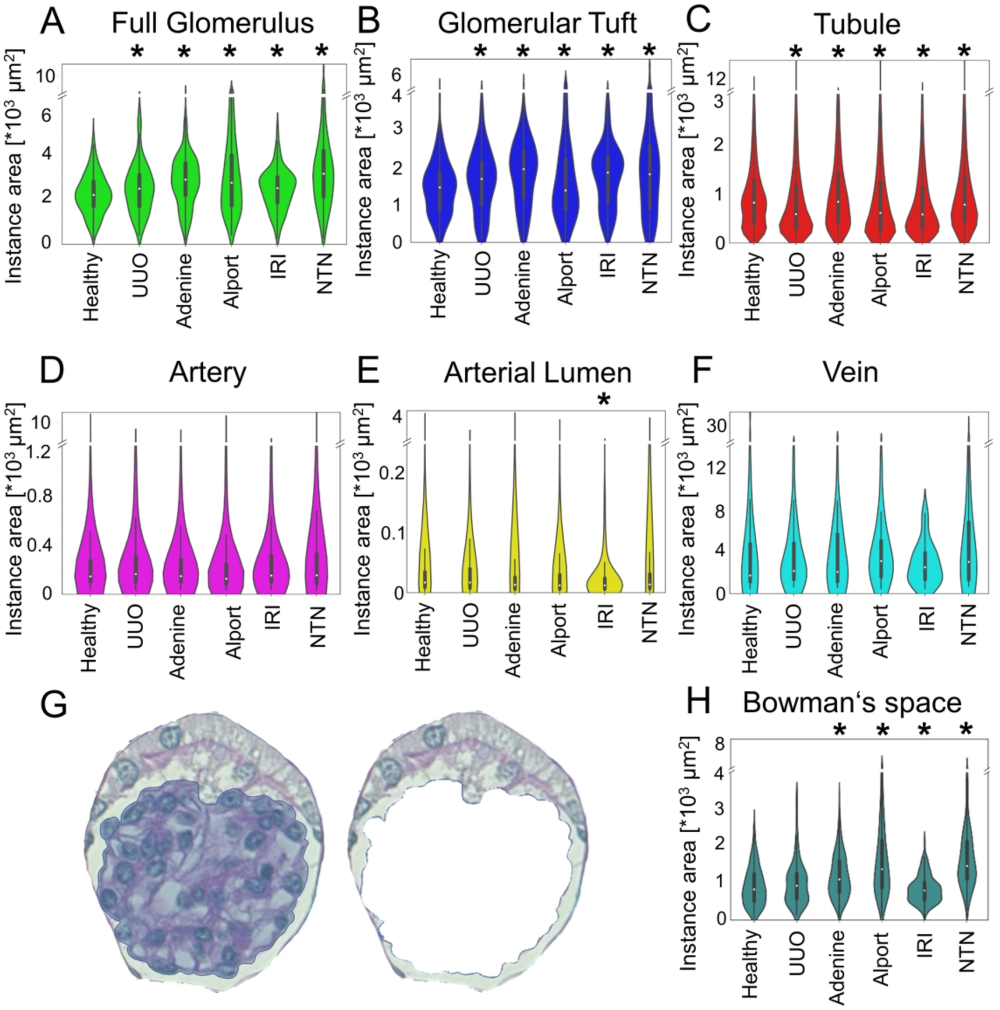


Figure 4. Instance sizes of each class.
Violine plots show the distribution pattern of cross-sectional instance sizes for each of the six automatically segmented classes: full glomerulus (A), glomerular tuft (B), tubule (C), artery (D), arterial lumen (E), vein (F) in healthy, UUO, IRI, adenine, Alport and NTN kidneys. In addition, we subtracted the glomerular tuft area from each glomerulus (G) to analyze size distribution of Bowman's space (H).
* = $p < 0.05$ vs. healthy. IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.

170x180mm (300 x 300 DPI)

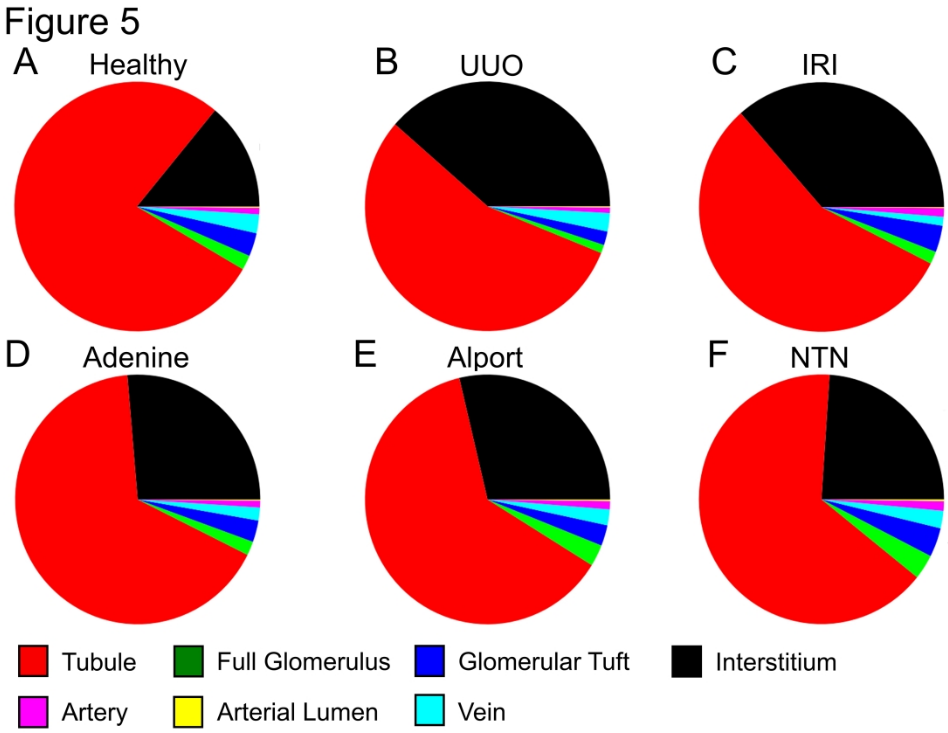


Figure 5. Relative area distributions of automatically segmented classes.
The relative area distributions in percent in healthy (A), UUO (B), IRI (C), adenine (D), Alport (E) and NTN (F) kidneys additionally give information on the proportion of remaining non-classified tubulointerstitial area (shown in black).
IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.

170x130mm (600 x 600 DPI)

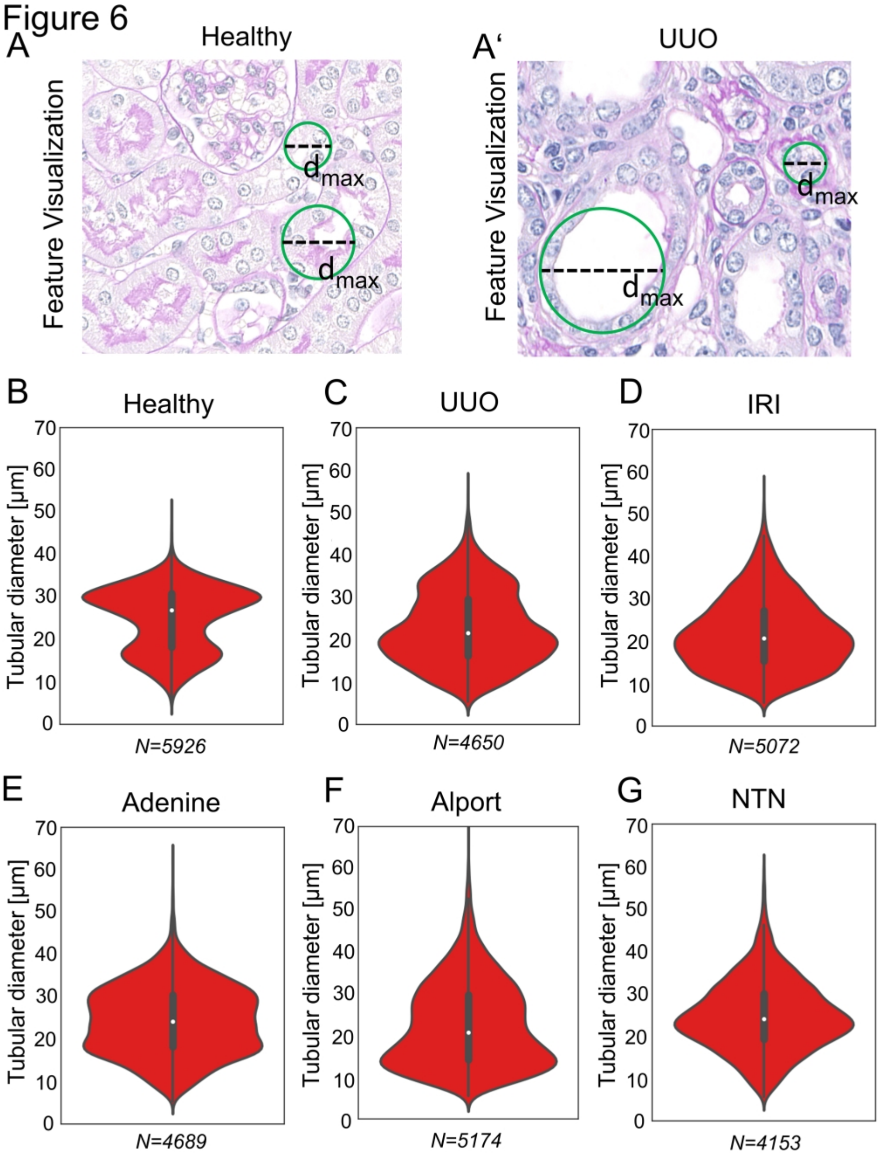


Figure 6. Quantitative analysis of tubular dilation.
An exemplary illustration of automated analysis of tubular dilation in PAS stainings of healthy (A) and UUO (A') mouse kidney (top). The maximum tubular diameter is defined as the diameter of the maximum sized circle that fits into a tubule segmentation. Violine plots show the distribution of the analyzed tubular diameter within each model, i.e. for healthy (B), UUO (C), IRI (D), adenine (E), Alport mice (F) and NTN (G).
IRI = ischemia-reperfusion injury, N=Number of analyzed tubule-instances, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.

170x220mm (600 x 600 DPI)

Figure 7

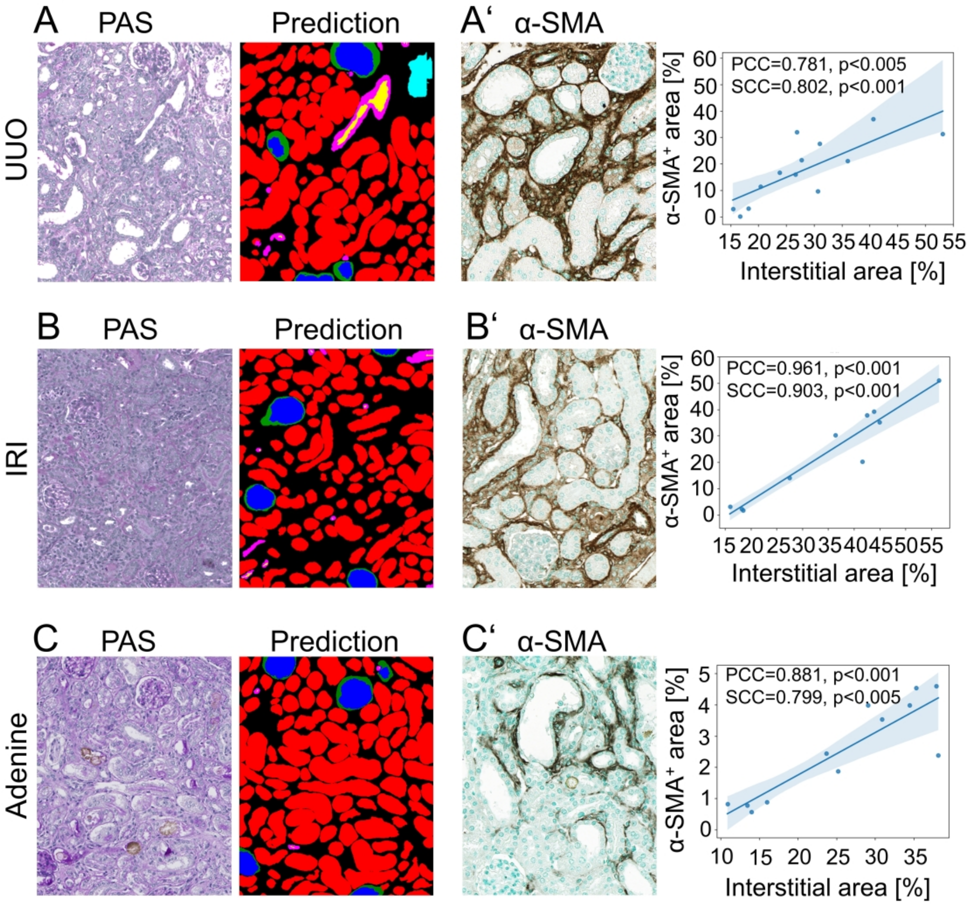


Figure 7. Correlation between segmentation and standard computer-assisted morphometric analyses. (A) Representative picture of the automated segmentation prediction in a murine UUO kidney section. The non-classified remaining tissue (black) correlates with α-SMA+ area (A’) quantified in immunostainings of the same kidneys. (B) Representative picture of the automated segmentation prediction on a murine IRI kidney section. The non-classified remaining tissue (black) correlates with α-SMA+ area (B’) quantified in immunostainings from the same kidneys. (C) Representative picture of the automated segmentation prediction on a murine adenine kidney section. The non-classified remaining tissue (black) correlates with α-SMA+ area (C’) quantified in immunostainings from the same kidneys.
IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, PCC = Pearsons correlation coefficient, SCC = Spearmans correlation coefficient, UUO = unilateral ureteral obstruction.

170x170mm (600 x 600 DPI)

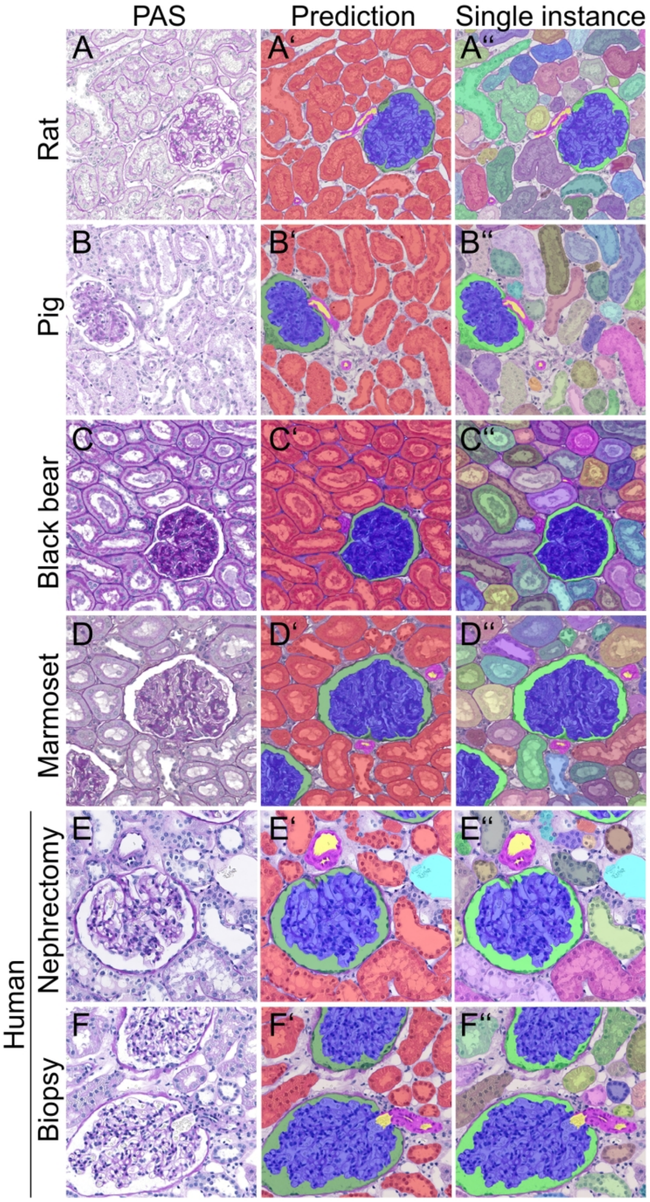


Figure 8

Figure 8. Automated segmentation of kidneys from various species.
Representative pictures illustrate the segmentation quality of the CNN in kidney tissue from rat (A-A''), pig (B-B''), black bear (C-C'') and marmoset (D-D''). Predictions (A', B', C', D') depict different classes, while A''-D'' display predictions on instance level for tubules. All classes are also correctly detected and segmented on human nephrectomy (E-E'') as well as smaller human biopsy (F-F'').

170x240mm (300 x 300 DPI)

Deep-Learning based ~~multi-disease, multi-species, multi-class~~ segmentation and quantification in experimental kidney histopathology

Running Title: DL in experimental nephropathology

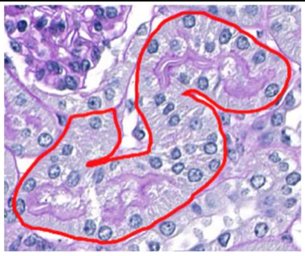
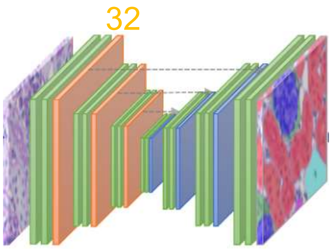
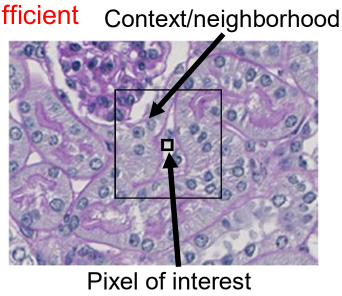
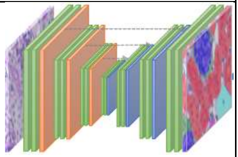
Nassim Bouteldja[2,*], Barbara M. Klinkhammer[1,3,*], Roman D. Bülow[1,*], Patrick Droste[1], Simon W. Otten[1], Saskia von Stillfried[1], Julia Moellmann[4], Susan M. Sheehan[5], Ron Korstanje[5], Sylvia Menzel[3], Peter Bankhead[6,7], Matthias Mietsch[8], Charis Drummer[9], Michael Lehrke[4], Rafael Kramann[3,10], Jürgen Floege[3], Peter Boor[1,3,*,#], Dorit Merhof[2,11,*]

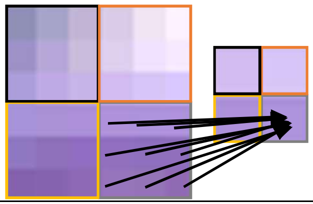
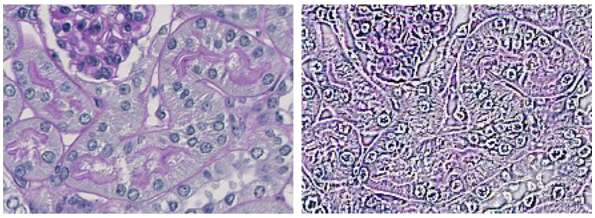
1 Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany
2 Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany
3 Department of Nephrology and Immunology, RWTH Aachen University Hospital, Aachen, Germany
4 Department of Cardiology and Vascular Medicine, RWTH Aachen University Hospital, Aachen, Germany
5 The Jackson Laboratory, Bar Harbor, Maine
6 Edinburgh Pathology, University of Edinburgh, Edinburgh, UK
7 Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
8 Laboratory Animal Science Unit, German Primate Center, Goettingen, Germany
9 Platform Degenerative Diseases, German Primate Center, Goettingen, Germany
10 Department of Internal Medicine, Nephrology and Transplantation, Erasmus Medical Center, Rotterdam, The Netherlands
11 Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

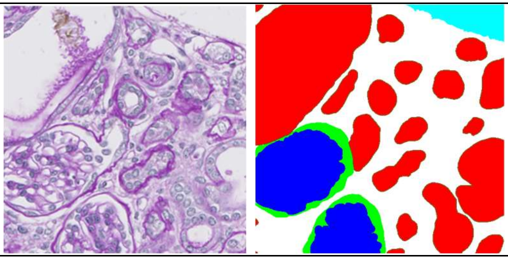
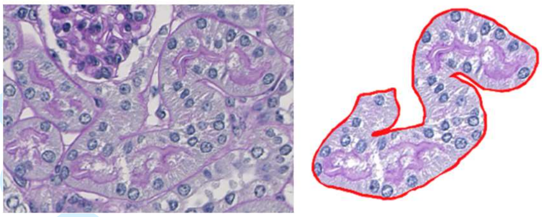
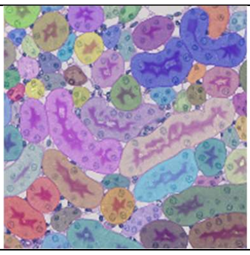
Supplementary material

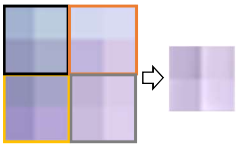
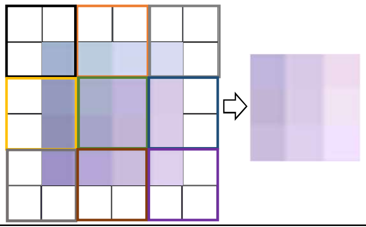
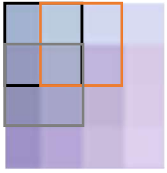
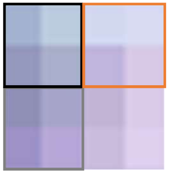
Content

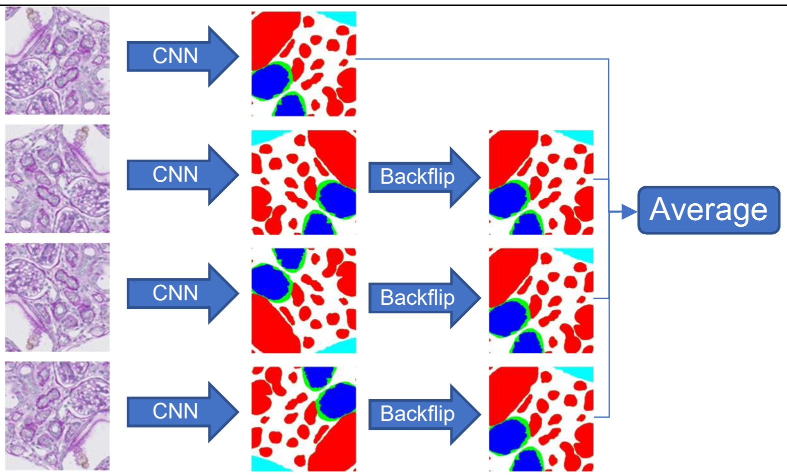
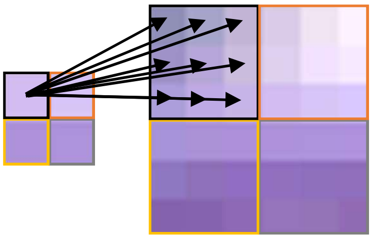
Supplementary Table 1. Glossary of technical terms.

Term	Description
Ablation study	Experiment with consecutively reduced input data. In more detail: A procedure where certain configurations of neural network architecture or training including modifications to data sets are changed to gain a better understanding of their importance and impact (mainly on overall performance).
Border class	 ->***Class*** comprising borders of structures. Example: The tubule's border marked in red is assigned to the border class. In more detail: Artificial class representing the border of specific structures. In our application, we make use of a border class, that especially represents the tubular basement membrane, to separate tubular (as well as glomerular or arterial) instances from each other, allowing for instance-level analysis.
Capacity	Amount of ->***parameters*** in a neural network. In more detail: A neural network consists of many trainable parameters. Its number represents the network's capacity. It is also associated with its complexity, i.e. the degree of complexity of patterns the model is able to learn. Note that a neural network represents a mathematical function including input variables and parameters. Thus, the parameters are here defined in a mathematical way.
Channel numbers	 Number of ->***feature maps***. Example: The channel number of the first, orange ->***convolutional layer*** is 32. In more detail: In convolutional neural networks, input data is subsequently propagated through ->***convolutional layers*** each producing multiple output ->***feature maps***. Their number represents the channel number of the layer.
Class	A group of structures. Example: All tubular structures belong to the "tubule"-class.
Context-awareness	 Ability of a method to incorporate sufficient spatial neighborhood information for the assessment / prediction of a pixel. In more detail: The more spatial context is considered for pixel prediction, the more context-aware is a technique. In our case, our network provides sufficient spatial context even for pixel prediction at patch border.
Convolutional layer	 Network layer performing convolutions to its input. Example: All green blocks represent such layers. In more detail: Such layers represent substantial components in CNNs. Convolutions are performed on input data resulting in multiple ->***feature maps***. Convolutions are mainly specified based on the following ->***parameters***:

	->***kernel size***, ->***stride*** and ->***padding***. As exemplary shown on the right, a convolution (with 3x3 kernel size) slides over the image and outputs a single value for each 3x3 region. 
Cross-entropy loss	Information-theoretical measure of the dissimilarity between network output and ->***ground truth***. In more detail: A commonly used ->***loss function*** when training segmentation or classification networks. The Cross-entropy loss (CE) is based on information theory and measures the difference between a target probability distribution (represented by ground truth annotations) and an estimated one (represented by model predictions). Its values range between 0 and 1. The smaller the loss, the higher the similarity. Thus, a perfect overlap results in a value of zero.
Dice loss / Dice score	The Dice score measures the similarity between network prediction and ->***ground truth*** based on their spatial overlap. In more detail: The Dice score is a metric to quantify the similarity between two binary segmentations X and Y as follows: $DSC = \frac{2\,\lvert X\cap Y\rvert}{\lvert X\rvert+\lvert Y\rvert}$. In other words, it roughly quantifies the amount of spatial overlap between both segmentations. For multi-label evaluation, binary representations of ground truth and prediction are compared for each class. Besides, the Dice loss is represented by the Dice score in the following way: $DSC_{loss} = 1 - DSC$, since neural networks require ->***loss functions*** instead of score functions.
Ensembling	->***Regularization*** technique to improve performance. In more detail: Instead of one single learning algorithm, multiple neural networks are differently trained, and thus form different predictors to reduce prediction variance. Final results are performed by merging the predictions of all networks.
Epoch	An epoch ends when all training samples have been fed through the network once.
Feature	An individual, measurable property, e.g. glomerular size is a feature of the glomerulus.
Feature map	Spatially arranged features that are generated by applying filters to the convolutional layer input, i.e. the input image or feature map outputs from the prior layer. Example: A convolutional filter has been applied to the left image resulting in a two-dimensional feature map highlighting its edges. 
Ground Truth	Target data we expect the network to predict. We annotate and classify structures according to *our* renal ->***class*** definitions in Supp. Table 2 and consider these annotations and classifications to correspond to reality, thus representing the ground truth. Example: Ground truth image of the left image is shown right.

	
Hyperparameter	Special ->***parameters*** to control e.g. the learning process or architecture of the deep learning model. They are determined by the experimentator before as well as dynamically during training. Examples are the amount of ->***epochs*** or the ->***kernel size.***
Image segmentation	Decomposition of an image into structures of interest. Example: Segmentation of a tubule. 
Instance	A single structure of a class. Example: All tubular instances are differently colored (Image from Supp. Fig. 5, third column). 
Instance normalization	->***Regularization*** technique applied in neural networks. In more detail: In contrast to the widely used batch normalization, instance normalization normalizes each ->***feature map*** independently providing zero mean and unit variance.
Kernel size	Specifies the size of a convolutional filter that is slid over the image.
Loss function	A mathematical function measuring the dissimilarity between network prediction and ->***ground truth***. In more detail: To train a neural network, a (differentiable) mathematical loss function representing a metric to measure the dissimilarity between prediction and ground-truth is required. During training, the network is consecutively optimized (with respect to the loss function) to lower the loss and thus to improve the similarity between prediction and ground-truth.
Negative slope	->***Hyperparameter*** in the mathematical LeakyReLU function. In more detail: The LeakyReLU function is defined as follows: $LeakyReLU(x) = \begin{cases} x, & x \geq 0 \\ negative\_slope * x, & otherwise \end{cases}$ Thus, the $negative\_slope$-hyperparameter specifies the slope of the LeakyReLU function for negative inputs, i.e. $x < 0$. Most commonly, $negative\_slope = 0.01$ is chosen by the experimentator.
Padding	An operation within convolutional layers to artificially enlarge the input data. In more detail: Specifies how much the input data is spatially padded around it. Padding an image with zeros exemplary means that zero values are added around it. Padding is used to counteract shrinkage of the input data caused by convolution.

	Example: without padding  with padding 
Parameter	Components of a (deep learning) system that fully define and characterize the system. In more detail: During network training, its trainable parameters are optimized. After training, all network parameters (trainable and non-trainable) are held constant, and the model is then used for prediction computation.
Receptive field	The prediction of a single output pixel only depends on a certain region of the input image. This region represents its receptive field. The size depends on the architecture of the network.
Reduce-On-Plateau	Technique to schedule the learning rate. In more detail: The learning rate represents an important ->***hyperparameter*** in neural networks that controls the speed of learning. This learning rate scheduler reduces the learning rate by a specific factor each time when the validation error has not decreased for a certain number of epochs.
Regularization	Regularization techniques are employed to improve network's generalization, i.e. reducing the error on test data. At the expense of increased training error, such techniques impose particularly designed constraints to the neural network preventing them to solely memorize the training data without having learned the underlying patterns.
ReLU	Stands for *rectified linear unit* and represents a mathematical function defined as follows: $ReLU(x) = \begin{cases} x, & x \geq 0 \\ 0, & otherwise \end{cases}$
Robustness	Describes the extent of input variability (e.g. in tissue morphology, staining, slide thickness, laboratory) an algorithm can cope with. Generally, it is measured by performance evaluation on those variabilities (usually held-out as in the current study).
Stride	An operation within convolutional layers to specify how many pixels the convolutional filter (or: ->***kernel***) is moved when slid over the image. Example: stride of "1" (shift of 1 pixel)  stride of "2" (shift of 2 pixels). 
Test-time augmentation	->***Regularization*** technique to improve performance. In more detail: Regularization technique that forwards flipped versions of the input through the network and averages their respectively back-flipped predictions to yield the final prediction. In contrast to ->***ensembling***, just a single network/predictor is used to perform multiple estimations.

	 CNN CNN CNN CNN Backflip Backflip Backflip Average
Transposed convolutions	The conventional convolution provides a many-to-one relationship between input and output, since many input pixels are connected to a single value in the output. In contrast, transposed convolutions make use of a reversed pixel connectivity (in backward direction) providing a one-to-many relationship. Thus, it is designed for image ->***upsampling***. 
Upsampling	Expansion or increase of the spatial resolution of an image. <u>In more detail:</u> Upsampling can be exemplarily performed by pixel interpolation meaning that new pixel values can be estimated between pixels by using their neighborhood, e.g. by averaging neighboring pixels values (ultimately yielding a denser image grid). The picture in ->***transposed convolutions*** exemplarily shows an upsampling of an artificial image.

Supplementary Table 2. Criteria for definition of classes.

Class	Criteria
Full glomerulus	- annotation along Bowman's capsule - if cross section showed urinary (or vascular) pole, glomerulus was encircled in round/oval shape
Glomerular tuft	- subclass of the full glomerulus class - annotation of glomerular tuft only (including podocytes) - for glomerular lesions: extracapillary proliferates (= crescents), parietal epithelial cells which migrated onto the tuft or tip lesions were not included
Tubule	- annotation along, but excluding, the basement membrane
Artery	- annotation of all arteries, including all arterial branches to arterioles - at least one visible vascular smooth muscle cell layer required
Arterial lumen	- subclass of the artery class - annotation of lumen only, excluding also the endothelium
Vein	- annotation of large "white" areas - only the lumen, i.e. the "white" area was annotated - for veins the definition of larger vessels next to arteries with a minimal diameter of 30µm - class includes non-tissue background and renal pelvis

Supplementary Table 3. Quantitative information on ground truth data.

Model / Species	Number of annotated patches / WSI	Train / val / test split of annotated patches	Train / val / test split of partially annotated WSI	Total number of instance annotations						Σ
				full glom.	glom. tuft	tubule	artery	arterial lumen	vein	
Healthy mouse	820 / 41	600 / 60 / 160	30 / 3 / 8	835	804	18536	1107	1416	609	23307
UUO	300 / 15	220 / 20 / 60	11 / 1 / 3	225	221	6795	301	314	177	8033
IRI	300 / 15	220 / 20 / 60	11 / 1 / 3	242	242	7555	354	397	102	8892
Adenine	300 / 15	220 / 20 / 60	11 / 1 / 3	257	256	5995	342	384	111	7345
Alport	300 / 15	220 / 20 / 60	11 / 1 / 3	413	368	7137	361	383	83	8745
NTN	300 / 15	220 / 20 / 60	11 / 1 / 3	247	237	5500	275	295	139	6693
db/db	30 / 3	0 / 0 / 30	0 / 0 / 3	27	27	652	27	22	10	765
Ext. UUO	30 / 3	0 / 0 / 30	0 / 0 / 3	46	43	879	42	27	8	1045
Human	230 / 12	200 / 0 / 30	10 / 0 / 2	123	148	1958	125	145	40	2539
Rat	80 / 8	50 / 0 / 30	5 / 0 / 3	56	59	1372	66	74	27	1654
Pig	80 / 6	50 / 0 / 30	5 / 0 / 1	50	49	900	57	67	23	1146
Marmoset	80 / 8	50 / 0 / 30	5 / 0 / 3	39	39	774	62	70	28	1012
Black bear	80 / 8	50 / 0 / 30	5 / 0 / 3	51	51	1240	85	91	28	1546
Σ	2930 / 164	2100 / 160 / 670	115 / 8 / 41	2611	2544	59293	3204	3685	1385	72722

IRI = ischemia reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction, val = validation

~~Supplementary Table 2. Quantitative information on ground truth data.~~

~~Model / Species~~	~~Number of annotated Patches/WSI~~	~~Total number of instance annotations~~						~~Σ~~
		~~full glomerulus~~	~~glomerular tuft~~	~~tubule~~	~~artery~~	~~arterial lumen~~	~~Vein~~	
~~Healthy mouse~~	~~820 / 41~~	~~835~~	~~804~~	~~18536~~	~~1107~~	~~1416~~	~~609~~	~~23307~~
~~UUO~~	~~300 / 15~~	~~225~~	~~221~~	~~6795~~	~~301~~	~~314~~	~~177~~	~~8033~~
~~IRI~~	~~300 / 15~~	~~242~~	~~242~~	~~7555~~	~~354~~	~~397~~	~~102~~	~~8892~~
~~Adenine~~	~~300 / 15~~	~~257~~	~~256~~	~~5995~~	~~342~~	~~384~~	~~111~~	~~7345~~
~~Alport~~	~~300 / 15~~	~~413~~	~~368~~	~~7137~~	~~361~~	~~383~~	~~83~~	~~8745~~
~~NTN~~	~~300 / 15~~	~~247~~	~~237~~	~~5500~~	~~275~~	~~295~~	~~139~~	~~6693~~
~~Human~~	~~200 / 10~~	~~108~~	~~126~~	~~1678~~	~~96~~	~~115~~	~~31~~	~~2154~~
~~Rat~~	~~50 / 5~~	~~32~~	~~31~~	~~895~~	~~33~~	~~34~~	~~14~~	~~1039~~
~~Pig~~	~~50 / 5~~	~~34~~	~~34~~	~~616~~	~~38~~	~~46~~	~~12~~	~~780~~
~~Marmoset~~	~~50 / 5~~	~~24~~	~~24~~	~~535~~	~~32~~	~~38~~	~~14~~	~~667~~
~~Black bear~~	~~50 / 5~~	~~30~~	~~32~~	~~689~~	~~49~~	~~55~~	~~13~~	~~868~~
~~Σ~~	~~2720 / 146~~	~~2447~~	~~2375~~	~~55931~~	~~2988~~	~~3477~~	~~1305~~	~~68523~~

~~IRI = ischemia reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction~~

Supplementary Table 4. Architecture of our CNN.

Network Architecture	Output size
Input image layer	640 x 640 x 3
Conv2d(i: 3, o: 32, k: 3, s: 1, p: 1) + IN(o: 32) + LeakyReLU(sl: 0.01)	640 x 640 x 32
Conv2d(i: 32, o: 32, k: 3, s: 1, p: 1) + IN(o: 32) + LeakyReLU(sl: 0.01)	640 x 640 x 32
MaxPool2d(k: 2, s: 2, p: 0)	320 x 320 x 32
Conv2d(i: 32, o: 64, k: 3, s: 1, p: 1) + IN(o: 64) + LeakyReLU(sl: 0.01)	320 x 320 x 64
Conv2d(i: 64, o: 64, k: 3, s: 1, p: 1) + IN(o: 64) + LeakyReLU(sl: 0.01)	320 x 320 x 64
MaxPool2d(k: 2, s: 2, p: 0)	160 x 160 x 64
Conv2d(i: 64, o: 128, k: 3, s: 1, p: 1) + IN(o: 128) + LeakyReLU(sl: 0.01)	160 x 160 x 128
Conv2d(i: 128, o: 128, k: 3, s: 1, p: 1) + IN(o: 128) + LeakyReLU(sl: 0.01)	160 x 160 x 128
MaxPool2d(k: 2, s: 2, p: 0)	80 x 80 x 128
Conv2d(i: 128, o: 256, k: 3, s: 1, p: 1) + IN(o: 256) + LeakyReLU(sl: 0.01)	80 x 80 x 256
Conv2d(i: 256, o: 256, k: 3, s: 1, p: 1) + IN(o: 256) + LeakyReLU(sl: 0.01)	80 x 80 x 256
MaxPool2d(k: 2, s: 2, p: 0)	40 x 40 x 256
Conv2d(i: 256, o: 512, k: 3, s: 1, p: 1) + IN(o: 512) + LeakyReLU(sl: 0.01)	40 x 40 x 512
Conv2d(i: 512, o: 512, k: 3, s: 1, p: 1) + IN(o: 512) + LeakyReLU(sl: 0.01)	40 x 40 x 512
MaxPool2d(k: 2, s: 2, p: 0)	20 x 20 x 512
Conv2d(i: 512, o: 1024, k: 3, s: 1, p: 1) + IN(o: 1024) + LeakyReLU(sl: 0.01)	20 x 20 x 1024
Conv2d(i: 1024, o: 1024, k: 3, s: 1, p: 1) + IN(o: 1024) + LeakyReLU(sl: 0.01)	20 x 20 x 1024
ConvTranspose2d(i: 1024, o: 1024, k: 2, s: 2)	40 x 40 x 1024
Conv2d(i: 1536, o: 512, k: 3, s: 1, p: 0) + IN(o: 512) + LeakyReLU(sl: 0.01)	38 x 38 x 512
Conv2d(i: 512, o: 512, k: 3, s: 1, p: 0) + IN(o: 512) + LeakyReLU(sl: 0.01)	36 x 36 x 512
ConvTranspose2d(i: 512, o: 512, k: 2, s: 2)	72 x 72 x 512
Conv2d(i: 768, o: 256, k: 3, s: 1, p: 0) + IN(o: 256) + LeakyReLU(sl: 0.01)	70 x 70 x 256
Conv2d(i: 256, o: 256, k: 3, s: 1, p: 0) + IN(o: 256) + LeakyReLU(sl: 0.01)	68 x 68 x 256
ConvTranspose2d(i: 256, o: 256, k: 2, s: 2)	136 x 136 x 256
Conv2d(i: 384, o: 128, k: 3, s: 1, p: 0) + IN(o: 128) + LeakyReLU(sl: 0.01)	134 x 134 x 128
Conv2d(i: 128, o: 128, k: 3, s: 1, p: 0) + IN(o: 128) + LeakyReLU(sl: 0.01)	132 x 132 x 128
ConvTranspose2d(i: 128, o: 128, k: 2, s: 2)	264 x 264 x 128
Conv2d(i: 192, o: 64, k: 3, s: 1, p: 0) + IN(o: 64) + LeakyReLU(sl: 0.01)	262 x 262 x 64
Conv2d(i: 64, o: 64, k: 3, s: 1, p: 0) + IN(o: 64) + LeakyReLU(sl: 0.01)	260 x 260 x 64
ConvTranspose2d(i: 64, o: 64, k: 2, s: 2)	520 x 520 x 64
Conv2d(i: 96, o: 32, k: 3, s: 1, p: 0) + IN(o: 32) + LeakyReLU(sl: 0.01)	518 x 518 x 32
Conv2d(i: 32, o: 32, k: 3, s: 1, p: 0) + IN(o: 32) + LeakyReLU(sl: 0.01)	516 x 516 x 32
Conv2d(i: 32, o: 8, k: 1, s: 1, p: 0)	516 x 516 x 8

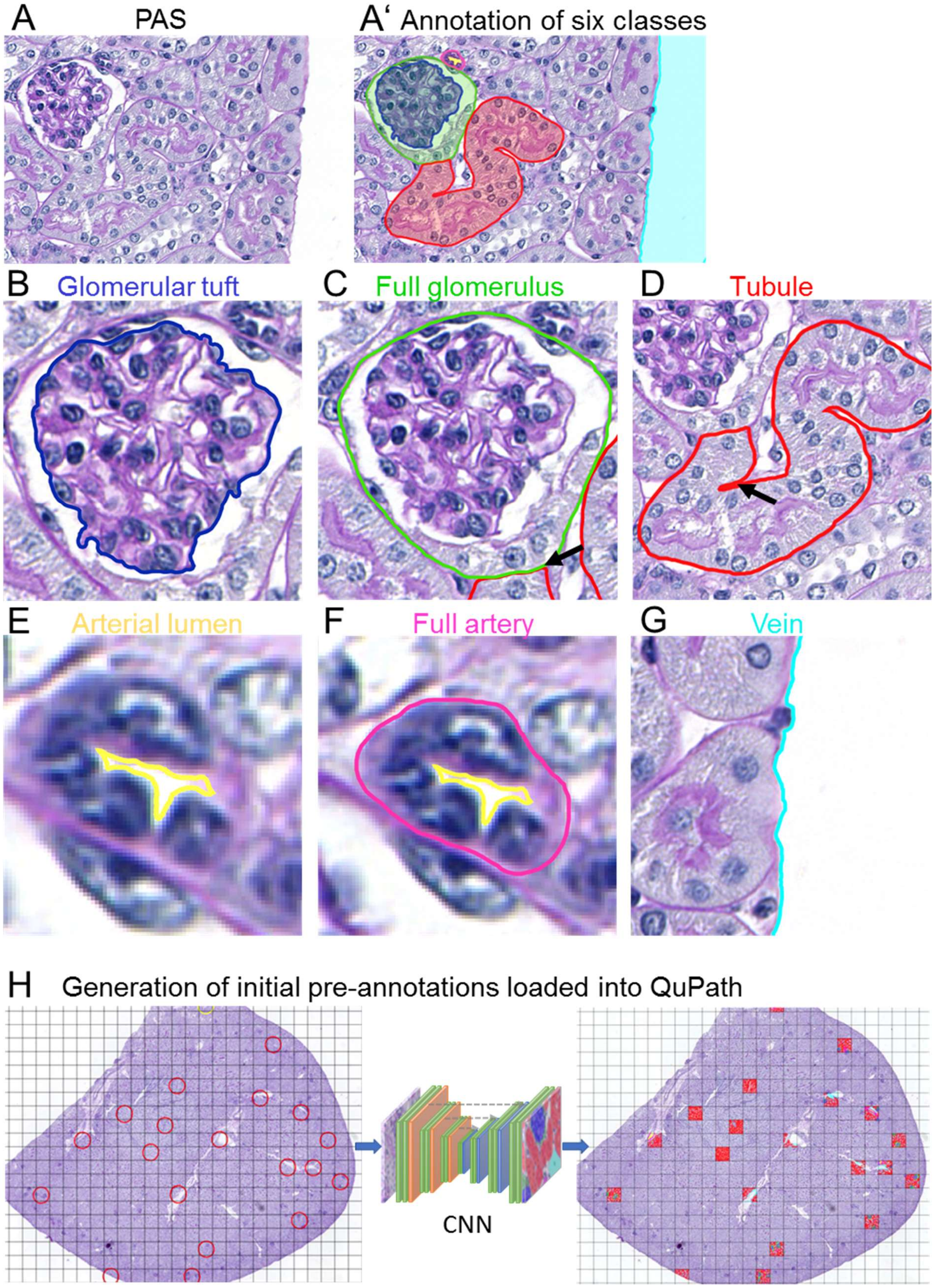
Conv2d = two-dimensional convolutional layer, IN = instance normalization, i = #input layers, o = #output layers, k = kernel size, s = stride, p = padding, sl = negative slope

Supplementary Table 5. Performance comparison of our model, its unmodified variant vanilla u-net, and state-of-the-art context-encoder.

Shown are mean object-level dice scores for our model / the unmodified variant vanilla u-net / state-of-the-art context-encoder. The highest Score is marked in bold. * p < 0.05 vs. vanilla u-net and ° p < 0.05 vs. context-encoder.

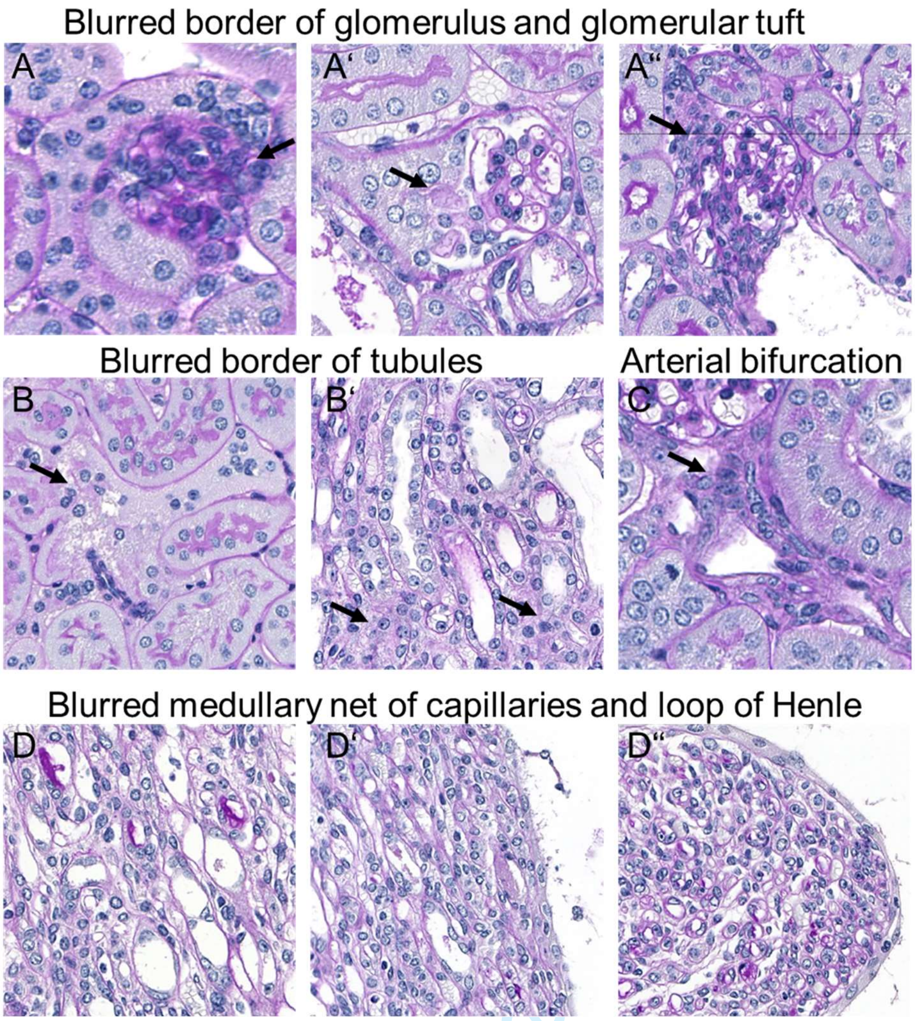
Mouse Model	Segmentation performance of our model / vanilla u-net / context-encoder					
	full glomerulus	glomerular tuft	tubule	artery	arterial lumen	vein
Healthy	**96.5** / 95.6 / 96.2	**93.8** / **93.8** / 93.5	**93.3** / 92.9 / 93.0	**88.1** / 87.4 / 87.8	80.3 / 80.0 / **80.6**	**94.3** / 88.9 / 92.0
UUO	**97.5** / 95.2 / 95.3	**95.6** / 93.9 / 94.5	90.8 / 90.8 / **91.3**	82.3 / 81.2 / **82.6**	**75.0** / 72.9 / 73.7	**97.6** / 95.4 / 94.6
IRI	96.0 / **97.7** / 95.7	**95.4** / 94.7 / 94.4	**90.2** / 89.1 / 89.9	**79.1** / 74.7 / 74.2	**73.5** / 62.3 / 61.7	**97.7** / 86.7 / 87.0
Adenine	**98.8** / 94.1 / 98.5	**97.2** / 94.1 / 97.1	**93.0** / 92.0 / 92.8	**87.9** / 83.3 / 83.2	**80.9** / 72.7 / 76.9	93.6 / 87.6 / **96.7**
Alport	94.7 / 95.5 / **96.3**	**91.3** / 86.4 / 87.6	**90.6** / 89.7 / 89.3	**80.3** / 74.2 / 72.0	**81.1** / 69.9 / 65.5	**89.2** / 83.2 / 81.7
NTN	95.5 / 91.5 / **96.3**	**94.8** / 93.9 / 93.9	**93.2** / 92.5 / 92.9	**86.8** / 82.7 / 83.9	78.2 / 73.9 / **79.1**	92.8 / 91.8 / **95.4**
ø	**96.4*** / 94.0 / 96.3	**94.2*** / 92.6 / 93.0	**92.0*** / 91.4 / 91.7	**85.3*°** / 82.8 / 82.9	**79.1*°** / 75.9 / 76.1	**94.3*** / 90.4 / 92.7

IRI = ischemia reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction

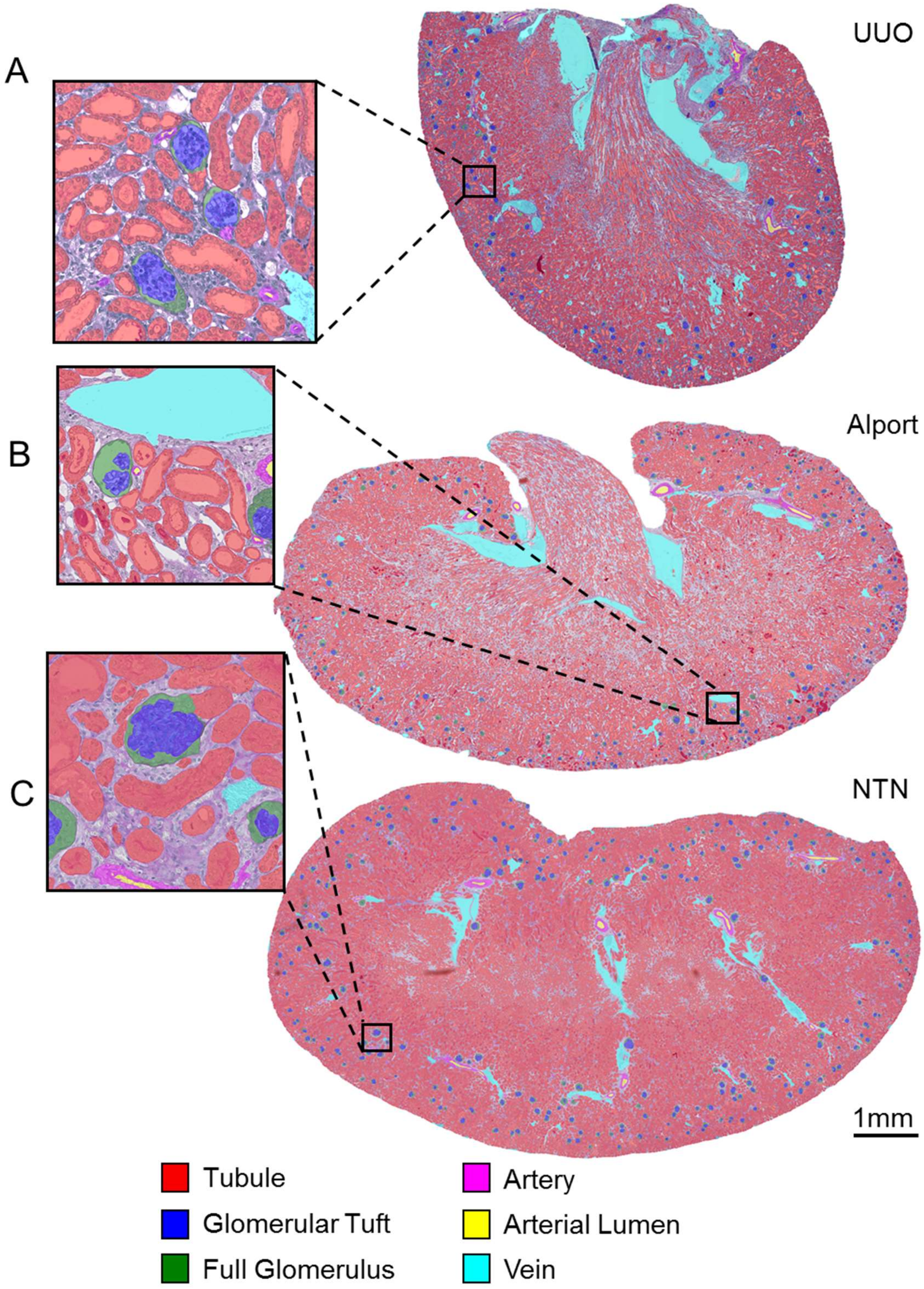

A
PAS
A' Annotation of six classes
B Glomerular tuft
C Full glomerulus
D Tubule
E Arterial lumen
F Full artery
G Vein
H Generation of initial pre-annotations loaded into QuPath
CNN

Supp. Fig. 1. Annotation procedure.

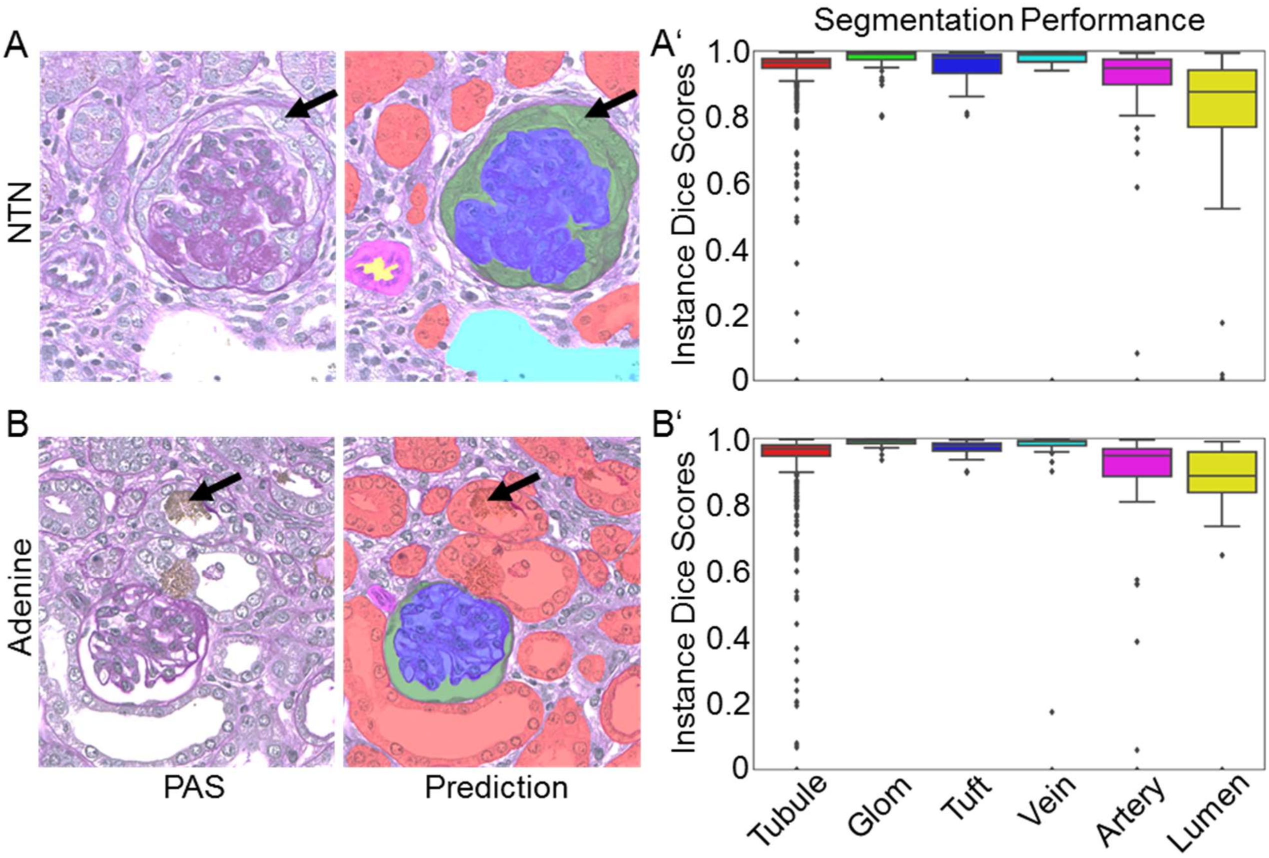
A representative picture of a PAS stained mouse kidney section (A) and an overlay with manual annotations for six classes (A'). The annotation of the "glomerular tuft" (blue (B)) included the capillary tuft, the mesangium and podocytes. A "full glomerulus" (green (C)) was annotated along bowman's capsule and included the tuft, bowman's space and parietal epithelial cells. The glomerular tuft was always a subclass of the full glomerulus. A full glomerulus always had a round or oval shape, this determined the separation from the proximal tubule (arrow). Tubules (red (D) were annotated along (but excluding) the tubular basement membrane, tangentially cut tubules without cytoplasm were excluded. The “arterial lumen” (yellow (D)) was always a subclass of the “artery” class (magenta (F)). Veins, background and renal pelvis were big “white” areas without tissue (cyan (G)). From the first manual annotations, we predicted initial pre-annotations for 20 patches per WSI and loaded them into Qupath for manual corrections facilitating annotation effort (H).


Blurred border of glomerulus and glomerular tuft
A
A'
A''
Blurred border of tubules
B
B'
Arterial bifurcation
C
Blurred medullary net of capillaries and loop of Henle
D
D'
D''

Supp. Fig. 2. Challenging morphology for manual and automated annotations. (A-A'') show examples of glomeruli in PAS stained murine kidney sections. On a sectional plane close to the vascular or urinary pole it was difficult to discriminate between glomerular tuft and arterioles (arrow, A), or the glomerular tuft and parietal epithelial cells or tubular epithelial cells (arrows, A',A''). Sometimes the tubular basement membrane appeared discontinuous (arrows in B, B'). The distinction of medial layers of arteries was harder when vessels run side by side (arrow, C). (D-D'') show medulla of murine kidneys with the network of capillaries and the tubular system, which in some cases was not easy to discriminate.

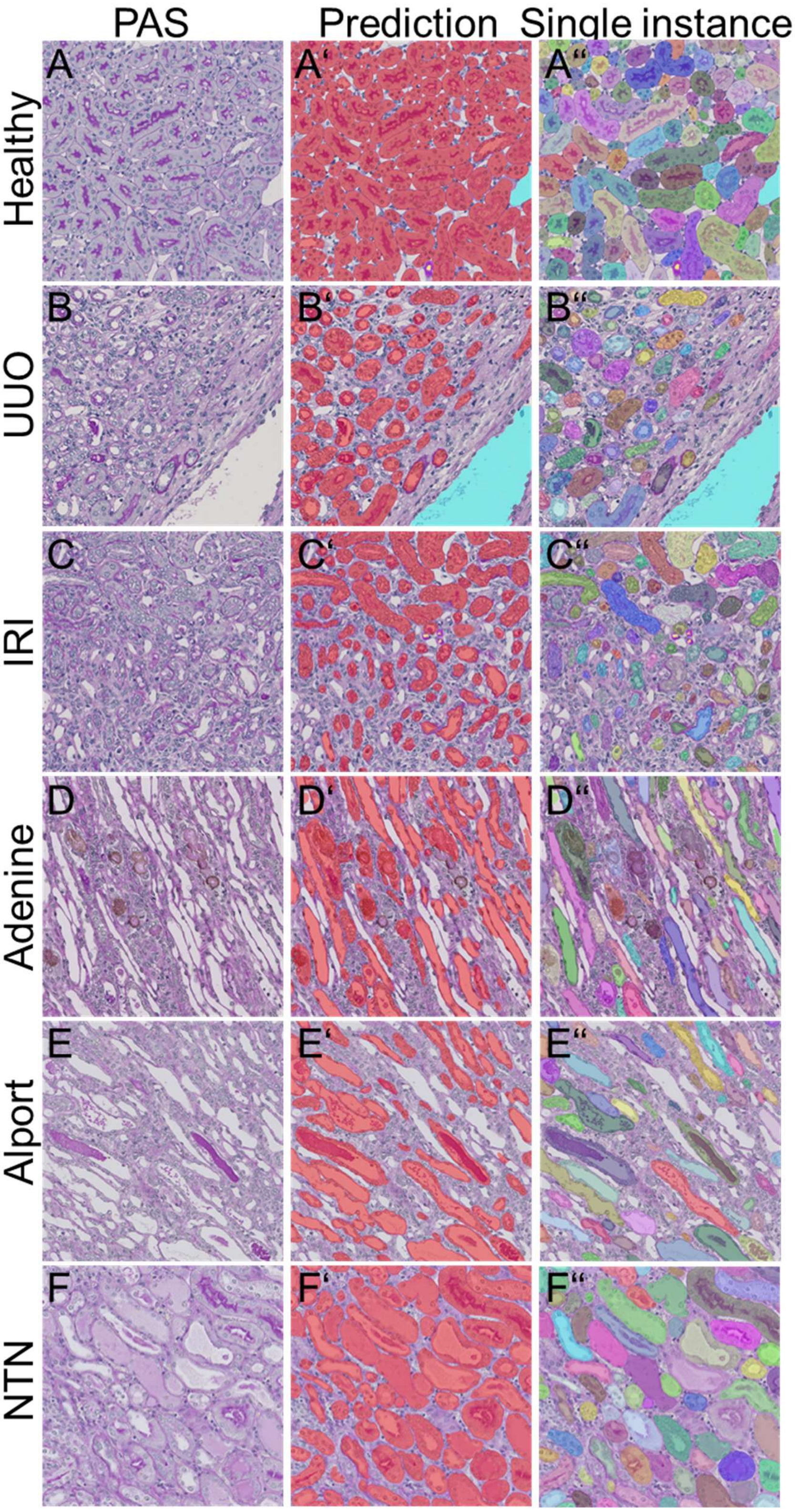


Supp. Fig. 3. Segmentation of WSI of UUO, Alport and NTN kidneys.
CNN generated segmentation predictions on a whole slide image (WSI) of an UUO (A), Alport (B) and NTN (C) mouse kidney. All six classes, were precisely segmented. NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.



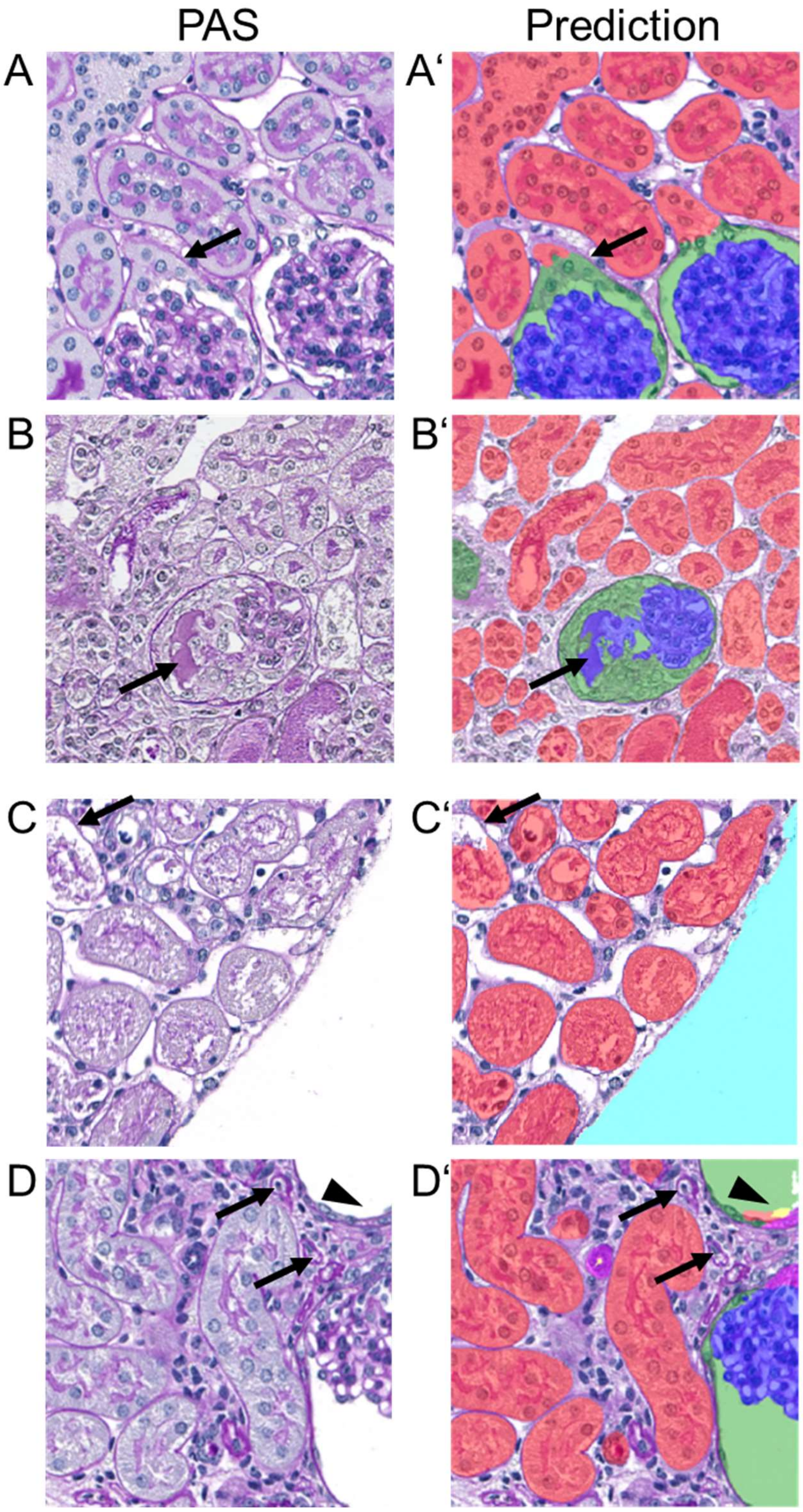
Supp. Fig. 4. Quantitative segmentation performance in murine NTN and adenine kidneys.

Representative PAS pictures and the corresponding segmentation prediction generated by our CNN for a murine NTN (A) and adenine kidney (B). Instance segmentation accuracy is shown by dice scores for each class in both models (A'-B'). Data are presented in Box plots with median, quartiles and whiskers. NTN = nephrotoxic nephropathy.

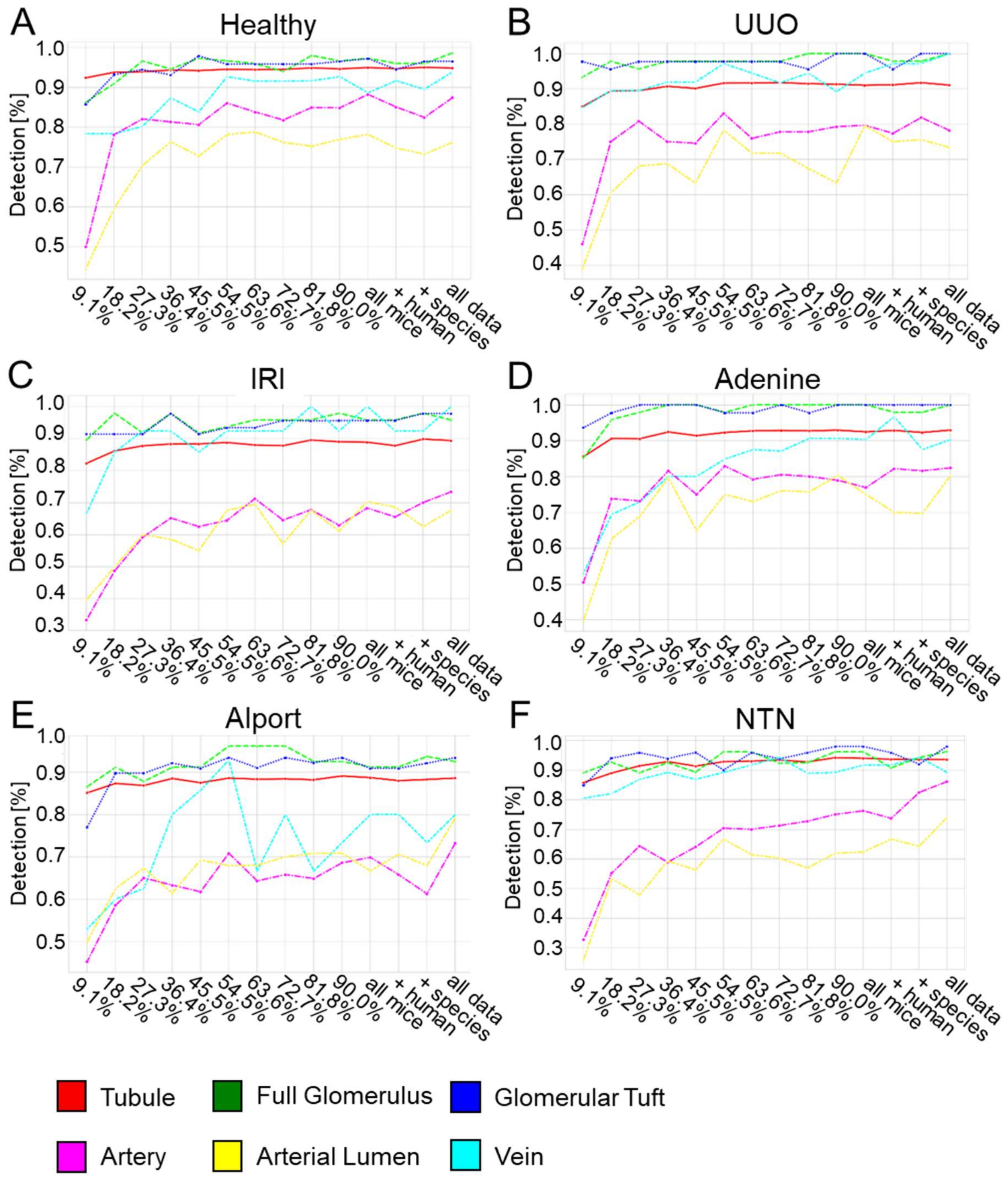


Supp. Fig. 5. Automated segmentation in the medulla of murine kidney sections. Representative PAS pictures and corresponding overlays with segmentation predictions showing either the different classes or every single instances for the medulla of murine healthy (A-A''), UUO (B-B''), IRI (C-C''), adenine (D-D''), Alport (E-E'') and NTN (F-F'') kidneys.

IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.



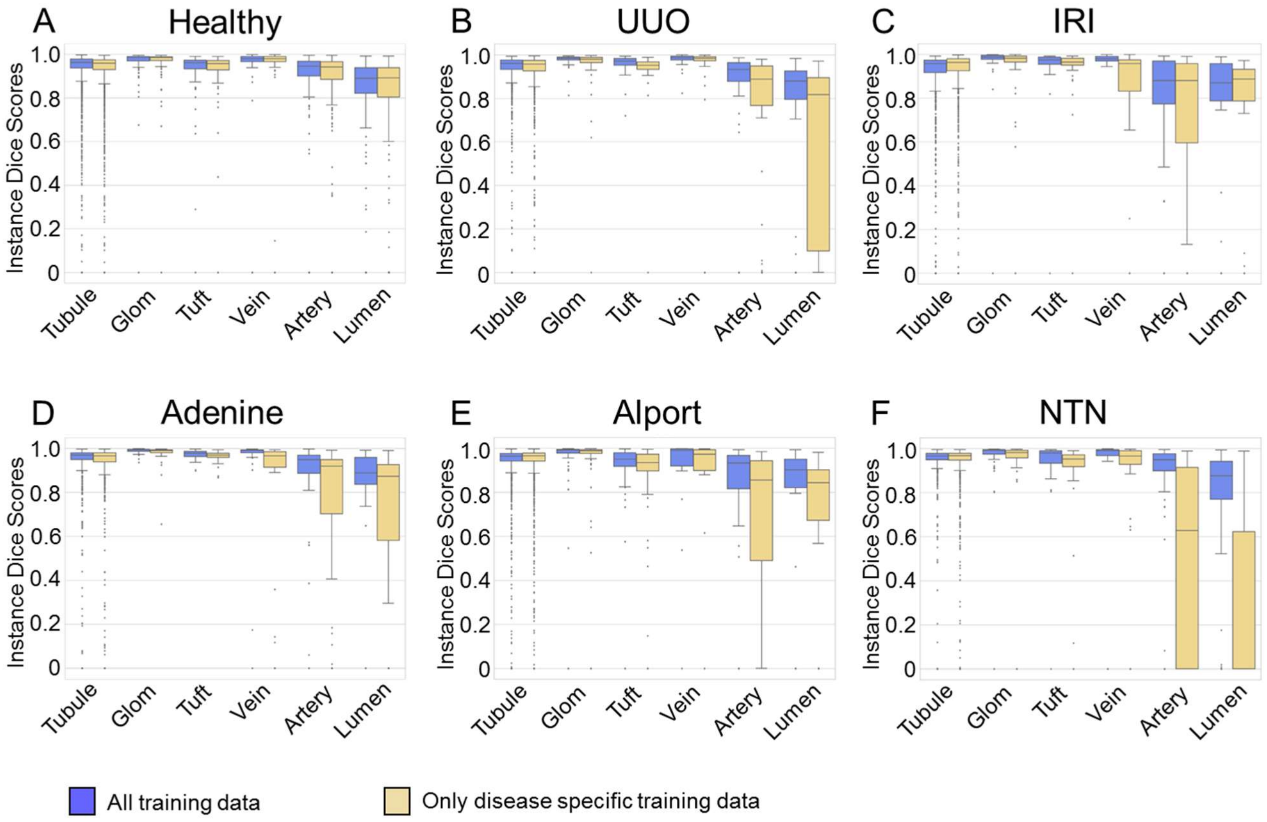
Supp. Fig. 6. Examples of ~~incorrectly segmented instances~~missclassifications. PAS photographs and prediction overlays show an incorrect separation of a "full glomerulus" and the connected proximal "tubule" (arrow in A, A'), a glomerular tuft that was inaccurately segmented with projections into the crescent (arrow in B, B') and an incompletely segmented tubule due to extensive necrosis (arrow in C,C'). Another example shows a strongly dilated tubule which is was incorrectly classified as full glomerulus and arterial lumen (arrowheads in D,D') and missing segmentations of atrophic tubules (arrows in D,D').



Supp. Fig. 7. Relation between amount of training data and detection performance.

The detection performance for all six classes in healthy (A), UUO (B), IRI (C), adenine (D), Alport (E) and NTN (F) was plotted against the amount of total data used for CNN training.

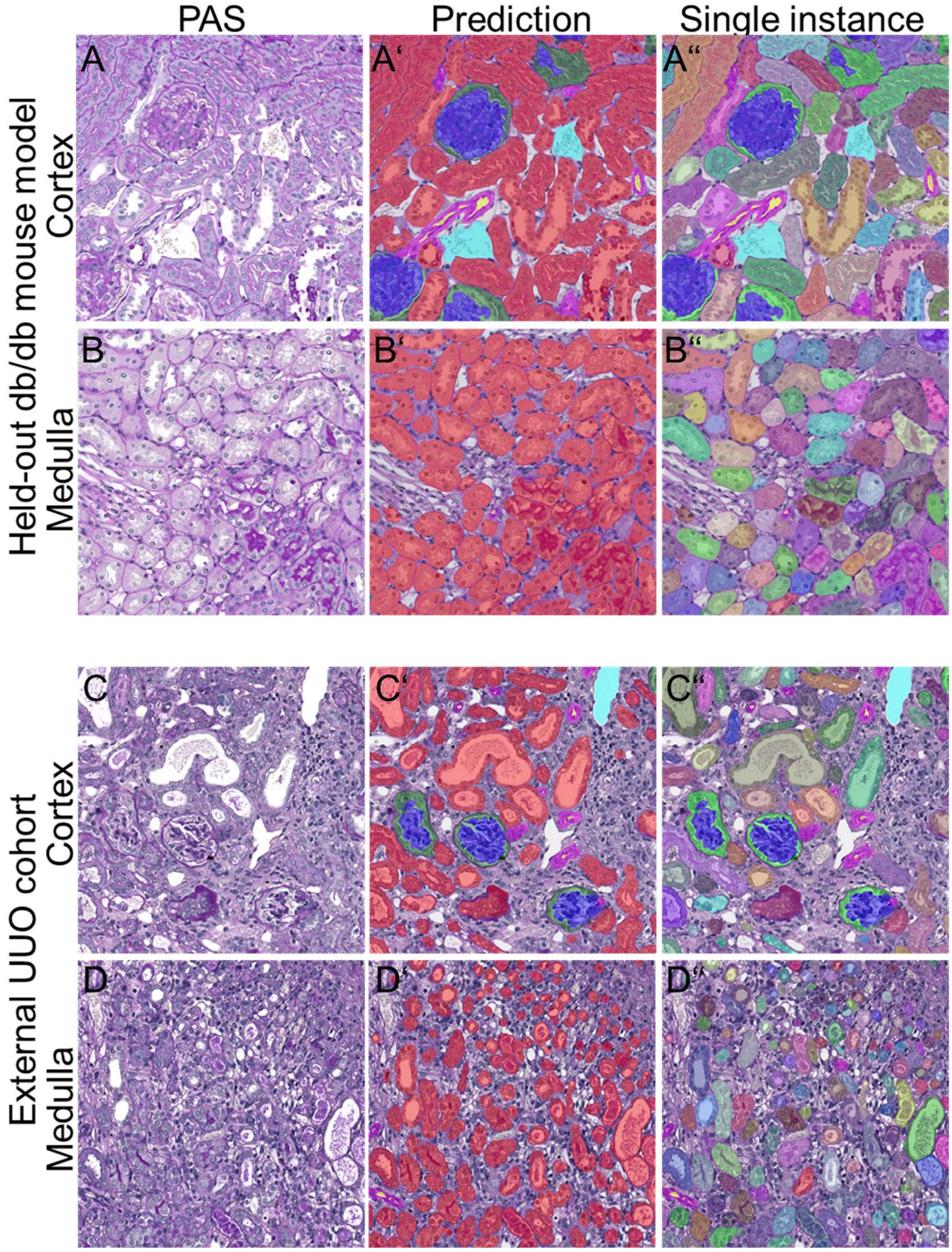
IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.



Supp. Fig. 8. Comparison between our full CNN and its variants independently trained on single models.~~the fully trained CNN and its variants.~~

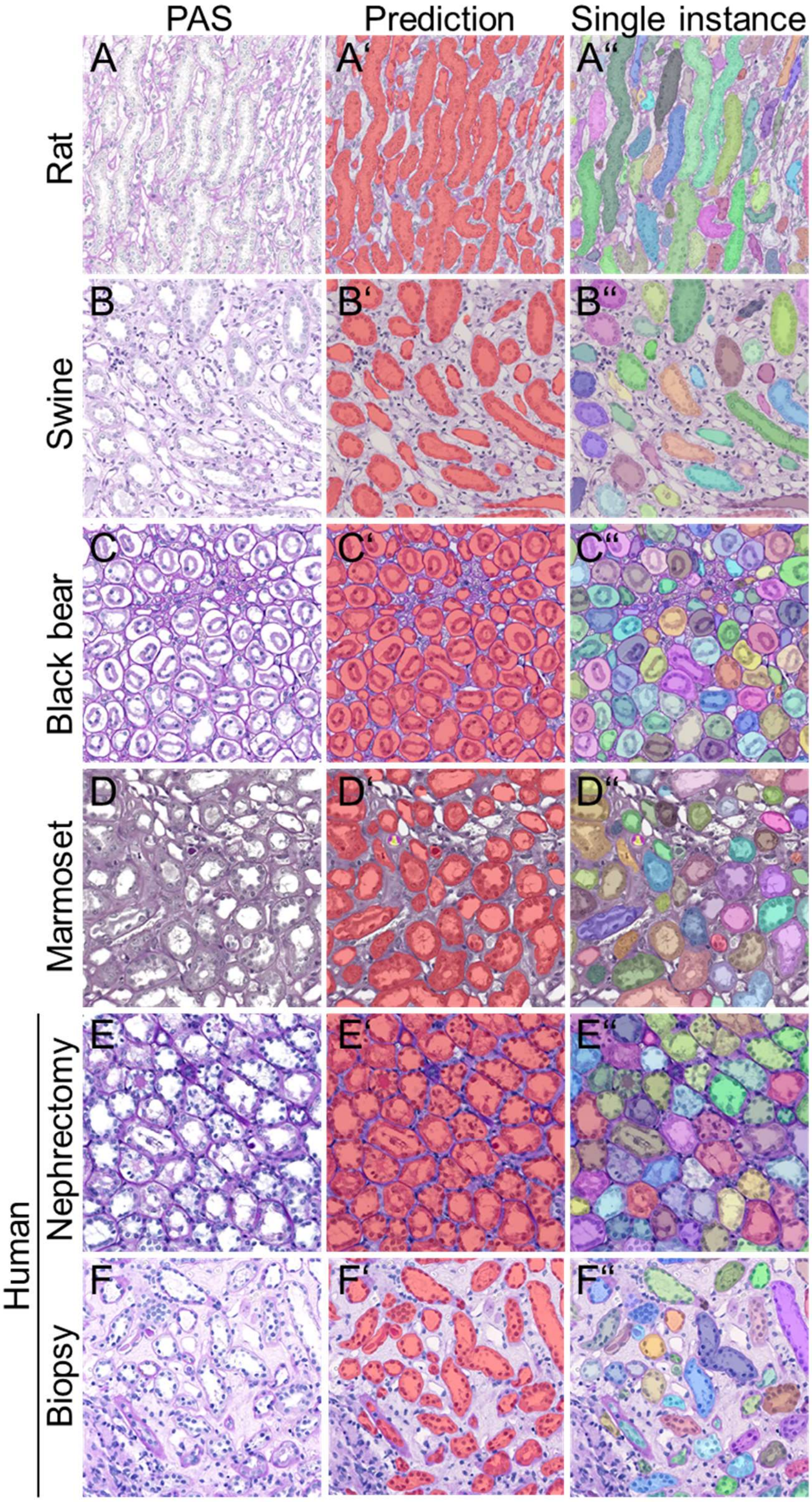
(A) Segmentation performance shown as instance dice scores for all six classes ~~in healthy kidneys~~was compared on our healthy kidney test data between our full~~y trained~~ CNN trained on all training data (blue) and its variant~~s~~ that ~~have~~has been solely trained with data from healthy kidneys (yellow). (B) The same comparison is shown for the UUO, in which the network variant was exclusively trained with annotations from UUO kidneys. Analogously, analyses are performed for IRI (C), adenine (D), Alport (E) and NTN (F).

Data are presented in Box plots with median, quartiles and whiskers. IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.

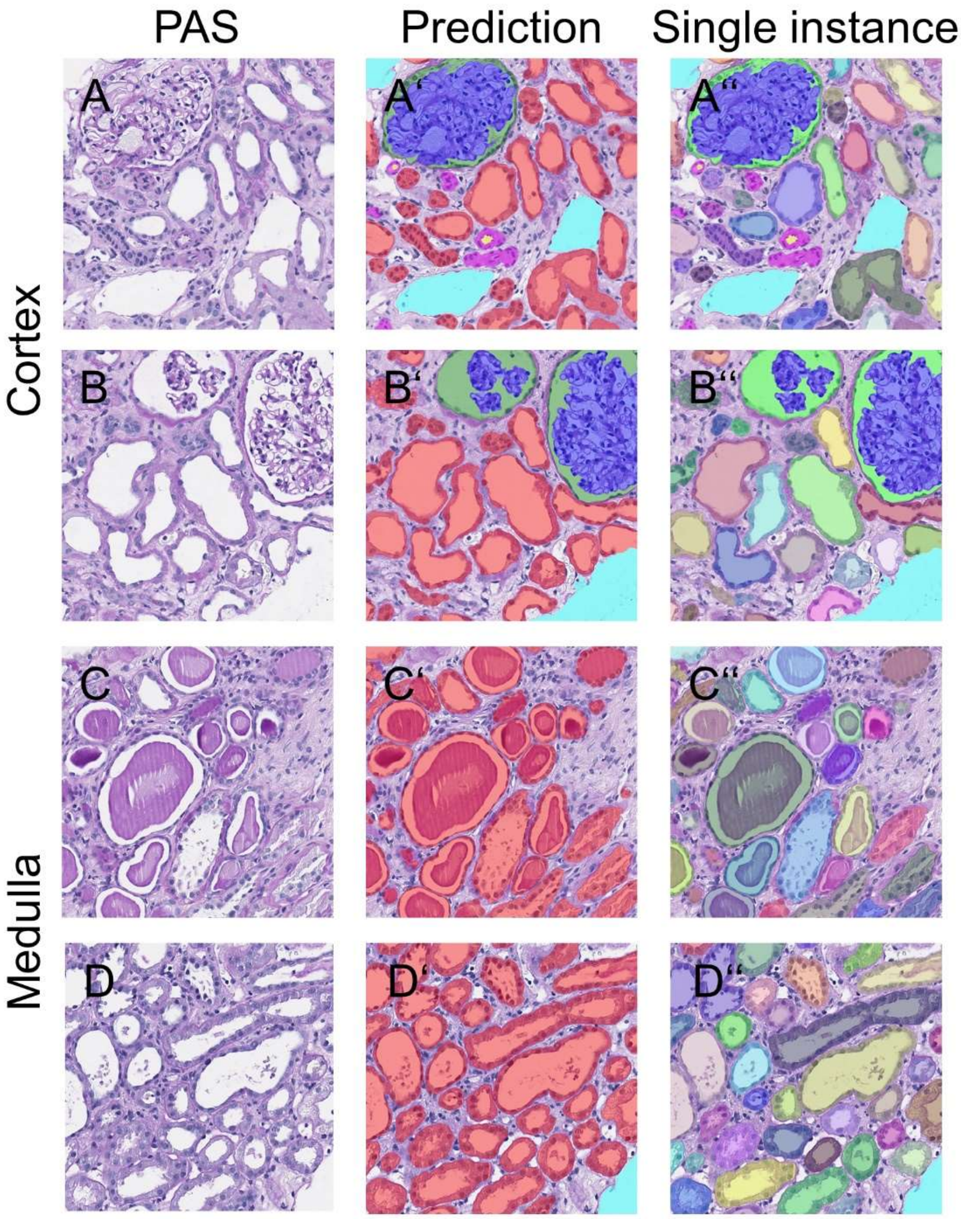


Supp. Fig. 9. Segmentation of non-trained and external murine kidney slides. Representative pictures show segmentation results for cortex (A-A'') and medulla (B-B'') for kidneys from db/db mice fed with high fat western diet. Predictions (A', B') depict different classes, while A'' and B'' display segmentation on single instance level. The CNN also accurately segments cortex (C-C'') and medulla (D-D'') from PAS slides of an external UUO cohort. Predictions (C', D') depict different classes, while C'' and D'' display segmentation on single instance level.

UUO = unilateral ureteral obstruction.



Supp. Fig. 10. Automated segmentation of renal medulla in different species. Representative PAS pictures and the corresponding overlays for segmentation predictions showing either the different classes or every single instance for the medulla of rat (A-A''), pig (B-B''), black bear (C-C''), marmoset (D-D'') and human (E-F'') kidneys. Segmentation is accurate on human nephrectomy (E-E'') as well as on biopsy specimens (F-F'').



Supp. Fig. 11. Automated segmentation of human biopsies presenting with acute tubular damage. Representative PAS-pictures and the respective segmentation prediction overlays from cortex (A-B'') and medulla (C-D'') of human biopsies with acute tubular damage.

Deep-Learning based segmentation and quantification in experimental kidney histopathology

Running Title: DL in experimental nephropathology

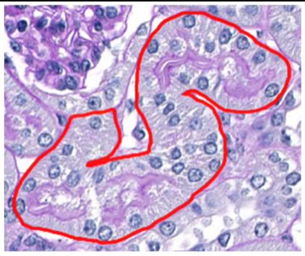
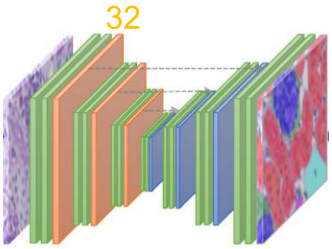
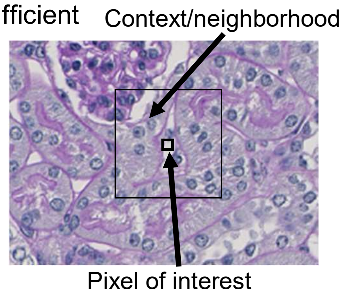
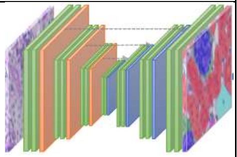
Nassim Bouteldja[2,*], Barbara M. Klinkhammer[1,3,*], Roman D. Bülow[1,*], Patrick Droste[1], Simon W. Otten[1], Saskia von Stillfried[1], Julia Moellmann[4], Susan M. Sheehan[5], Ron Korstanje[5], Sylvia Menzel[3], Peter Bankhead[6,7], Matthias Mietsch[8], Charis Drummer[9], Michael Lehrke[4], Rafael Kramann[3,10], Jürgen Floege[3], Peter Boor[1,3,*,#], Dorit Merhof[2,11,*]

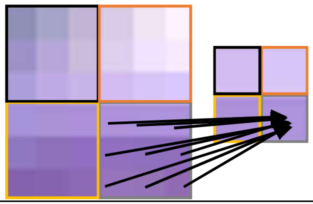
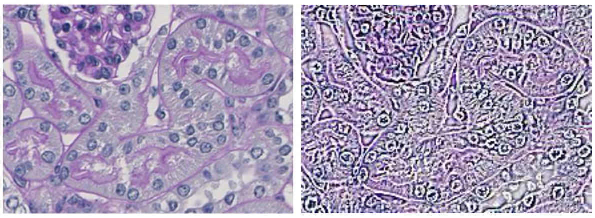
1 Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany
2 Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany
3 Department of Nephrology and Immunology, RWTH Aachen University Hospital, Aachen, Germany
4 Department of Cardiology and Vascular Medicine, RWTH Aachen University Hospital, Aachen, Germany
5 The Jackson Laboratory, Bar Harbor, Maine
6 Edinburgh Pathology, University of Edinburgh, Edinburgh, UK
7 Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
8 Laboratory Animal Science Unit, German Primate Center, Goettingen, Germany
9 Platform Degenerative Diseases, German Primate Center, Goettingen, Germany
10 Department of Internal Medicine, Nephrology and Transplantation, Erasmus Medical Center, Rotterdam, The Netherlands
11 Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

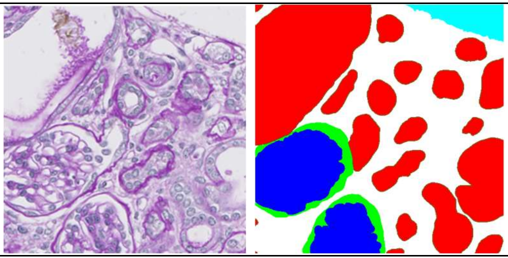
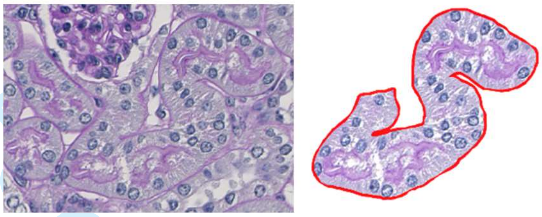
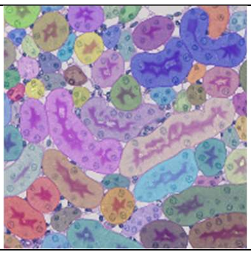
Supplementary material

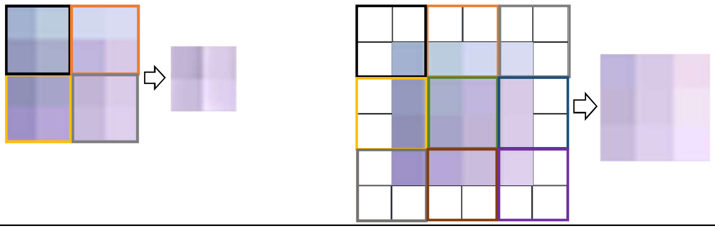
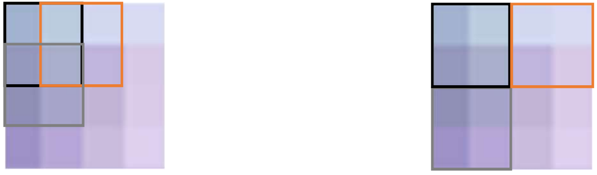
Content

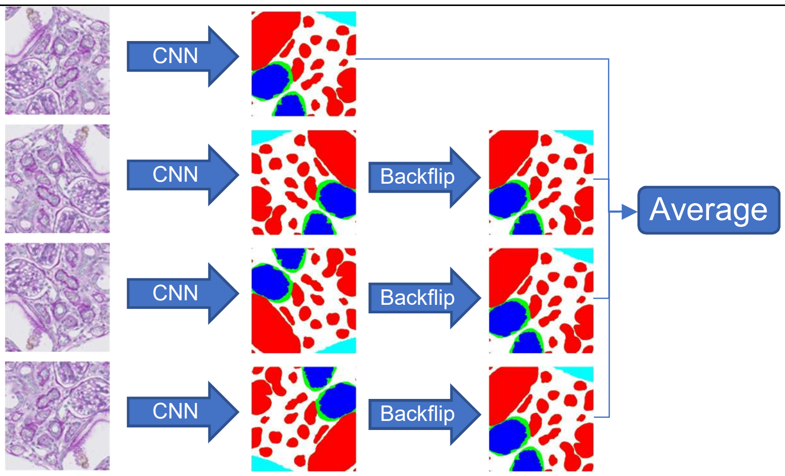
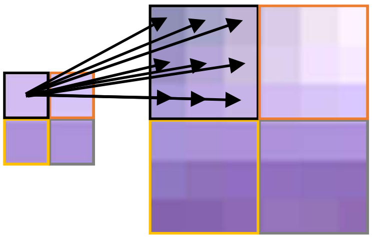
Supplementary Table 1. Glossary of technical terms.

Term	Description
Ablation study	Experiment with consecutively reduced input data. In more detail: A procedure where certain configurations of neural network architecture or training including modifications to data sets are changed to gain a better understanding of their importance and impact (mainly on overall performance).
Border class	->***Class*** comprising borders of structures. Example: The tubule's border marked in red is assigned to the border class.  In more detail: Artificial class representing the border of specific structures. In our application, we make use of a border class, that especially represents the tubular basement membrane, to separate tubular (as well as glomerular or arterial) instances from each other, allowing for instance-level analysis.
Capacity	Amount of ->***parameters*** in a neural network. In more detail: A neural network consists of many trainable parameters. Its number represents the network's capacity. It is also associated with its complexity, i.e. the degree of complexity of patterns the model is able to learn. Note that a neural network represents a mathematical function including input variables and parameters. Thus, the parameters are here defined in a mathematical way.
Channel numbers	Number of ->***feature maps***. Example: The channel number of the first, orange ->***convolutional layer*** is 32.  In more detail: In convolutional neural networks, input data is subsequently propagated through ->***convolutional layers*** each producing multiple output ->***feature maps***. Their number re-presents the channel number of the layer.
Class	A group of structures. Example: All tubular structures belong to the "tubule"-class.
Context-awareness	Ability of a method to incorporate sufficient spatial neighborhood information for the assessment / prediction of a pixel.  In more detail: The more spatial context is considered for pixel prediction, the more context-aware is a technique. In our case, our network provides sufficient spatial context even for pixel prediction at patch border.
Convolutional layer	Network layer performing convolutions to its input. Example: All green blocks represent such layers.  In more detail: Such layers represent substantial components in CNNs. Convolutions are performed on input data resulting in multiple ->***feature maps***. Convolutions are mainly specified based on the following ->***parameters***:

	->***kernel size***, ->***stride*** and ->***padding***. As exemplary shown on the right, a convolution (with 3x3 kernel size) slides over the image and outputs a single value for each 3x3 region. 
Cross-entropy loss	Information-theoretical measure of the dissimilarity between network output and ->***ground truth***. In more detail: A commonly used ->***loss function*** when training segmentation or classification networks. The Cross-entropy loss (CE) is based on information theory and measures the difference between a target probability distribution (represented by ground truth annotations) and an estimated one (represented by model predictions). Its values range between 0 and 1. The smaller the loss, the higher the similarity. Thus, a perfect overlap results in a value of zero.
Dice loss / Dice score	The Dice score measures the similarity between network prediction and ->***ground truth*** based on their spatial overlap. In more detail: The Dice score is a metric to quantify the similarity between two binary segmentations X and Y as follows: $DSC = \frac{2\,\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$. In other words, it roughly quantifies the amount of spatial overlap between both segmentations. For multi-label evaluation, binary representations of ground truth and prediction are compared for each class. Besides, the Dice loss is represented by the Dice score in the following way: $DSC_{loss} = 1 - DSC$, since neural networks require ->***loss functions*** instead of score functions.
Ensembling	->***Regularization*** technique to improve performance. In more detail: Instead of one single learning algorithm, multiple neural networks are differently trained, and thus form different predictors to reduce prediction variance. Final results are performed by merging the predictions of all networks.
Epoch	An epoch ends when all training samples have been fed through the network once.
Feature	An individual, measurable property, e.g. glomerular size is a feature of the glomerulus.
Feature map	Spatially arranged features that are generated by applying filters to the convolutional layer input, i.e. the input image or feature map outputs from the prior layer. Example: A convolutional filter has been applied to the left image resulting in a two-dimensional feature map highlighting its edges. 
Ground Truth	Target data we expect the network to predict. We annotate and classify structures according to *our* renal ->***class*** definitions in Supp. Table 2 and consider these annotations and classifications to correspond to reality, thus representing the ground truth. Example: Ground truth image of the left image is shown right.

	
Hyperparameter	Special ->***parameters*** to control e.g. the learning process or architecture of the deep learning model. They are determined by the experimentator before as well as dynamically during training. Examples are the amount of ->***epochs*** or the ->***kernel size.***
Image segmentation	Decomposition of an image into structures of interest. Example: Segmentation of a tubule. 
Instance	A single structure of a class. Example: All tubular instances are differently colored (Image from Supp. Fig. 5, third column). 
Instance normalization	->***Regularization*** technique applied in neural networks. In more detail: In contrast to the widely used batch normalization, instance normalization normalizes each ->***feature map*** independently providing zero mean and unit variance.
Kernel size	Specifies the size of a convolutional filter that is slid over the image.
Loss function	A mathematical function measuring the dissimilarity between network prediction and ->***ground truth***. In more detail: To train a neural network, a (differentiable) mathematical loss function representing a metric to measure the dissimilarity between prediction and ground-truth is required. During training, the network is consecutively optimized (with respect to the loss function) to lower the loss and thus to improve the similarity between prediction and ground-truth.
Negative slope	->***Hyperparameter*** in the mathematical LeakyReLU function. In more detail: The LeakyReLU function is defined as follows: $LeakyReLU(x) = \begin{cases} x, & x \geq 0 \\ negative\_slope * x, & otherwise \end{cases}$ Thus, the $negative\_slope$-hyperparameter specifies the slope of the LeakyReLU function for negative inputs, i.e. $x < 0$. Most commonly, $negative\_slope = 0.01$ is chosen by the experimentator.
Padding	An operation within convolutional layers to artificially enlarge the input data. In more detail: Specifies how much the input data is spatially padded around it. Padding an image with zeros exemplary means that zero values are added around it. Padding is used to counteract shrinkage of the input data caused by convolution.

	Example: without padding with padding 
Parameter	Components of a (deep learning) system that fully define and characterize the system. In more detail: During network training, its trainable parameters are optimized. After training, all network parameters (trainable and non-trainable) are held constant, and the model is then used for prediction computation.
Receptive field	The prediction of a single output pixel only depends on a certain region of the input image. This region represents its receptive field. The size depends on the architecture of the network.
Reduce-On-Plateau	Technique to schedule the learning rate. In more detail: The learning rate represents an important ->***hyperparameter*** in neural networks that controls the speed of learning. This learning rate scheduler reduces the learning rate by a specific factor each time when the validation error has not decreased for a certain number of epochs.
Regularization	Regularization techniques are employed to improve network's generalization, i.e. reducing the error on test data. At the expense of increased training error, such techniques impose particularly designed constraints to the neural network preventing them to solely memorize the training data without having learned the underlying patterns.
ReLU	Stands for *rectified linear unit* and represents a mathematical function defined as follows: $ReLU(x) = \begin{cases} x, & x \geq 0 \\ 0, & otherwise \end{cases}$
Robustness	Describes the extent of input variability (e.g. in tissue morphology, staining, slide thickness, laboratory) an algorithm can cope with. Generally, it is measured by performance evaluation on those variabilities (usually held-out as in the current study).
Stride	An operation within convolutional layers to specify how many pixels the convolutional filter (or: ->***kernel***) is moved when slid over the image. Example: stride of "1" (shift of 1 pixel) stride of "2" (shift of 2 pixels). 
Test-time augmentation	->***Regularization*** technique to improve performance. In more detail: Regularization technique that forwards flipped versions of the input through the network and averages their respectively back-flipped predictions to yield the final prediction. In contrast to ->***ensembling***, just a single network/predictor is used to perform multiple estimations.

	 CNN CNN CNN CNN Backflip Backflip Backflip Average
Transposed convolutions	The conventional convolution provides a many-to-one relationship between input and output, since many input pixels are connected to a single value in the output. In contrast, transposed convolutions make use of a reversed pixel connectivity (in backward direction) providing a one-to-many relationship. Thus, it is designed for image ->***upsampling***. 
Upsampling	Expansion or increase of the spatial resolution of an image. <u>In more detail:</u> Upsampling can be exemplarily performed by pixel interpolation meaning that new pixel values can be estimated between pixels by using their neighborhood, e.g. by averaging neighboring pixels values (ultimately yielding a denser image grid). The picture in ->***transposed convolutions*** exemplarily shows an upsampling of an artificial image.

Supplementary Table 2. Criteria for definition of classes.

Class	Criteria
Full glomerulus	- annotation along Bowman's capsule - if cross section showed urinary (or vascular) pole, glomerulus was encircled in round/oval shape
Glomerular tuft	- subclass of the full glomerulus class - annotation of glomerular tuft only (including podocytes) - for glomerular lesions: extracapillary proliferates (= crescents), parietal epithelial cells which migrated onto the tuft or tip lesions were not included
Tubule	- annotation along, but excluding, the basement membrane
Artery	- annotation of all arteries, including all arterial branches to arterioles - at least one visible vascular smooth muscle cell layer required
Arterial lumen	- subclass of the artery class - annotation of lumen only, excluding also the endothelium
Vein	- annotation of large "white" areas - only the lumen, i.e. the "white" area was annotated - for veins the definition of larger vessels next to arteries with a minimal diameter of 30µm - class includes non-tissue background and renal pelvis

Supplementary Table 3. Quantitative information on ground truth data.

Model / Species	Number of annotated patches / WSI	Train / val / test split of annotated patches	Train / val / test split of partially annotated WSI	Total number of instance annotations						Σ
				full glom.	glom. tuft	tubule	artery	arterial lumen	vein	
Healthy mouse	820 / 41	600 / 60 / 160	30 / 3 / 8	835	804	18536	1107	1416	609	23307
UUO	300 / 15	220 / 20 / 60	11 / 1 / 3	225	221	6795	301	314	177	8033
IRI	300 / 15	220 / 20 / 60	11 / 1 / 3	242	242	7555	354	397	102	8892
Adenine	300 / 15	220 / 20 / 60	11 / 1 / 3	257	256	5995	342	384	111	7345
Alport	300 / 15	220 / 20 / 60	11 / 1 / 3	413	368	7137	361	383	83	8745
NTN	300 / 15	220 / 20 / 60	11 / 1 / 3	247	237	5500	275	295	139	6693
db/db	30 / 3	0 / 0 / 30	0 / 0 / 3	27	27	652	27	22	10	765
Ext. UUO	30 / 3	0 / 0 / 30	0 / 0 / 3	46	43	879	42	27	8	1045
Human	230 / 12	200 / 0 / 30	10 / 0 / 2	123	148	1958	125	145	40	2539
Rat	80 / 8	50 / 0 / 30	5 / 0 / 3	56	59	1372	66	74	27	1654
Pig	80 / 6	50 / 0 / 30	5 / 0 / 1	50	49	900	57	67	23	1146
Marmoset	80 / 8	50 / 0 / 30	5 / 0 / 3	39	39	774	62	70	28	1012
Black bear	80 / 8	50 / 0 / 30	5 / 0 / 3	51	51	1240	85	91	28	1546
Σ	2930 / 164	2100 / 160 / 670	115 / 8 / 41	2611	2544	59293	3204	3685	1385	72722

IRI = ischemia reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction, val = validation

Supplementary Table 4. Architecture of our CNN.

Network Architecture	Output size
Input image layer	640 x 640 x 3
Conv2d(i: 3, o: 32, k: 3, s: 1, p: 1) + IN(o: 32) + LeakyReLU(sl: 0.01)	640 x 640 x 32
Conv2d(i: 32, o: 32, k: 3, s: 1, p: 1) + IN(o: 32) + LeakyReLU(sl: 0.01)	640 x 640 x 32
MaxPool2d(k: 2, s: 2, p: 0)	320 x 320 x 32
Conv2d(i: 32, o: 64, k: 3, s: 1, p: 1) + IN(o: 64) + LeakyReLU(sl: 0.01)	320 x 320 x 64
Conv2d(i: 64, o: 64, k: 3, s: 1, p: 1) + IN(o: 64) + LeakyReLU(sl: 0.01)	320 x 320 x 64
MaxPool2d(k: 2, s: 2, p: 0)	160 x 160 x 64
Conv2d(i: 64, o: 128, k: 3, s: 1, p: 1) + IN(o: 128) + LeakyReLU(sl: 0.01)	160 x 160 x 128
Conv2d(i: 128, o: 128, k: 3, s: 1, p: 1) + IN(o: 128) + LeakyReLU(sl: 0.01)	160 x 160 x 128
MaxPool2d(k: 2, s: 2, p: 0)	80 x 80 x 128
Conv2d(i: 128, o: 256, k: 3, s: 1, p: 1) + IN(o: 256) + LeakyReLU(sl: 0.01)	80 x 80 x 256
Conv2d(i: 256, o: 256, k: 3, s: 1, p: 1) + IN(o: 256) + LeakyReLU(sl: 0.01)	80 x 80 x 256
MaxPool2d(k: 2, s: 2, p: 0)	40 x 40 x 256
Conv2d(i: 256, o: 512, k: 3, s: 1, p: 1) + IN(o: 512) + LeakyReLU(sl: 0.01)	40 x 40 x 512
Conv2d(i: 512, o: 512, k: 3, s: 1, p: 1) + IN(o: 512) + LeakyReLU(sl: 0.01)	40 x 40 x 512
MaxPool2d(k: 2, s: 2, p: 0)	20 x 20 x 512
Conv2d(i: 512, o: 1024, k: 3, s: 1, p: 1) + IN(o: 1024) + LeakyReLU(sl: 0.01)	20 x 20 x 1024
Conv2d(i: 1024, o: 1024, k: 3, s: 1, p: 1) + IN(o: 1024) + LeakyReLU(sl: 0.01)	20 x 20 x 1024
ConvTranspose2d(i: 1024, o: 1024, k: 2, s: 2)	40 x 40 x 1024
Conv2d(i: 1536, o: 512, k: 3, s: 1, p: 0) + IN(o: 512) + LeakyReLU(sl: 0.01)	38 x 38 x 512
Conv2d(i: 512, o: 512, k: 3, s: 1, p: 0) + IN(o: 512) + LeakyReLU(sl: 0.01)	36 x 36 x 512
ConvTranspose2d(i: 512, o: 512, k: 2, s: 2)	72 x 72 x 512
Conv2d(i: 768, o: 256, k: 3, s: 1, p: 0) + IN(o: 256) + LeakyReLU(sl: 0.01)	70 x 70 x 256
Conv2d(i: 256, o: 256, k: 3, s: 1, p: 0) + IN(o: 256) + LeakyReLU(sl: 0.01)	68 x 68 x 256
ConvTranspose2d(i: 256, o: 256, k: 2, s: 2)	136 x 136 x 256
Conv2d(i: 384, o: 128, k: 3, s: 1, p: 0) + IN(o: 128) + LeakyReLU(sl: 0.01)	134 x 134 x 128
Conv2d(i: 128, o: 128, k: 3, s: 1, p: 0) + IN(o: 128) + LeakyReLU(sl: 0.01)	132 x 132 x 128
ConvTranspose2d(i: 128, o: 128, k: 2, s: 2)	264 x 264 x 128
Conv2d(i: 192, o: 64, k: 3, s: 1, p: 0) + IN(o: 64) + LeakyReLU(sl: 0.01)	262 x 262 x 64
Conv2d(i: 64, o: 64, k: 3, s: 1, p: 0) + IN(o: 64) + LeakyReLU(sl: 0.01)	260 x 260 x 64
ConvTranspose2d(i: 64, o: 64, k: 2, s: 2)	520 x 520 x 64
Conv2d(i: 96, o: 32, k: 3, s: 1, p: 0) + IN(o: 32) + LeakyReLU(sl: 0.01)	518 x 518 x 32
Conv2d(i: 32, o: 32, k: 3, s: 1, p: 0) + IN(o: 32) + LeakyReLU(sl: 0.01)	516 x 516 x 32
Conv2d(i: 32, o: 8, k: 1, s: 1, p: 0)	516 x 516 x 8

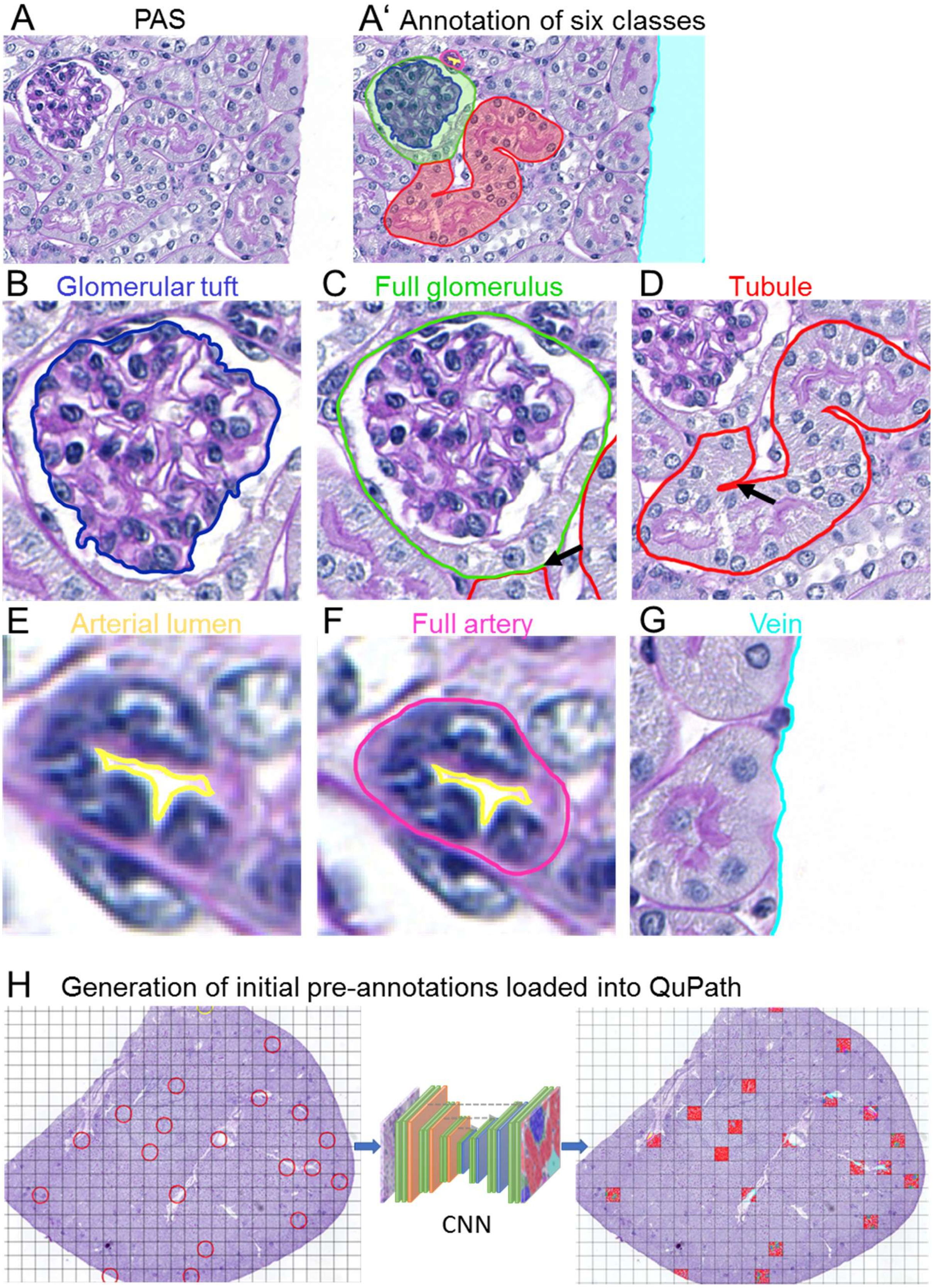
Conv2d = two-dimensional convolutional layer, IN = instance normalization, i = #input layers, o = #output layers, k = kernel size, s = stride, p = padding, sl = negative slope

Supplementary Table 5. Performance comparison of our model, its unmodified variant vanilla u-net, and state-of-the-art context-encoder.

Shown are mean object-level dice scores for our model / the unmodified variant vanilla u-net / state-of-the-art context-encoder. The highest Score is marked in bold. * p < 0.05 vs. vanilla u-net and ° p < 0.05 vs. context-encoder.

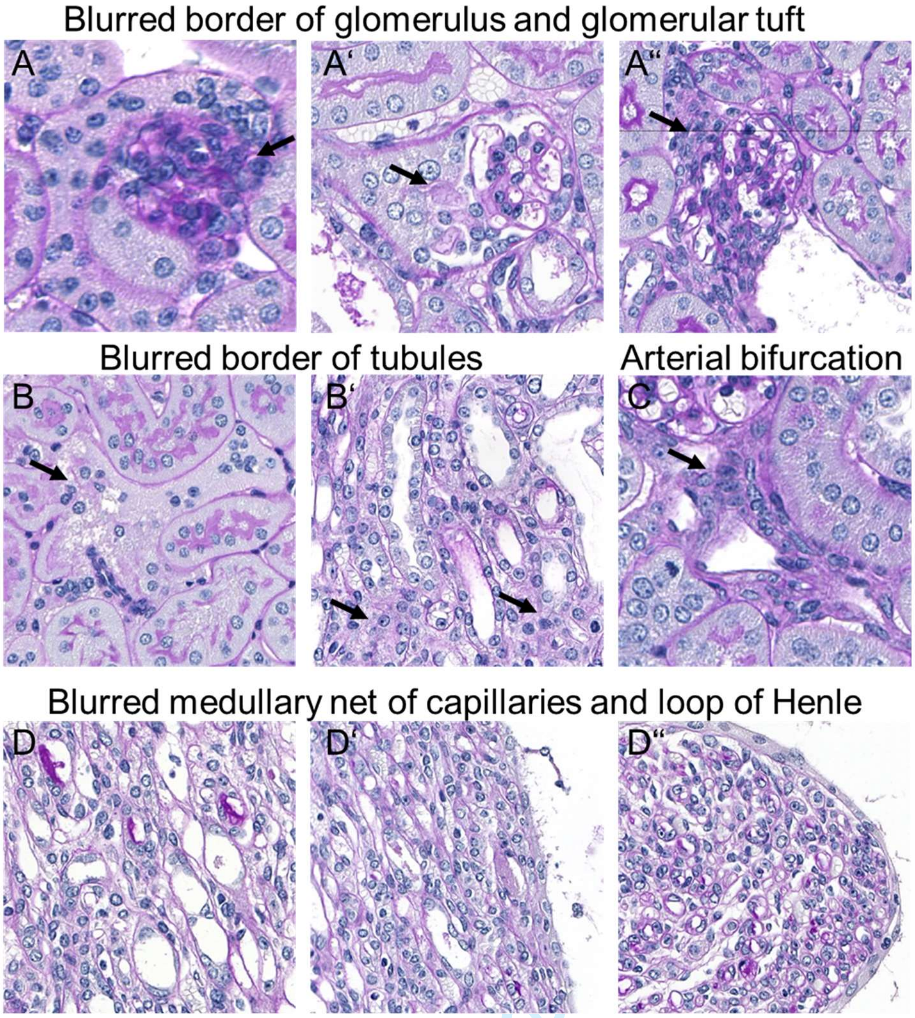
Mouse Model	Segmentation performance of our model / vanilla u-net / context-encoder					
	full glomerulus	glomerular tuft	tubule	artery	arterial lumen	vein
Healthy	**96.5** / 95.6 / 96.2	**93.8** / **93.8** / 93.5	**93.3** / 92.9 / 93.0	**88.1** / 87.4 / 87.8	80.3 / 80.0 / **80.6**	**94.3** / 88.9 / 92.0
UUO	**97.5** / 95.2 / 95.3	**95.6** / 93.9 / 94.5	90.8 / 90.8 / **91.3**	82.3 / 81.2 / **82.6**	**75.0** / 72.9 / 73.7	**97.6** / 95.4 / 94.6
IRI	96.0 / **97.7** / 95.7	**95.4** / 94.7 / 94.4	**90.2** / 89.1 / 89.9	**79.1** / 74.7 / 74.2	**73.5** / 62.3 / 61.7	**97.7** / 86.7 / 87.0
Adenine	**98.8** / 94.1 / 98.5	**97.2** / 94.1 / 97.1	**93.0** / 92.0 / 92.8	**87.9** / 83.3 / 83.2	**80.9** / 72.7 / 76.9	93.6 / 87.6 / **96.7**
Alport	94.7 / 95.5 / **96.3**	**91.3** / 86.4 / 87.6	**90.6** / 89.7 / 89.3	**80.3** / 74.2 / 72.0	**81.1** / 69.9 / 65.5	**89.2** / 83.2 / 81.7
NTN	95.5 / 91.5 / **96.3**	**94.8** / 93.9 / 93.9	**93.2** / 92.5 / 92.9	**86.8** / 82.7 / 83.9	78.2 / 73.9 / **79.1**	92.8 / 91.8 / **95.4**
ø	**96.4*** / 94.0 / 96.3	**94.2*** / 92.6 / 93.0	**92.0*** / 91.4 / 91.7	**85.3*°** / 82.8 / 82.9	**79.1*°** / 75.9 / 76.1	**94.3*** / 90.4 / 92.7

IRI = ischemia reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction

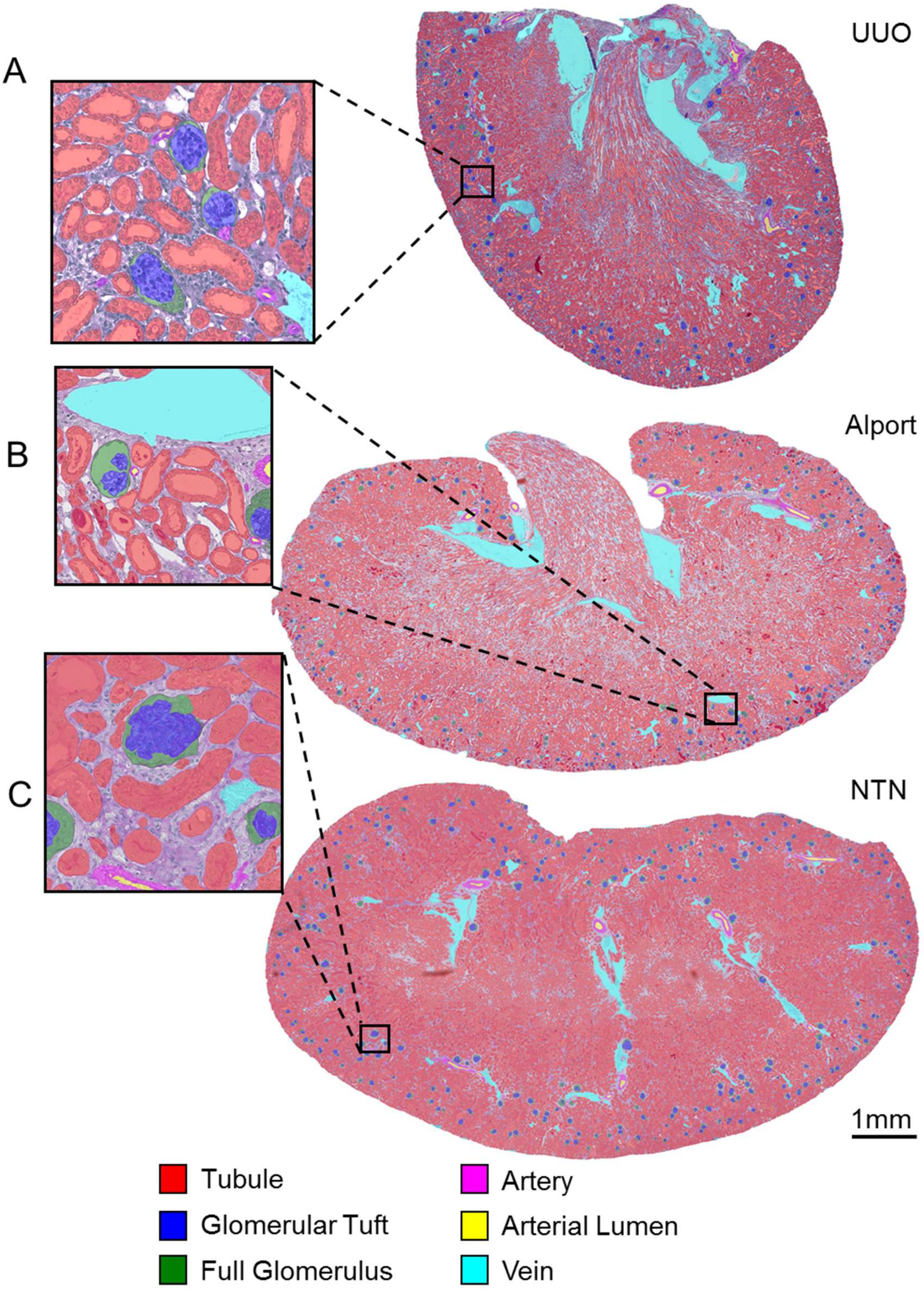

A
PAS
A‘ Annotation of six classes
B Glomerular tuft
C Full glomerulus
D Tubule
E Arterial lumen
F Full artery
G Vein
H Generation of initial pre-annotations loaded into QuPath
CNN

Supp. Fig. 1. Annotation procedure.

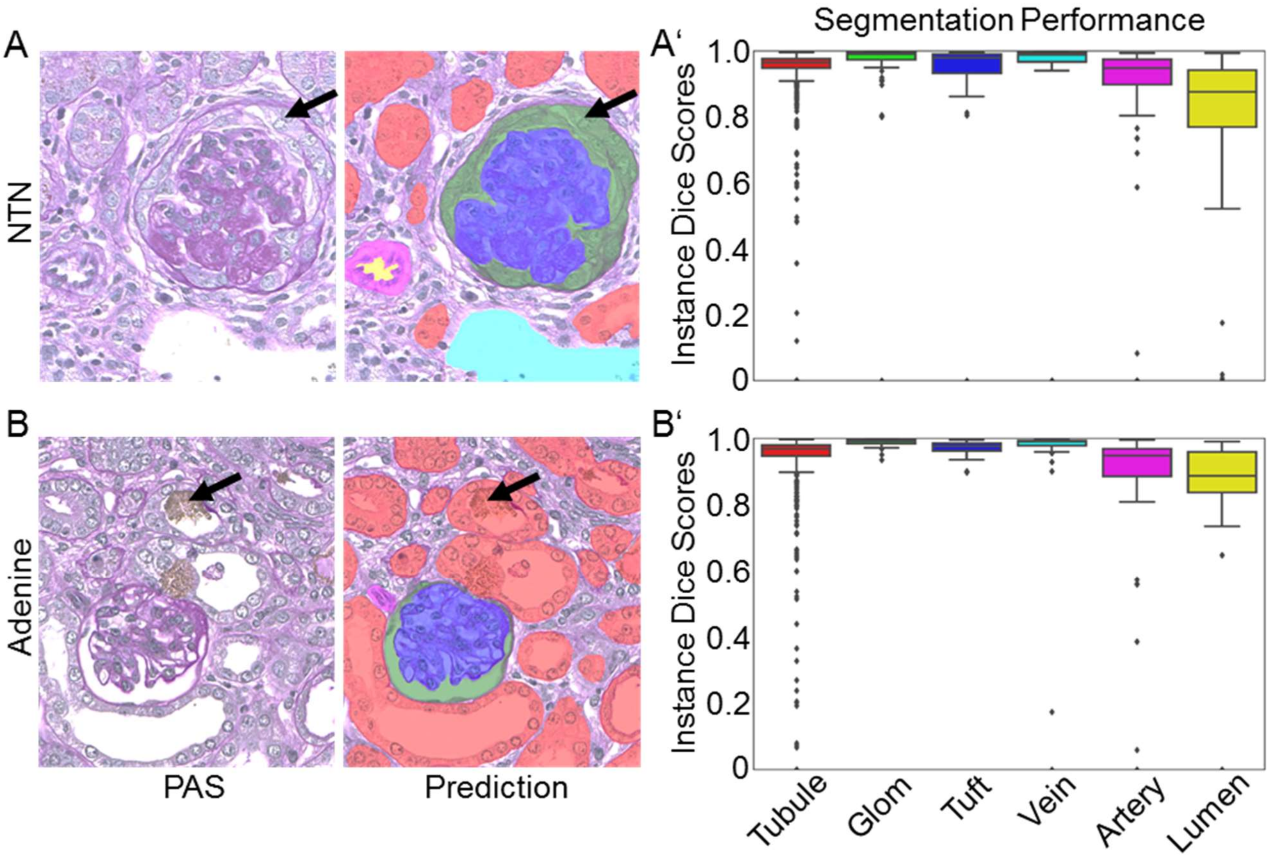
A representative picture of a PAS stained mouse kidney section (A) and an overlay with manual annotations for six classes (A’). The annotation of the “glomerular tuft” (blue (B)) included the capillary tuft, the mesangium and podocytes. A “full glomerulus” (green (C)) was annotated along bowman’s capsule and included the tuft, bowman’s space and parietal epithelial cells. The glomerular tuft was always a subclass of the full glomerulus. A full glomerulus always had a round or oval shape, this determined the separation from the proximal tubule (arrow). Tubules (red (D) were annotated along (but excluding) the tubular basement membrane, tangentially cut tubules without cytoplasm were excluded. The “arterial lumen” (yellow (D)) was always a subclass of the “artery” class (magenta (F)). Veins, background and renal pelvis were big “white” areas without tissue (cyan (G)). From the first manual annotations, we predicted initial pre-annotations for 20 patches per WSI and loaded them into Qupath for manual corrections facilitating annotation effort (H).


Blurred border of glomerulus and glomerular tuft
A
A‘
A“
Blurred border of tubules
B
B‘
Arterial bifurcation
C
Blurred medullary net of capillaries and loop of Henle
D
D‘
D“

Supp. Fig. 2. Challenging morphology for manual and automated annotations. (A-A’’) show examples of glomeruli in PAS stained murine kidney sections. On a sectional plane close to the vascular or urinary pole it was difficult to discriminate between glomerular tuft and arterioles (arrow, A), or the glomerular tuft and parietal epithelial cells or tubular epithelial cells (arrows, A’,A’’). Sometimes the tubular basement membrane appeared discontinuous (arrows in B, B’). The distinction of medial layers of arteries was harder when vessels run side by side (arrow, C). (D-D’’) show medulla of murine kidneys with the network of capillaries and the tubular system, which in some cases was not easy to discriminate.

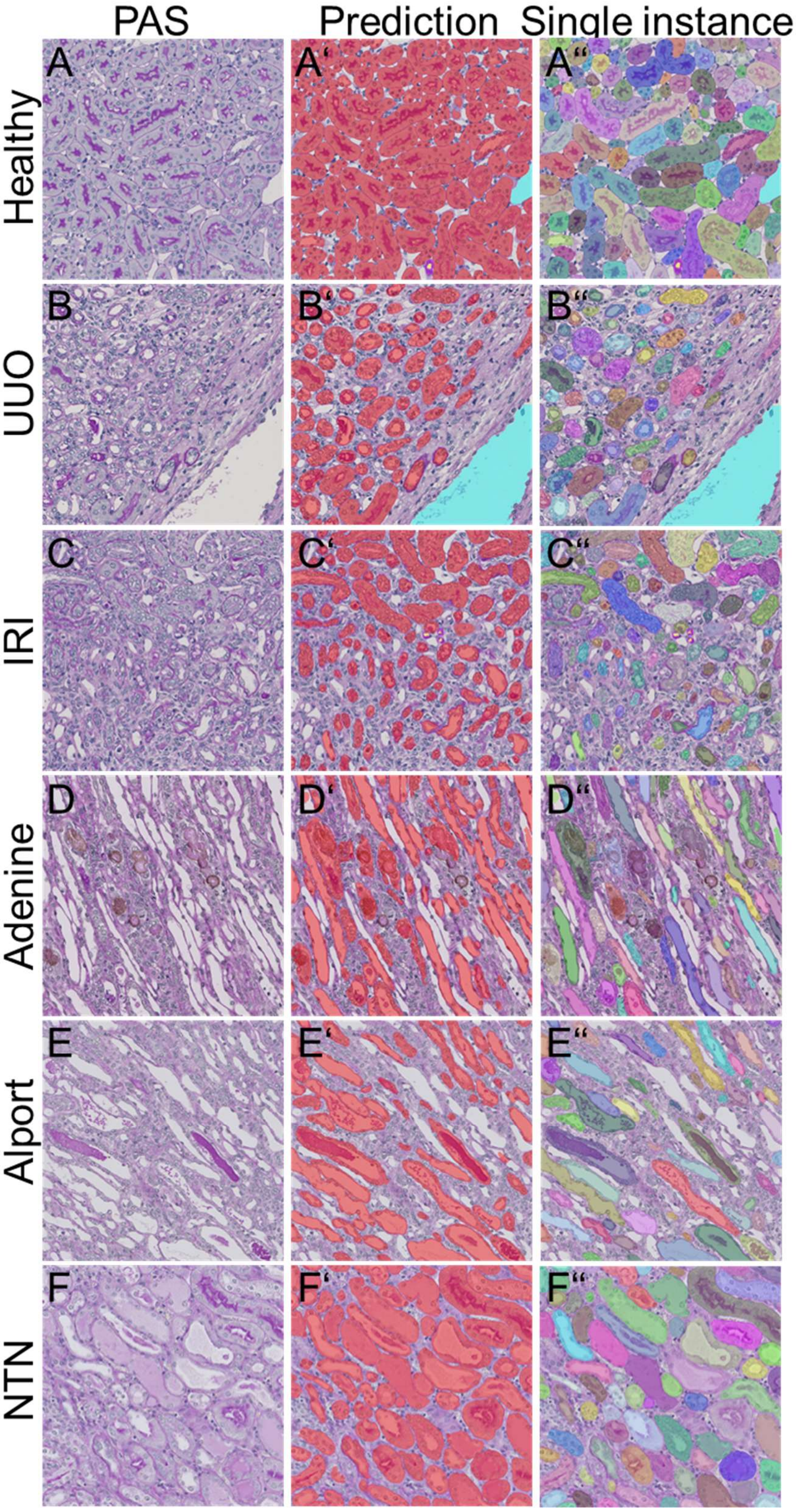


Supp. Fig. 3. Segmentation of WSI of UUO, Alport and NTN kidneys.
CNN generated segmentation predictions on a whole slide image (WSI) of an UUO (A), Alport (B) and NTN (C) mouse kidney. All six classes, were precisely segmented. NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.



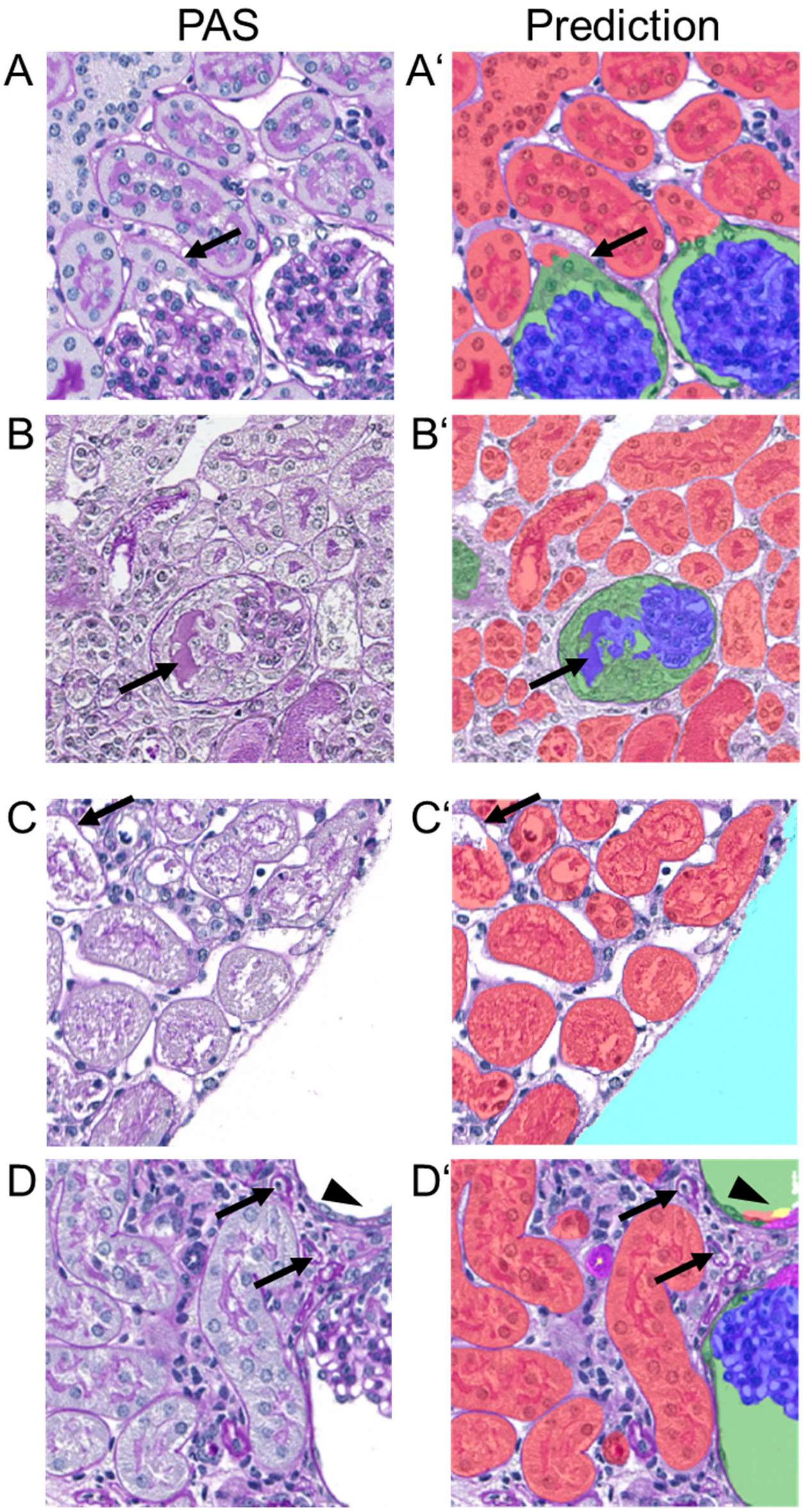
Supp. Fig. 4. Quantitative segmentation performance in murine NTN and adenine kidneys.

Representative PAS pictures and the corresponding segmentation prediction generated by our CNN for a murine NTN (A) and adenine kidney (B). Instance segmentation accuracy is shown by dice scores for each class in both models (A'-B'). Data are presented in Box plots with median, quartiles and whiskers. NTN = nephrotoxic nephropathy.



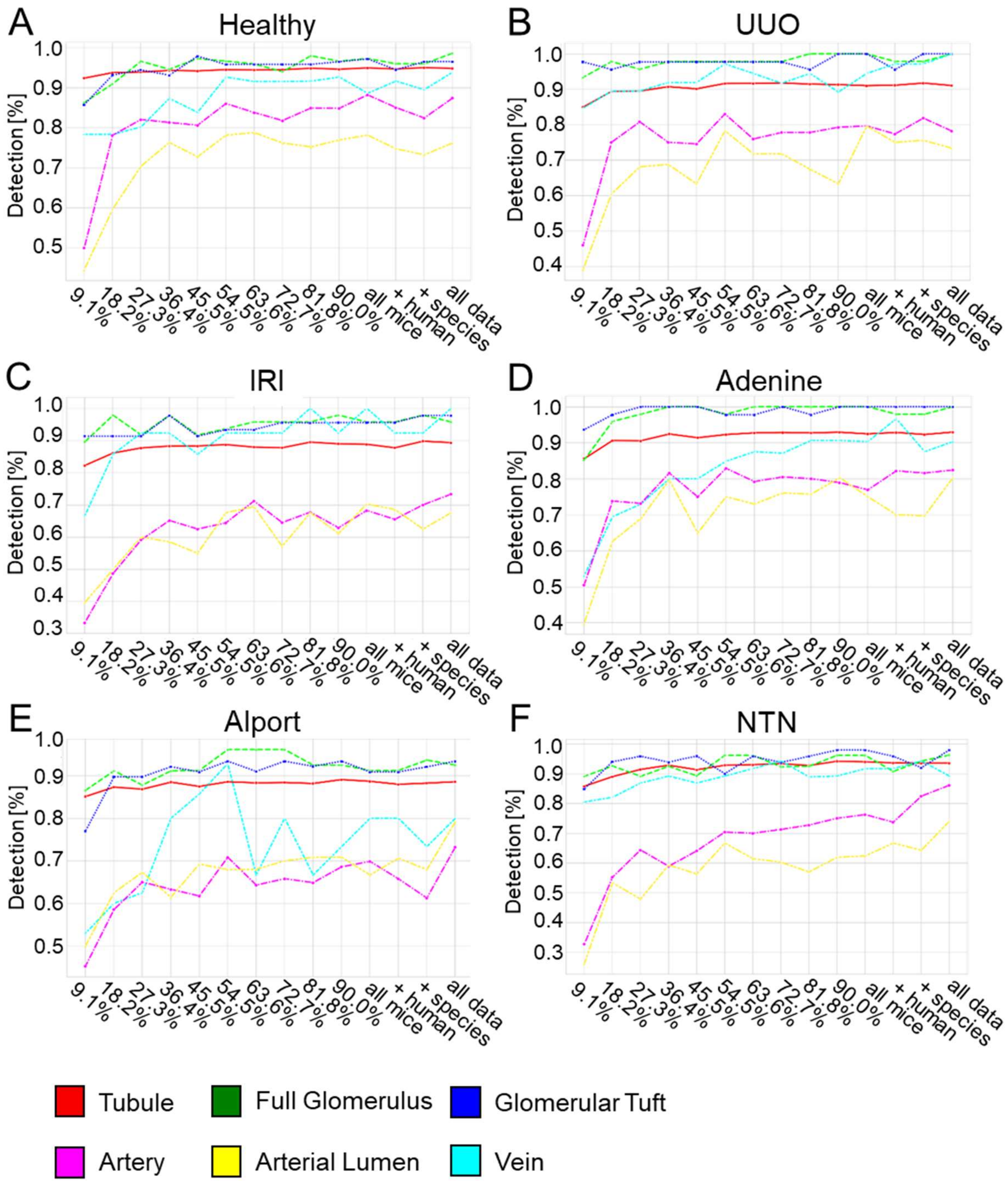
Supp. Fig. 5. Automated segmentation in the medulla of murine kidney sections. Representative PAS pictures and corresponding overlays with segmentation predictions showing either the different classes or every single instances for the medulla of murine healthy (A-A’’), UUO (B-B’’), IRI (C-C’’), adenine (D-D’’), Alport (E-E’’) and NTN (F-F’’) kidneys.

IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.



Supp. Fig. 6. Examples of missclassifications.

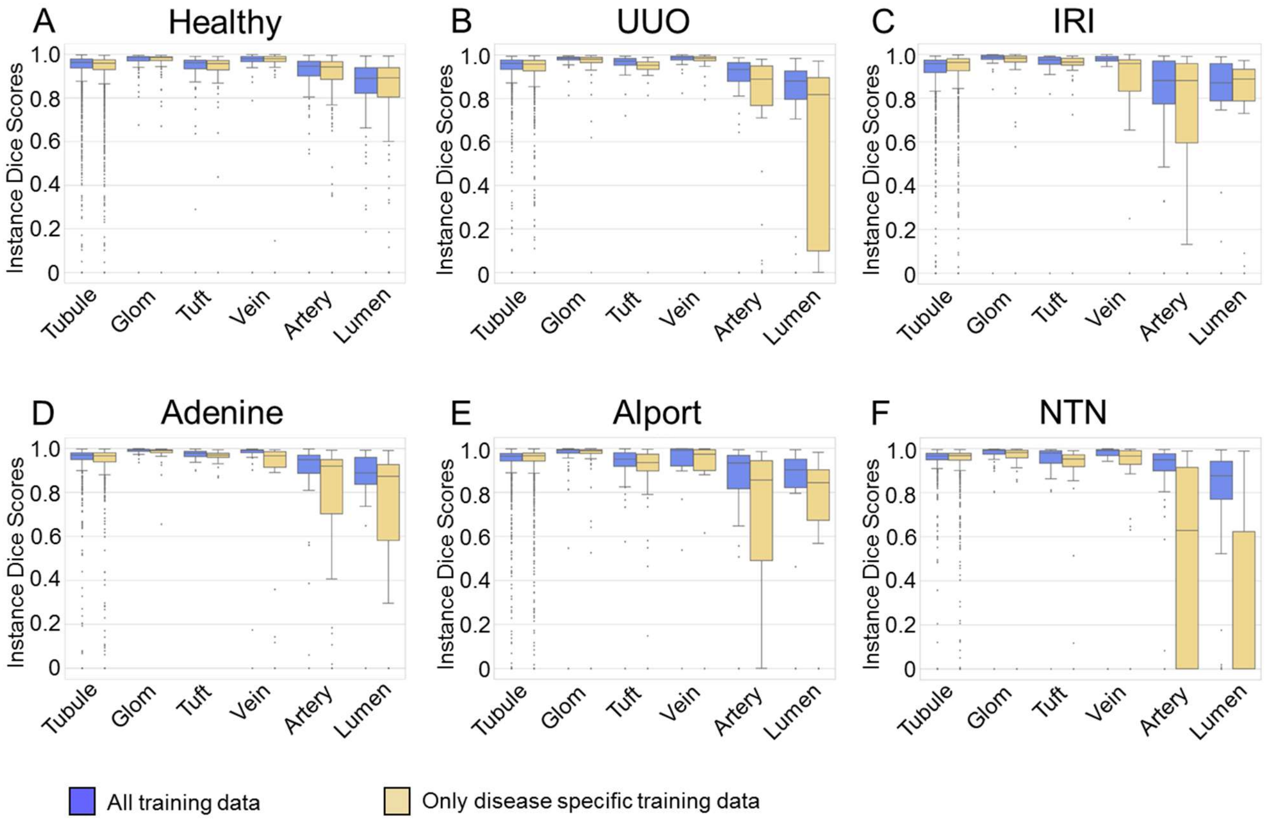
PAS photographs and prediction overlays show an incorrect separation of a “full glomerulus” and the connected proximal “tubule” (arrow in A, A’), a glomerular tuft that was inaccurately segmented with projections into the crescent (arrow in B, B’) and an incompletely segmented tubule due to extensive necrosis (arrow in C,C’). Another example shows a strongly dilated tubule which is was incorrectly classified as full glomerulus and arterial lumen (arrowheads in D,D’) and missing segmentations of atrophic tubules (arrows in D,D’).



Supp. Fig. 7. Relation between amount of training data and detection performance.

The detection performance for all six classes in healthy (A), UUO (B), IRI (C), adenine (D), Alport (E) and NTN (F) was plotted against the amount of total data used for CNN training.

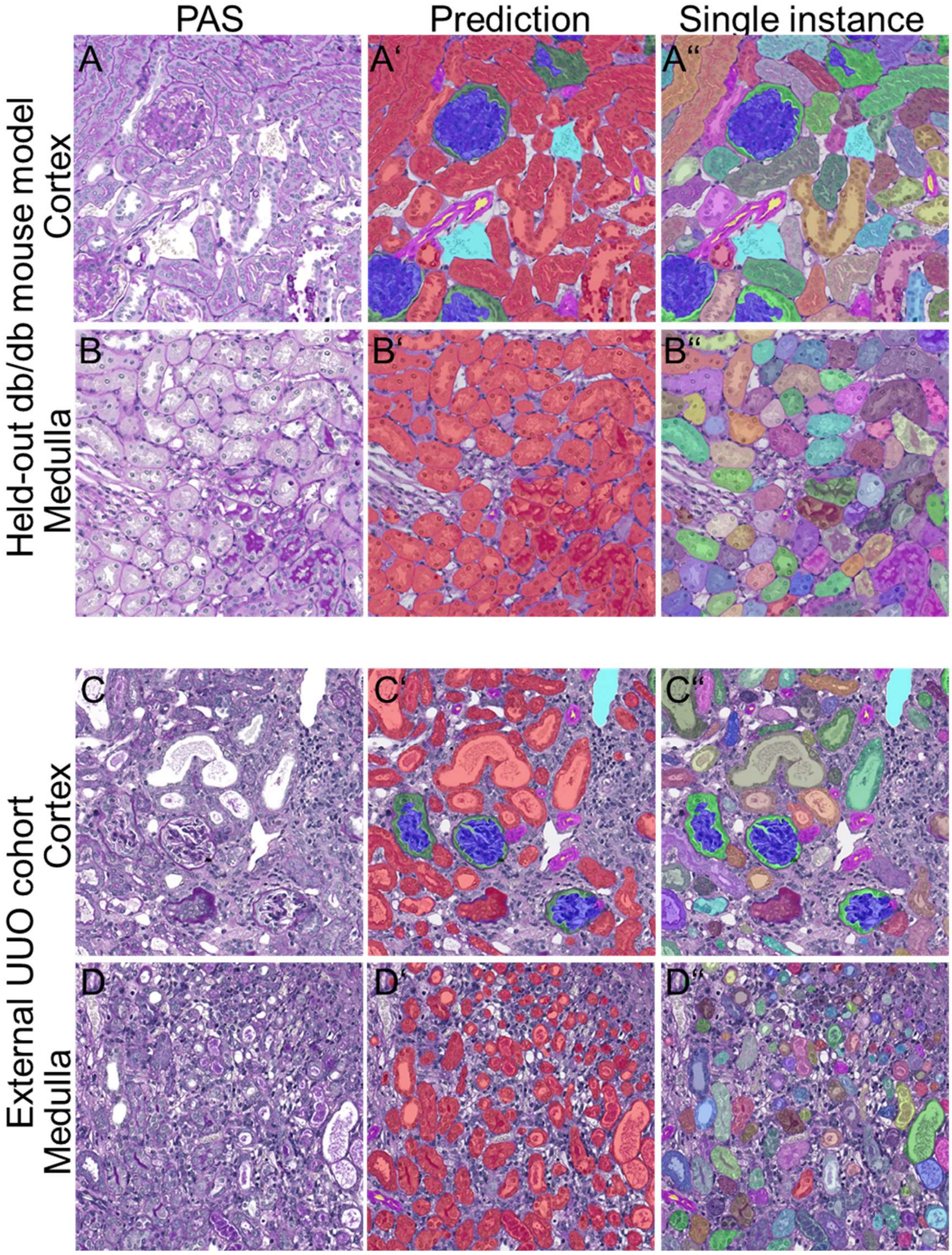
IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.



Supp. Fig. 8. Comparison between our full CNN and its variants independently trained on single models.

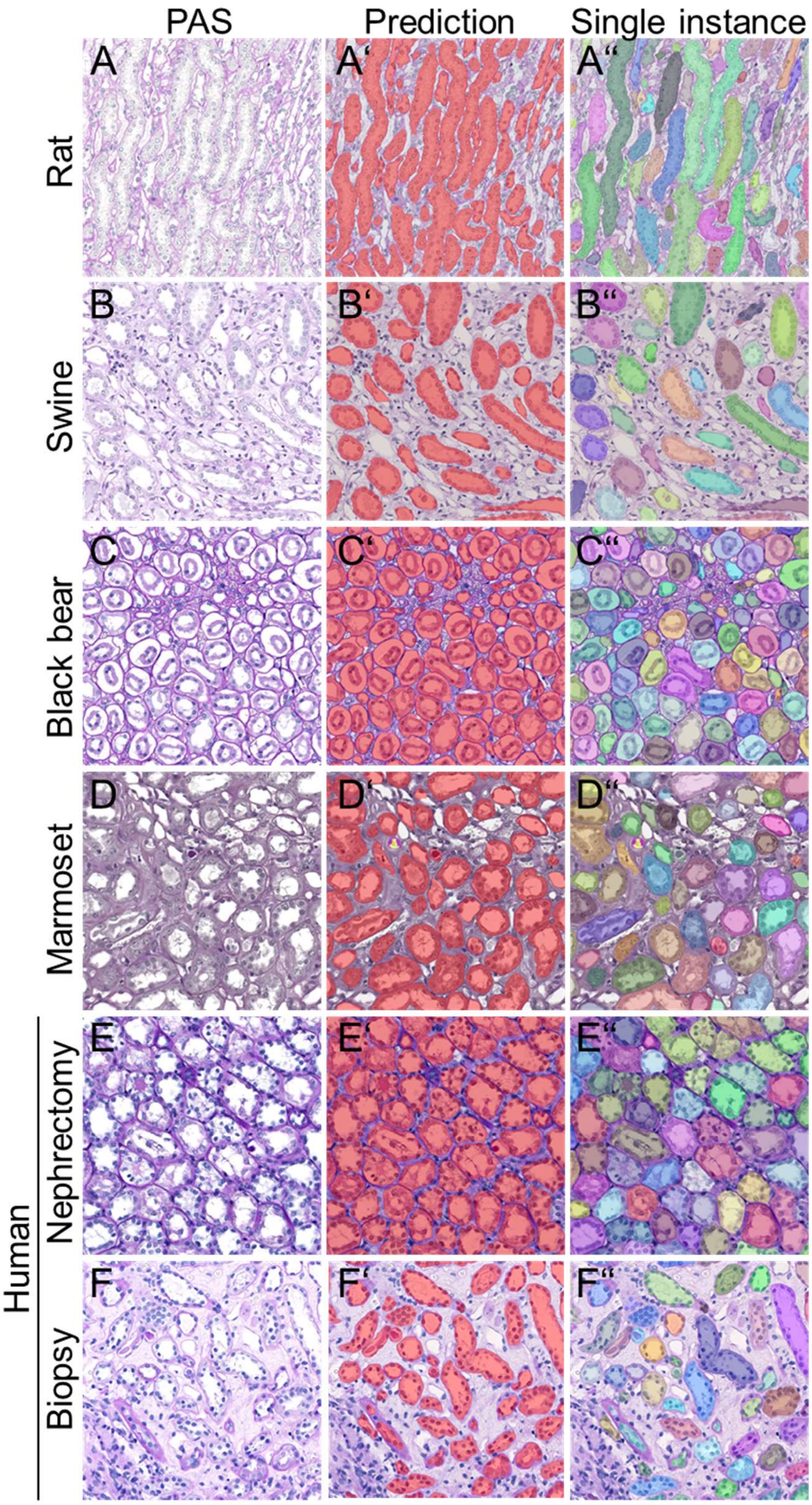
(A) Segmentation performance shown as instance dice scores for all six classes was compared on our healthy kidney test data between our full CNN trained on all training data (blue) and its variant that has been solely trained with data from healthy kidneys (yellow). (B) The same comparison is shown for the UUO, in which the network variant was exclusively trained with annotations from UUO kidneys. Analogously, analyses are performed for IRI (C), adenine (D), Alport (E) and NTN (F).

Data are presented in Box plots with median, quartiles and whiskers. IRI = ischemia-reperfusion injury, NTN = nephrotoxic nephropathy, UUO = unilateral ureteral obstruction.

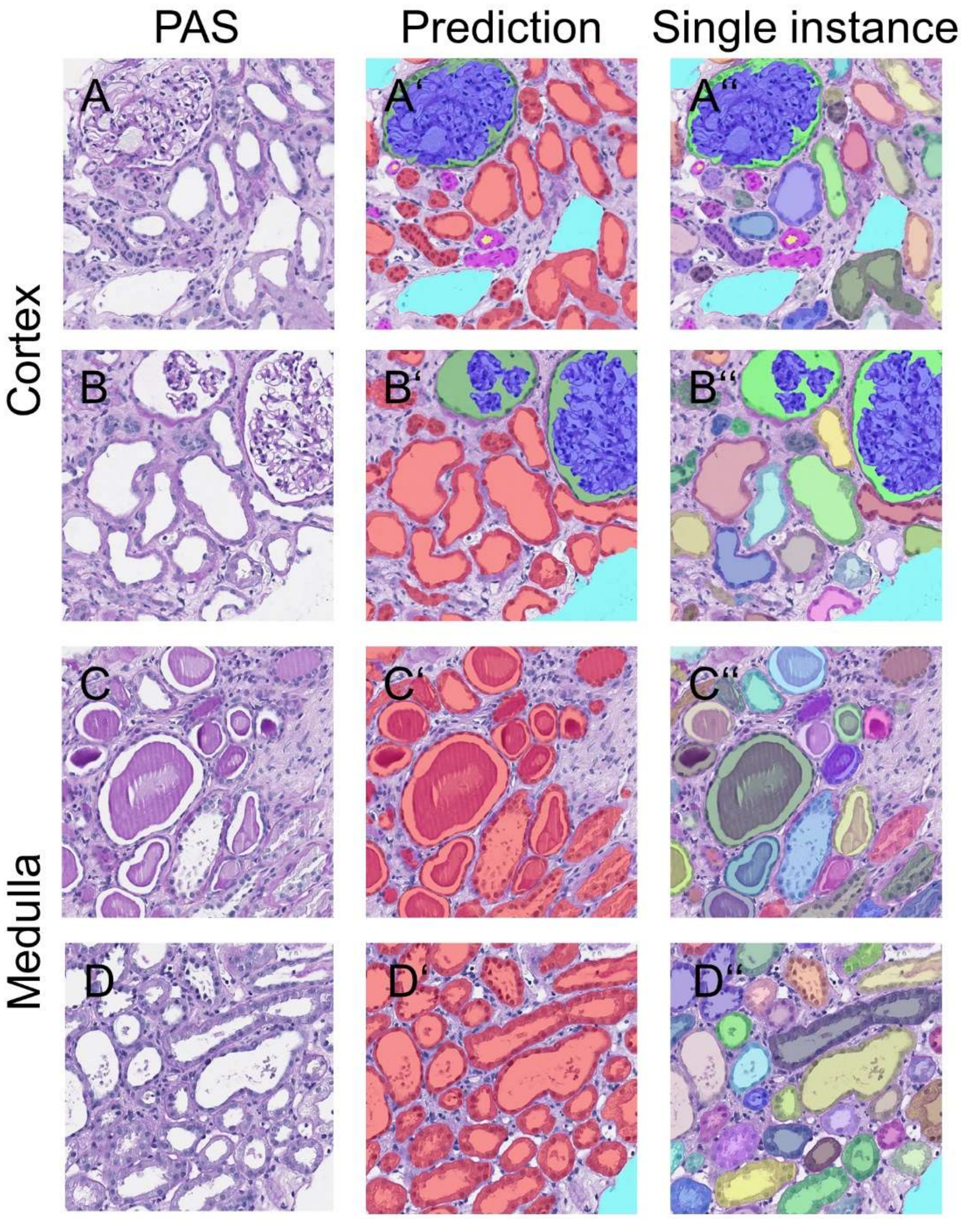


Supp. Fig. 9. Segmentation of non-trained and external murine kidney slides. Representative pictures show segmentation results for cortex (A-A'') and medulla (B-B'') for kidneys from db/db mice fed with high fat western diet. Predictions (A', B') depict different classes, while A'' and B'' display segmentation on single instance level. The CNN also accurately segments cortex (C-C'') and medulla (D-D'') from PAS slides of an external UUO cohort. Predictions (C', D') depict different classes, while C'' and D'' display segmentation on single instance level.

UUO = unilateral ureteral obstruction.



Supp. Fig. 10. Automated segmentation of renal medulla in different species. Representative PAS pictures and the corresponding overlays for segmentation predictions showing either the different classes or every single instance for the medulla of rat (A-A''), pig (B-B''), black bear (C-C''), marmoset (D-D'') and human (E-F'') kidneys. Segmentation is accurate on human nephrectomy (E-E'') as well as on biopsy specimens (F-F'').



Supp. Fig. 11. Automated segmentation of human biopsies presenting with acute tubular damage. Representative PAS-pictures and the respective segmentation prediction overlays from cortex (A-B’’) and medulla (C-D’’) of human biopsies with acute tubular damage.