

Autonomous Quality and Hallucination Assessment for Virtual Tissue Staining and Digital Pathology

Luzhe Huang^{1,2,3,7}, Yuzhu Li^{1,2,3,7}, Nir Pillar^{1,2,3}, Tal Keidar Haran⁴, William Dean Wallace⁵,
Aydogan Ozcan^{1,2,3,6,*}

¹ Electrical and Computer Engineering Department, University of California, Los Angeles, CA 90095, USA.

² Bioengineering Department, University of California, Los Angeles, CA 90095, USA.

³ California NanoSystems Institute (CNSI), University of California, Los Angeles, CA 90095, USA.

⁴ Department of Pathology, Hadassah Hebrew University Medical Center, Jerusalem 91120, Israel.

⁵ Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA.

⁶ Department of Surgery, University of California, Los Angeles, CA 90095, USA.

⁷ These authors contributed equally: Luzhe Huang, Yuzhu Li.

* ozcan@ucla.edu

Abstract

Histopathological staining of human tissue slides is essential in the diagnosis of various diseases. The recent advances in virtual tissue staining technologies using AI alleviate some of the costly and tedious steps involved in the traditional histochemical staining process, permitting multiplexed rapid staining of label-free tissue without using staining reagents, while also preserving the tissue block. However, potential hallucinations and artifacts in these virtually stained tissue images pose concerns, especially for the clinical utility of these AI-driven approaches. Quality assessment of histology images is, in general, performed by human experts, which can be subjective and depends on the training level of the expert. Here, we present an autonomous quality and hallucination assessment method (termed AQuA), mainly designed for virtual tissue staining, while also being applicable to histochemical staining. Leveraging a novel architecture, AQuA autonomously achieves 99.8% accuracy when detecting acceptable and unacceptable virtually stained tissue images without access to ground truth, also presenting an agreement of 98.5% with the manual assessments made by a group of board-certified pathologists. Besides, AQuA achieves super-human performance in identifying realistic-looking, virtually stained hallucinatory images that would normally mislead human diagnosticians by deceiving them into diagnosing patients that never existed. We further demonstrate the wide adaptability of AQuA across various virtually and histochemically stained human tissue images and showcase its strong external generalization to detect unseen hallucination patterns of virtual staining network models as well as artifacts observed in the traditional histochemical staining workflow. This framework creates new opportunities to greatly

enhance the reliability of virtual staining and will provide quality assurance for various image generation and transformation tasks in the fields of digital pathology and computational imaging.

Introduction

Histopathological examination is considered the gold standard diagnostic procedure for most lesions of the human body. Prior to microscopic evaluation, extracted tissue samples undergo multistep processing that includes fixation, grossing, embedding, sectioning, staining, mounting and cover-slipping. Artifacts, defined as abnormal structures or features produced by the processing of tissue samples¹, are common in histopathology. These artifacts alter tissue morphology, interfere with the diagnostic process, and can even preclude reaching a definite diagnosis. To minimize tissue slides heavily affected by artifacts reaching the pathologist's table, histotechnicians and lab specialists constantly perform manual quality assurance after each processing step. These inspections include comparing the tissue slide gross morphology to the tissue block morphology, histochemical stain (**HS**) color quality assessment, and checking the presence of tissue folds, holes, and other external material masking the tissue.

Over the last decade, pathology has undergone significant technological developments, with the inception of digital and computational pathology. The introduction of deep learning models allows for the analysis and stratification of histopathology images in ways that were not available before^{2,3}. Recently, artificial intelligence-based techniques have created new opportunities to transform traditional century-old histochemical staining

methods. Deep learning models have been developed to virtually replicate the images of chemically stained slides using only the microscopic images of unlabeled/label-free tissue samples⁴⁻¹⁷, eliminating the time-consuming, laborious, and costly chemical staining procedures. These methods, which digitally generate histological stains using trained deep neural networks, are collectively termed as virtual staining (**VS**) and have been explored for both label-free staining^{4-7,9,14-16,18-23} and stain-to-stain transformations²⁴⁻³⁰. Although these VS techniques offer major benefits such as reduced stain variability, high-speed, cost-effectiveness, reduced labor and multiplexing, they also introduce risks of potential hallucinations and artifacts in the generated virtually stained images. One known pitfall of generative models is their susceptibility to hallucinate by producing information/images that are not based on factual data or reality. These hallucinations can range from subtle structural and/or color inconsistencies to entirely fabricated content. In the context of virtual histopathology and VS, these hallucinations can be categorized into two main areas: (1) 'technical and unrealistic' hallucinations, and (2) 'realistic' hallucinations. 'Unrealistic' hallucinations introduced by technical issues include blurred areas, folded tissue regions or aberrantly stained areas, which are similar to traditional artifacts that are frequently seen in glass slide-based pathology; these forms of hallucinations and artifacts would normally be spotted by an expert histotechnologist or pathologist examining the image. In the second category, however, such hallucinations would appear as 'realistic', such as e.g. the replacement and/or addition of some tissue components with imaginary ones. These realistic hallucinations might mislead pathologists, deceiving them to diagnose features that do not appear in the actual tissue specimen, although looking realistic and believable

from the perspective of tissue staining quality. These types of realistic hallucinations may result in a misinterpretation of tissue features, misleading the diagnosis process (e.g., hallucinated tumor cells appearing inside a benign tissue section), altering tumor grade (e.g., hallucinated mitotic figures within a tumor), and affecting the predicted response to treatment (e.g., hallucinated lymphocytes in the tumor microenvironment), among many other possibilities. While several deep learning-based tools have been developed for detecting ‘technical’ or ‘unrealistic’ artifacts in standard pathology slides³¹⁻³⁴ that are histochemically stained, their accuracy vary and constant human-in-the-loop inspections are required for quality assurance. Furthermore, none of these tools have been trained or evaluated specifically for their ability to detect hallucinations induced by virtual staining models. Even more critically, there is currently no available tool to evaluate and detect virtually stained tissue images that present realistic and believable hallucinations, which may significantly influence tissue interpretation and deceive pathologists into diagnosing patients that never existed.

Here, we present AQuA, an automated image quality and hallucination assessment framework as the first of its kind computational tool that is specifically designed for recognizing morphological artifacts and hallucinations created by generative VS models (as depicted in Fig. 1a), including technical/unrealistic and, more importantly, realistic hallucinations. AQuA-Net was trained, validated and tested using a set of virtually stained Hematoxylin & Eosin (H&E) tissue slides taken from human renal and lung biopsies. Using a unique data assembly methodology (described in the Methods section), a wide variety of morphological hallucinations and error modes within VS images were introduced into

kidney and lung tissue datasets. Leveraging a novel architecture design (Fig. 1b), AQuA-Net automatically detects these morphological hallucinations and low-quality VS images without the need for any ground truth information. By introducing a label-free autofluorescence (AF) based virtual tissue staining model (i.e., VS: AF \rightarrow H&E) and its corresponding reverse image transformation, i.e., virtual autofluorescence model (VAF: H&E \rightarrow AF), we employed virtual iterations between the H&E and AF domains, which were utilized as the input of AQuA-Net to autonomously detect artifacts and hallucinations in virtually stained H&E images, without access to the ground truth histochemical H&E images. Blindly tested on human kidney and lung tissue samples from new patients, AQuA successfully classified each virtually stained image as having an acceptable or unacceptable stain quality with high specificity and sensitivity; it also showcased an exceptional external generalization to new types and styles of hallucinations and artifacts generated by poorly trained VS models that were never used before. In a blinded comparison against a group of board-certified pathologists, AQuA scored super-human accuracy, especially detecting numerous realistic hallucinations that were scored as good stained images by expert pathologists, who would normally be misled into diagnosing patients using these hallucinated tissue images. Additionally, we adapted AQuA to discriminating good and bad histochemically stained (HS), standard H&E images and demonstrated its superior accuracy to detect staining artifacts that routinely appear in the traditional clinical workflow, also beating various analytical hand-crafted image quality assessment metrics.

Our analyses revealed AQuA’s potential as a fully autonomous, widely generalizable virtual tissue staining quality and hallucination assessment framework with super-human performance. We believe that AQuA can be extended beyond the realm of virtual H&E staining to virtual immunohistochemistry and immunofluorescence staining, which pose similar stain-quality assessment challenges. This autonomous hallucination detection and uncertainty estimation tool for deep neural networks can substantially save time and professional labor conventionally induced by human supervision, and simultaneously mitigate potential errors caused by subjective interpretations, which would be valuable for enhancing the reliability of a plethora of deep learning-based generative models and image translation methods.

Results

Image-level autonomous quality and hallucination assessment of kidney tissue virtual staining

We first validated our method on virtual H&E staining of human kidney tissue samples. Starting with the establishment of good- and poor-staining models, we trained one VS model using paired AF-HS images (see the Methods section) until it converged. With the convergence thresholds determined on both the validation loss and training epochs, early stopped checkpoints with fewer epochs and higher validation losses were defined as poor VS models, and the other checkpoints with more epochs and lower validation losses were regarded as good VS models. Figure 2(a) illustrates negative and positive VS images generated by good-staining models (free of artifacts and hallucinations – i.e., **negative**) and

poor-staining models (artifacts and hallucinations detected – i.e., **positive**), respectively. The registered HS images are also shown here, representing the ground truth – although they are not used in the inference phase since AQUA is an autonomous framework. As pointed by the black arrows, poor VS models produce hallucinations, which are hard to distinguish without the histochemical ground truth reference. In fact, even with the ground truth HS images, it is not easy to discriminate hallucinated images from high-quality VS images with sufficient accuracy as presented in Fig. 2b. Since VS models were trained under the generative adversarial network (GAN)³⁵ framework, traditional structural quality assessment metrics are incapable of distinguishing semantic hallucination. In terms of commonly used structural image quality metrics, including mean square error (MSE), Pearson correlation coefficient (PCC) and peak signal-to-noise ratio (PSNR), the distributions of the negative and positive VS images present strong overlap and cannot be well separated with simple thresholding. More importantly, in VS workflow, the histochemically stained images (the ground truth) would be unavailable, and therefore such image quality metrics cannot be used in the deployment of the VS workflow. In contrast, AQUA performs chemistry-free, unsupervised quality and hallucination assessment and better discriminates the two populations of VS images (positive vs. negative) without the need for ground truth HS images. Figure 2c showcases the confidence score distribution of the two populations and the confusion matrix with respect to the ground truth labels. Based on a validation set, we selected the threshold $\alpha = 0.5253$ scoring 100% sensitivity, and then blindly applied it to the test set (containing ~2,100 VS images) as shown by the green dashed line in Fig. 2c. Both the scatter plot and the

confusion matrix confirm a very good classification performance provided by our method on the two populations, with an accuracy of 99.8% and a sensitivity of 99.8%. Figure 2d further compares the performances of our method and histochemical ground truth-based quality assessment metrics, in terms of the absolute value of t-statistics and Kullback-Leibler (KL) divergence between the populations of positive and negative VS images. Our unsupervised quality and hallucination monitor provides significantly better positive/negative separation than all the supervised quality assessment metrics that demand access to the ground truth HS images.

In addition to these comparisons against traditional structural HS-based metrics, we further assessed the performance of our method against the decisions of three board-certified pathologists, covering the condition that no histochemical reference images are accessible. Each pathologist independently and blindly evaluated the quality of a set of VS images of human kidney samples. A consensus was reached on $N = 127$ images, among which $N_G = 66$ images came from good-staining VS models and $N_P = 61$ images from poor-staining VS models. Specifically, the pathologists were asked to score each image on three metrics: the stain quality of nuclei, cytoplasm and extracellular space from 1 to 4, with higher scores indicating better quality (refer to the evaluation details in the Methods section). Figure 3a reports the overall agreements between our method's confidence scores and the pathologists' scores on all $N = 127$ test images. Here, the range for the average score over three pathologists for an acceptable stain quality was set as $(2,4]$. Figure 3a reveals that our method agrees well with the pathologists' opinion (average stain quality score), corresponding to the blue and green areas in the confusion matrices shown

in Fig. 3a. We further break down Fig. 3a into Figs. 3b and 3c, comparing our method against the pathologist scores on the good-staining and poor-staining VS models, respectively. As illustrated in Fig. 3b, AQuA achieves 100% acceptance rate without any human intervention or supervision; in other words, all VS images from good-staining VS models are accepted. As a reference, the pathologists agreed on **98.5%** of the classification of our method (green zones) over the three separate stain quality scores and the average stain quality score. In Fig. 3c, when detecting hallucinated VS images coming from poor-staining models, AQuA scores **100% rejection** rate without any human supervision or access to histochemically stained ground truth images. In contrast, when presented with these VS images featuring realistic hallucinations (yellow zones), pathologists failed to flag these hallucinated images and could not identify the mismatch between VS images and their HS counterparts when the access to HS ground truth was unavailable. Examples of these realistic hallucinations that mislead the pathologists are shown in Extended Data Fig. 1a, and further analysis of pathologists' failure on these images are detailed in the Discussion section. This comparison clearly demonstrated the super-human level performance of AQuA as an autonomous virtual stain quality monitor and hallucination detector, rendering its potential in substituting human supervision in a virtual staining workflow where histochemical staining is eliminated and not available. Furthermore, AQuA is effective on detecting various failure patterns of VS models, including suboptimal convergence and generalization failure. In addition to testing our AQuA model on hallucinated images from early stopped checkpoints, we conducted an external generalization test to an overfitted VS model using the same classifier described

above, which has never seen such overfitted VS models. As shown in Extended Data Fig. 2a, the overfitted VS model, well trained on a small subset of the training data, fails to generalize on testing data and exhibits a distinct staining failure pattern compared with the early stopped VS models, such as missing a large portion of nuclei and generating blurry features. Despite the difficulty of detecting these staining failure patterns never seen before during the classifier training, AQuA successfully discriminated all VS images from the overfitted model and the early stopped model as illustrated in Extended Data Fig. 2b. This generalization test further demonstrates the robustness and generalizability of AQuA to various unseen, unknown staining failure patterns of VS models.

Image-level autonomous quality and hallucination assessment of lung tissue virtual staining

To showcase that AQuA's performance is not organ-dependent, we also applied AQuA to human lung tissue samples. Following a similar training strategy as detailed above, AQuA was trained on human lung VS images, and tested on a set of ~2,400 VS images excluded from the training stage (see the details in the Methods section). Figure 4 summarizes the performance of AQuA and compares it against histochemical-based, supervised quality assessment metrics that used HS ground truth images. Figure 4a showcases typical negative and positive VS images of human lung tissues, and the corresponding HS image is also shown as a reference. A poor VS model usually generates hallucinatory nuclei as pointed by the black arrows in Fig. 4a, which could be hard to distinguish without the HS image. Even with the supervision from HS ground truths, traditional quality assessment metrics such as MSE, PCC and PSNR cannot separate the populations of positive and

negative VS images as depicted in Fig. 4b. On the contrary, AQuA successfully discriminates the two groups of VS images as shown in Fig. 4c, and achieves a high accuracy of 97.8% and a sensitivity of 99.5% based on a threshold of $\alpha = 0.6263$. In terms of quantitative measures between the distributions of negative and positive VS images, our unsupervised method considerably outperforms supervised metrics and separates the two distributions much better, scoring significantly higher t-statistics and KL divergence values, even though it does not use ground truth HS images.

A blinded comparison with pathologists' scores was also implemented on human lung tissue samples, following the previously outlined methodology. On a set of $N = 99$ testing VS images, where pathologists reached consensus, our method shows a good agreement with their scoring as shown in Fig. 5a. Figure 5b and 5c illustrate the breakdown of Fig. 5a into the two cases on good-staining and poor-staining VS models, respectively. As shown in Fig. 5b, AQuA correctly identifies 100% of the negative VS images generated by good-staining models ($N_G = 59$), on which board-certified pathologists also accepted 98.3% of the same VS images and rejected only one disagreement case consisting almost entirely of red blood cells (RBCs). The disagreement case is further shown in Extended Data Fig. 1b and analyzed in the Discussion section. On the poor-staining VS models with $N_P = 40$ images, AQuA successfully rejected all hallucinatory images generated by these poor-staining models, including unrealistic (blue zone) and realistic (yellow zone) hallucinations. In contrast, board-certified pathologists failed to identify the VS images generated by poor-staining models harboring realistic hallucinations in the yellow zone. An example of these realistic hallucination is shown in Extended Data Fig. 1a. These comparisons further

confirm the super-human performance of AQuA's staining quality and hallucination assessment of VS models without access to ground truth HS images.

Model-level autonomous quality assessment of virtual staining

So far we have focused on implementing image-level autonomous quality monitoring of VS images; this scenario broadly applies to cases where already approved and rigorously validated VS models routinely and continuously generate VS images on new patient specimens, eliminating the standard histochemical staining workflow. Therefore, in these cases, AQuA serves as an autonomous and unsupervised “watchdog” to detect when the virtually stained images should not be used, even if the VS model has already been validated and approved for a given workflow.

Another important scenario that we will consider here is the VS model pre-approval process; here we will address the question of should we accept a given VS model into our workflow or not. To answer this question, we created an autonomous method for VS Model quality assessment, termed *M-AQuA* – which is also based on the AQuA framework. In practice, the performance of VS models, even if they are pre-approved, should be periodically evaluated, as part of a quality assurance routine, for generalization failures, which may result from a variety of causes including e.g., changes in imaging and sample preparation protocols, replacement or aberrations of imaging hardware as well as variations in biological content and storage conditions of tissue, among other factors. Hence, the periodic assessment of each VS model is necessary to assure the quality of the whole VS workflow on top of autonomous quality monitoring of each VS image. Of course, the use of *M-AQuA* is supposed to be much less frequent compared to image-level VS

quality assessments performed through AQuA as they serve two different functions; the former (*M*-AQuA) decides on the quality of the VS model, whether e.g., it should be approved for use or trained with more data; the latter (AQuA), however, decides if the generated VS image should be used for expert examination or not.

The workflow of *M*-AQuA is established upon AQuA-based examination on a set of N distinct VS images from the same VS model (that is under investigation). In this workflow, the previously trained image-level AQuA is independently and simultaneously applied on each VS image in the set, giving the image-level confidence scores. Then, the set of N confidence scores is compared to the confidence score distributions of good- and poor-staining VS models in the training set, to determine if the current VS model under test is generalizing well or not. Based on the theory of linear discriminant analysis (LDA), this is equivalent to comparing the average confidence score \bar{s} (over N) to a threshold β . This threshold β is determined on the training set following LDA such that the distributions of good- and poor-staining models have equal probability density at β (see the details in the Methods section). Figure 6a depicts the workflow of *M*-AQuA. As shown in Fig. 6a, N VS images generated by the same VS model under test are fed into AQuA in parallel, and the resulting \bar{s} is compared to β to determine whether the VS model should be accepted or rejected.

We applied this VS model quality assessment workflow on $M = 10$ test VS models of human lung tissue, among which half were good-staining VS models and the remaining ones were poor-staining models. We randomly sampled N images for each model $R = 100$ times, and summarized the repeatability of our test results; see Fig. 6b. The mean and

standard deviation of \bar{s} over $R = 100$ repetitions are shown under various N (from 2 to 20). The distribution of \bar{s} squeezes and eventually converges over increasing N , indicating that more test images (larger N) improve the discrimination accuracy of M -AQuA. Nevertheless, even for $N = 2$, the groups of good- and poor-staining models can be completely separated by a threshold of $\beta = 0.6159$. The accuracy and KL divergence between the \bar{s} distributions of good- and poor-staining models are also reported in Fig. 6, confirming the benefits of increasing N . In Figs. 6c and 6d, we further conducted a scalability test to M -AQuA with increasing the number of VS models (M). Figure 6c presents the histograms of \bar{s} of M randomly and independently selected VS test models, where the good-staining and poor-staining VS models each form 50% of the models. In all these tests with various M , all the good- and poor-staining models were correctly discriminated, scoring 100% accuracy as reported to Fig. 6d. The KL divergence between the \bar{s} distributions of good- and poor-staining models are also reported in Fig. 6d. These findings demonstrate the success and robustness of our model-level virtual staining quality assessment method (M -AQuA). We also tested the external generalizability of M -AQuA framework to poor-staining VS models that exhibit staining failure patterns never seen before, including, for instance, some VS models that overfitted to small training data. This might happen if HS staining (the ground truth needed for model training) is scarce, expensive or exhibit variations in quality, with a small number of tissue slides achieving acceptable staining quality, as usually encountered in e.g., immunohistochemistry and immunofluorescence tissue staining. Extended Data Figure 2c summarizes the external generalization performance of M -AQuA on 20 overfitted VS models, 20 early stopped models as well as 20 good-staining VS

models – all never seen before. The *M*-AQuA classifier and threshold established using only good-staining and early stopped VS models can successfully generalize to overfitted VS models, scoring 100% accuracy at the model assessment level. This experiment further confirms the robustness of *M*-AQuA to detecting various unseen types of VS model failure, including model suboptimality and generalization errors.

Autonomous staining quality assessment on histochemically stained (HS) images

In addition to autonomous staining quality and hallucination assessment on VS images produced by generative AI models, AQuA framework can also be extended to assessing the quality of HS slides of tissue samples, automatically identifying artifacts existing in traditional chemically-stained histology images. Such artifacts can normally arise because of a variety of factors and exhibit different types of imperfections, including loss of image contrast due to erroneous chemical staining operations, reduced image quality because of defocusing and optical aberrations, among other factors. Figure 7a presents typical examples of HS images of good and bad staining quality labeled by board-certified pathologists, where the well-stained images demonstrate normal histological features, with inconspicuous numbers, sizes, and distributions of cells, along with intra-tissue stain variability. In contrast, inadequately stained HS images show silhouettes of cells with lack of basophilia and loss of distinct nuclear contours. Depending on the clinical context, the poorly stained HS images could have been taken from a lung affected by conditions such as pulmonary infarction or organized pneumonia. Alternatively, they could result from poorly fixed, autolyzed tissue, or out-of-focus regions caused by autofocusing failure of the scanning optical microscope or thickness variations within the sectioned tissue. Figure 7b

reveals that the use of well-defined, hand-crafted analytical metrics, such as normalized nuclei count and average nuclei area (see the Methods section for the definitions of these metrics), often employed in quality assurance in histopathology, cannot effectively differentiate issues observed in poor vs. good HS images. However, the application of autonomous stain quality evaluation using AQuA framework leads to successful discrimination between the well-stained histology images and those laden with artifacts. Here AQuA was trained on a set of HS images of human lung tissue samples excluded from the testing stage (see the Methods section). This demonstration once again confirms the effectiveness of AQuA framework and its advantages over existing quality assessment methods, clearly highlighting the versatility of AQuA for autonomous assessment and detection of various artifacts, appearing in either virtually stained or histochemically stained tissue images.

Discussion

We reported AQuA, the first autonomous quality and hallucination assessment tool for virtually stained tissue images, and showcased its superior ability in identifying morphological hallucinations formed by generative AI-based VS models. We also highlighted AQuA's ability at the model level by accurately detecting poor quality VS models with types of staining failures never seen before. These demonstrate its role as a gatekeeper for AI-based virtual staining of tissue in digital pathology.

The rapid rise and wide spread use of deep learning have caused concerns regarding the reliability and quality control of neural network outputs^{36,37}, especially in critical biomedical

applications like virtual tissue staining^{17,38-40}. The generative nature of virtual staining models not only brings up risks from new attack strategies⁴¹⁻⁴³, but also cast difficulty in detecting the failure modes of these models using traditional and supervised evaluation metrics. In fact, for virtual staining of tissue, in the deployment phase of the VS model there would be *no* histochemical staining available and therefore supervised evaluation metrics based on ground truth images *cannot* be used in the VS workflow. In VS related initial studies, laborious manual quality assessments performed by pathologists based on high-level semantic features and domain expertise were critical to assure the quality of VS models; this is not practical for the deployment of a VS model, which is expected to function autonomously. Therefore, AQuA provides a much needed, transformative tool for VS quality assessment and hallucination detection without access to histochemically stained ground truth images. Through blind testing on human kidney and lung tissue samples, AQuA achieved 99.8% accuracy and 99.8% sensitivity based on its chemistry-free, unsupervised quality assessment, outperforming common structural, supervised quality assessment metrics that used HS ground truth images. In comparison with a group of board-certified pathologists, the classification of AQuA reached 100% agreement on negative VS images, generated by good VS models, also manifesting a super-human performance when detecting realistic-hallucinatory VS images that would normally mislead pathologists to diagnose realistic-looking VS tissue images that never existed in real life.

The success of AQuA is closely linked to the novel design of AQuA-Net. First, by leveraging successive iterations between the VS and VAF models, inference uncertainty within the VS

image gradually accumulates and therefore becomes easier to accurately detect. The advantages of these VS-VAF iterations are studied in additional experiments performed on the hyperparameter T , as summarized in Extended Data Fig. 3. Compared to the AQuA-Net performance without such VS-VAF iterations ($T = 1$), the introduction of these iterations significantly improves both the classification accuracy and the KL divergence between the positive and negative VS image distributions. Second, the use of majority voting mechanism considerably enhances the classification performance in terms of accuracy and sensitivity. As reported in Extended Data Table 3, compared to a single classifier ($C = 1$), majority voting mechanism with $C > 1$ effectively boosts both the accuracy and the sensitivity. The optimal hyperparameters T and C were empirically determined and highlighted in Extended Data Table 1 for our experiments using human kidney and lung tissue samples. Furthermore, compared to drastically increasing the scale of the network model using complicated architectures and more trainable parameters, the majority voting mechanism provides an efficient and effective strategy to improve inference performance, as highlighted in Extended Data Fig. 4 and Extended Data Table 2.

H&E staining has remained unchanged for over a century due to its reproducibility with a variety of fixatives, displaying a broad range of cytoplasmic, nuclear, and extracellular matrix features. In a typical tissue sample, nuclei are stained blue/purple (with Hematoxylin), while the cytoplasm and extracellular matrix exhibit varying degrees of pink staining (with Eosin). Notably, cells display significant intranuclear details, featuring cell-type-specific patterns of nuclear condensation that hold diagnostic significance. The ability of virtual tissue staining technologies to replicate these intricate details using

generative AI models harbors a concerning aspect of inducing hallucinations and generating visual artifacts. Many of these artifacts would classify as unrealistic hallucinations, resembling traditional histology artifacts that are easy to spot by experts—e.g., morphological changes that may arise in any tissue processing step, such as unstained/overstained areas, tissue folding, tears and holes. These unrealistic and easy-to-spot artifacts present a sharp contrast from well stained areas. On the other hand, realistic hallucinations are much harder to detect due to their tendency to blend in the stained tissue. Hence, realistic hallucinations could lead to misinterpreting tissue characteristics, potentially misleading the diagnostic process by evaluating tissue images that never existed. To better highlight this, Extended Data Fig. 1a provides examples of realistic hallucinations that were predicted as having acceptable quality by a group of pathologists, which were correctly flagged by AQuA. Two pairs of images involved kidney tissue samples, wherein several nuclei in the glomeruli and renal tubules are missing on the VS images compared to their histochemical counterparts. Human practitioners/pathologists determined that these images were properly stained (ready to be diagnosed), while AQuA labeled these images as poor staining quality (i.e., cannot be used for diagnostics). Another example originates from a lung tissue sample, where a number of type 2 alveolar cells and macrophages are understained in the VS images compared to traditional histology. Similar to the kidney case, these images were approved by pathologists but rejected by AQuA without access to HS ground truth images. The only instance of a false negative evaluation (compared to pathologist consensus) that we observed is reported in Extended Data Fig. 1b; it is taken from another lung sample, where the entire patch is composed of RBCs and

scattered lymphocytes. The under-representation of RBCs in the training data of normal lung tissue slides led to AQuA accepting this image field-of-view, whereas human pathologists, who are familiar with similar morphologies in lung specimens, disapproved its quality.

In conclusion, as an autonomous tool, AQuA provides a robust, scalable, and efficient quality assurance framework as a hallucination detector for VS images, and provides a transformative advancement toward more reliable, trustworthy AI in virtual staining-related pathology applications.

References

1. McInnes, E. Artefacts in histopathology. *Comp. Clin. Pathol.* **13**, 100–108 (2005).
2. Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *Lancet Oncol.* **20**, e253–e261 (2019).
3. Jahn, S. W., Plass, M. & Moinfar, F. Digital Pathology: Advantages, Limitations and Emerging Perspectives. *J. Clin. Med.* **9**, 3697 (2020).
4. Borhani, N., Bower, A. J., Boppart, S. A. & Psaltis, D. Digital staining through the application of deep neural networks to multi-modal multi-photon microscopy. *Biomed. Opt. Express* **10**, 1339 (2019).
5. Rivenson, Y. *et al.* Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nat. Biomed. Eng.* **3**, 466–477 (2019).
6. Rivenson, Y. *et al.* PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning. *Light Sci. Appl.* **8**, 23 (2019).
7. Nygate, Y. N. *et al.* Holographic virtual staining of individual biological cells. *Proc. Natl. Acad. Sci.* **117**, 9223–9231 (2020).
8. Rivenson, Y., de Haan, K., Wallace, W. D. & Ozcan, A. Emerging Advances to Transform Histopathology Using Virtual Staining. *BME Front.* **2020**, 9647163 (2020).
9. Li, D. *et al.* Deep Learning for Virtual Histological Staining of Brightfield Microscopic Images of Unlabeled Carotid Artery Tissue. *Mol. Imaging Biol.* **22**, 1301–1309 (2020).
10. Li, X. *et al.* Unsupervised content-preserving transformation for optical microscopy. *Light Sci. Appl.* **10**, 44 (2021).

11. Pradhan, P. *et al.* Computational tissue staining of non-linear multimodal imaging using supervised and unsupervised deep learning. *Biomed. Opt. Express* **12**, 2280 (2021).
12. Picon, A. *et al.* Autofluorescence Image Reconstruction and Virtual Staining for In-Vivo Optical Biopsying. *IEEE Access* **9**, 32081–32093 (2021).
13. Meng, X., Li, X. & Wang, X. A Computationally Virtual Histological Staining Method to Ovarian Cancer Tissue by Deep Generative Adversarial Networks. *Comput. Math. Methods Med.* **2021**, 1–12 (2021).
14. Abraham, T., Costa, P. C., Filan, C. E., Robles, F. & Levenson, R. M. Mode-mapping qOBM microscopy to virtual hematoxylin and eosin (H&E) histology via deep learning. in *Unconventional Optical Imaging III* (eds. Georges, M. P., Popescu, G. & Verrier, N.) 58 (SPIE, Strasbourg, France, 2022). doi:10.1117/12.2622160.
15. Soltani, S. *et al.* Prostate cancer histopathology using label-free multispectral deep-UV microscopy quantifies phenotypes of tumor aggressiveness and enables multiple diagnostic virtual stains. *Sci. Rep.* **12**, 9329 (2022).
16. Cao, R. *et al.* Label-free intraoperative histology of bone tissue via deep-learning-assisted ultraviolet photoacoustic microscopy. *Nat. Biomed. Eng.* **7**, 124–134 (2022).
17. Bai, B. *et al.* Deep learning-enabled virtual histological staining of biological samples. *Light Sci. Appl.* **12**, 57 (2023).
18. Liu, Y., Yuan, H., Wang, Z. & Ji, S. Global Pixel Transformers for Virtual Staining of Microscopy Images. *IEEE Trans. Med. Imaging* **39**, 2256–2266 (2020).
19. Zhang, Y. *et al.* Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue. *Light Sci. Appl.* **9**, 78 (2020).

20. Li, J. *et al.* Biopsy-free in vivo virtual histology of skin using deep learning. *Light Sci. Appl.* **10**, 233 (2021).
21. Zhang, Y. *et al.* Virtual Staining of Defocused Autofluorescence Images of Unlabeled Tissue Using Deep Neural Networks. *Intell. Comput.* **2022**, 2022/9818965 (2022).
22. Bai, B. *et al.* Label-Free Virtual HER2 Immunohistochemical Staining of Breast Tissue using Deep Learning. *BME Front.* **2022**, 9786242 (2022).
23. Li, Y. *et al.* Virtual histological staining of unlabeled autopsy tissue. *Nat. Commun.* **15**, 1684 (2024).
24. Burlingame, E. A. *et al.* SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning. *Sci. Rep.* **10**, 17507 (2020).
25. de Haan, K. *et al.* Deep learning-based transformation of H&E stained tissues into special stains. *Nat. Commun.* **12**, 4884 (2021).
26. Xie, W. *et al.* Prostate Cancer Risk Stratification via Nondestructive 3D Pathology with Deep Learning-Assisted Gland Analysis. *Cancer Res.* **82**, 334–345 (2022).
27. Ghahremani, P. *et al.* Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification. *Nat. Mach. Intell.* **4**, 401–412 (2022).
28. Bouteldja, N., Klinkhammer, B. M., Schlaich, T., Boor, P. & Merhof, D. Improving unsupervised stain-to-stain translation using self-supervision and meta-learning. *J. Pathol. Inform.* **13**, 100107 (2022).
29. Zhang, R. *et al.* MVFStain: Multiple virtual functional stain histopathology images generation based on specific domain mapping. *Med. Image Anal.* **80**, 102520 (2022).

30. Yang, X. *et al.* Virtual Stain Transfer in Histology via Cascaded Deep Neural Networks. *ACS Photonics* **9**, 3134–3143 (2022).
31. Hang Wu, *et al.* Detection of blur artifacts in histopathological whole-slide images of endomyocardial biopsies. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.* **2015**, 727–730 (2015).
32. Foucart, A., Debeir, O. & Decaestecker, C. Artifact Identification in Digital Pathology from Weak and Noisy Supervision with Deep Residual Networks. in *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)* 1–6 (IEEE, Brussels, Belgium, 2018). doi:10.1109/CloudTech.2018.8713350.
33. Kanwal, N., Perez-Bueno, F., Schmidt, A., Engan, K. & Molina, R. The Devil is in the Details: Whole Slide Image Acquisition and Processing for Artifacts Detection, Color Variation, and Data Augmentation: A Review. *IEEE Access* **10**, 58821–58844 (2022).
34. Kanwal, N. *et al.* Are you sure it's an artifact? Artifact detection and uncertainty quantification in histological images. *Comput. Med. Imaging Graph. Off. J. Comput. Med. Imaging Soc.* **112**, 102321 (2024).
35. Goodfellow, I. J. *et al.* Generative Adversarial Networks. *ArXiv14062661 Cs Stat* (2014).
36. Abdar, M. *et al.* A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* **76**, 243–297 (2021).
37. Gawlikowski, J. *et al.* A Survey of Uncertainty in Deep Neural Networks. (2021) doi:10.48550/ARXIV.2107.03342.

38. Levy, J. J. *et al.* A large-scale internal validation study of unsupervised virtual trichrome staining technologies on nonalcoholic steatohepatitis liver biopsies. *Mod. Pathol.* **34**, 808–822 (2021).
39. Vasiljević, J., Nisar, Z., Feuerhake, F., Wemmert, C. & Lampert, T. CycleGAN for virtual stain transfer: Is seeing really believing? *Artif. Intell. Med.* **133**, 102420 (2022).
40. Kreiss, L. *et al.* Digital staining in optical microscopy using deep learning - a review. *PhotoniX* **4**, 34 (2023).
41. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and Harnessing Adversarial Examples. Preprint at <http://arxiv.org/abs/1412.6572> (2015).
42. Papernot, N. *et al.* The Limitations of Deep Learning in Adversarial Settings. Preprint at <http://arxiv.org/abs/1511.07528> (2015).
43. Huang, S., Papernot, N., Goodfellow, I., Duan, Y. & Abbeel, P. Adversarial Attacks on Neural Network Policies. Preprint at <http://arxiv.org/abs/1702.02284> (2017).

Figures

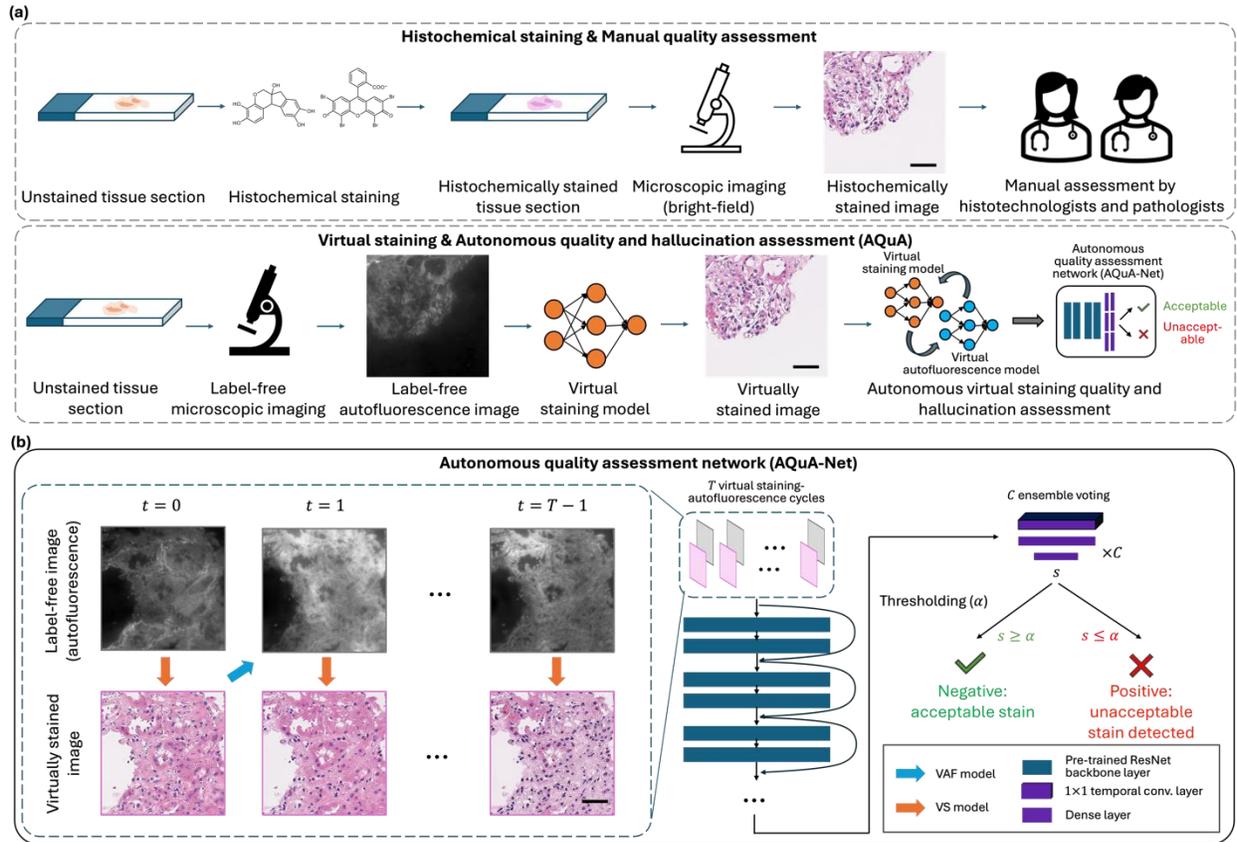


Figure 1 Workflow of histochemical staining (HS) and virtual staining (VS) of tissue and corresponding quality assessment. (a) Traditional histochemical staining workflow with manual quality assessment by histotechnologists and pathologists, and virtual staining workflow with autonomous quality and hallucination assessment by AQUA. (b) Details of AQUA-Net. The image sequences generated by T successive virtual staining-autofluorescence cycles are fed into AQUA-Net, passing the pre-trained neural network backbone. C voting heads consisting of a temporal convolutional layer and dense layers generate the output confidence score to determine the quality of the VS image after thresholding (α). Scale bar: $50\mu\text{m}$.

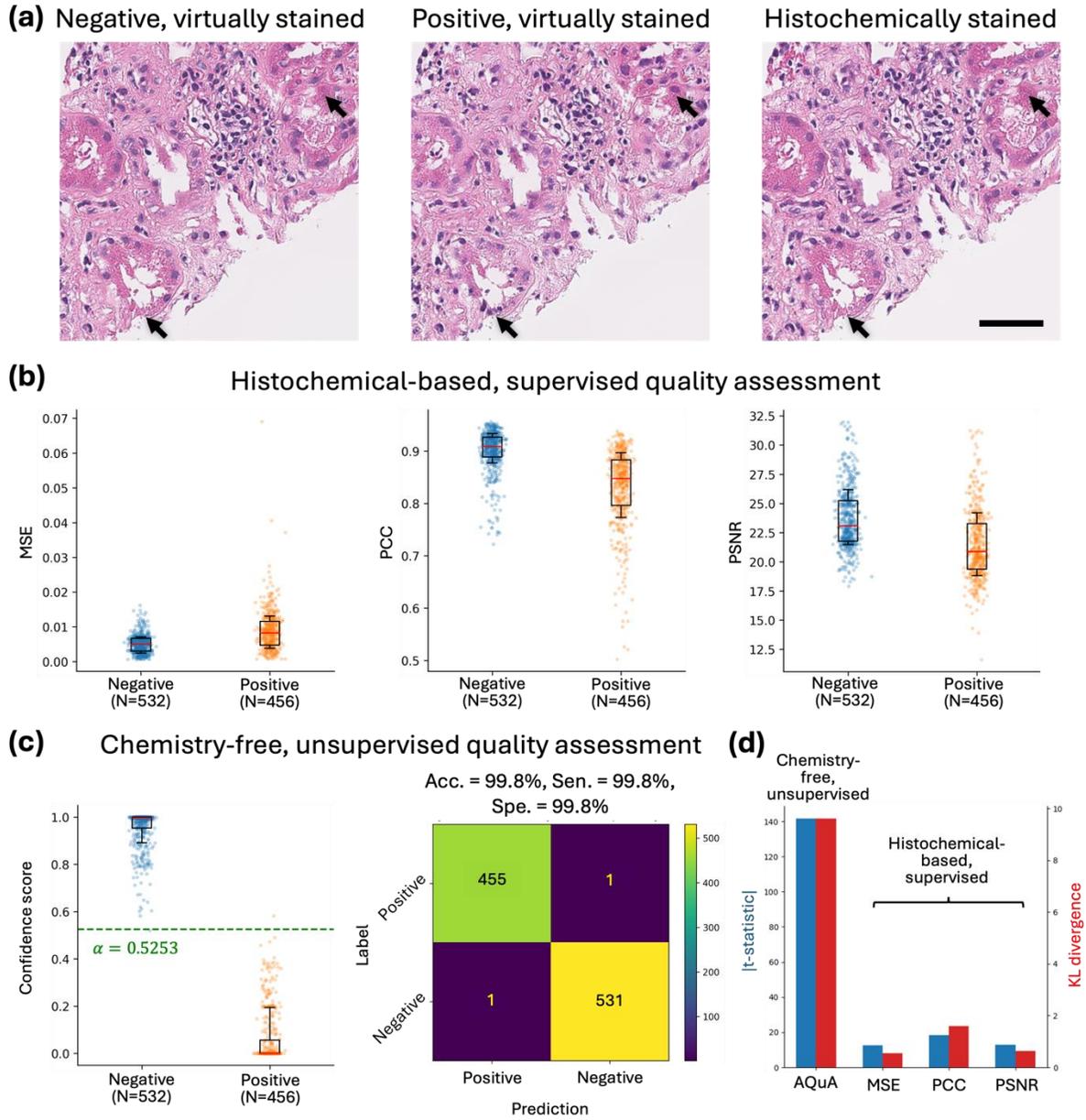


Figure 2 Image-level autonomous quality and hallucination assessment of kidney tissue virtual staining using AQa. (a) Examples of a negative VS image generated by a good-staining model, a positive VS image generated by a poor-staining model, and the corresponding histochemically stained HE image for reference (ground truth). (b) Performance of histochemical-based, supervised quality assessment metrics on the test set of VS images of human kidney samples. MSE, PCC, PSNR are calculated between the

VS images and the corresponding HS ground truth images. (c) Chemistry-free, unsupervised quality and hallucination assessment by AQuA on the same test set. Confidence score averaged from C=4 ensemble voters with T=5 virtual staining-autofluorescence cycles. The green dashed line illustrates the 100% sensitivity threshold α determined on the validation set. (d) Despite not using HS ground truth images in its evaluation, AQuA provides better discrimination between the positive and negative VS image populations compared to histochemical-based, supervised image quality metrics. Scale bar: 50 μ m.

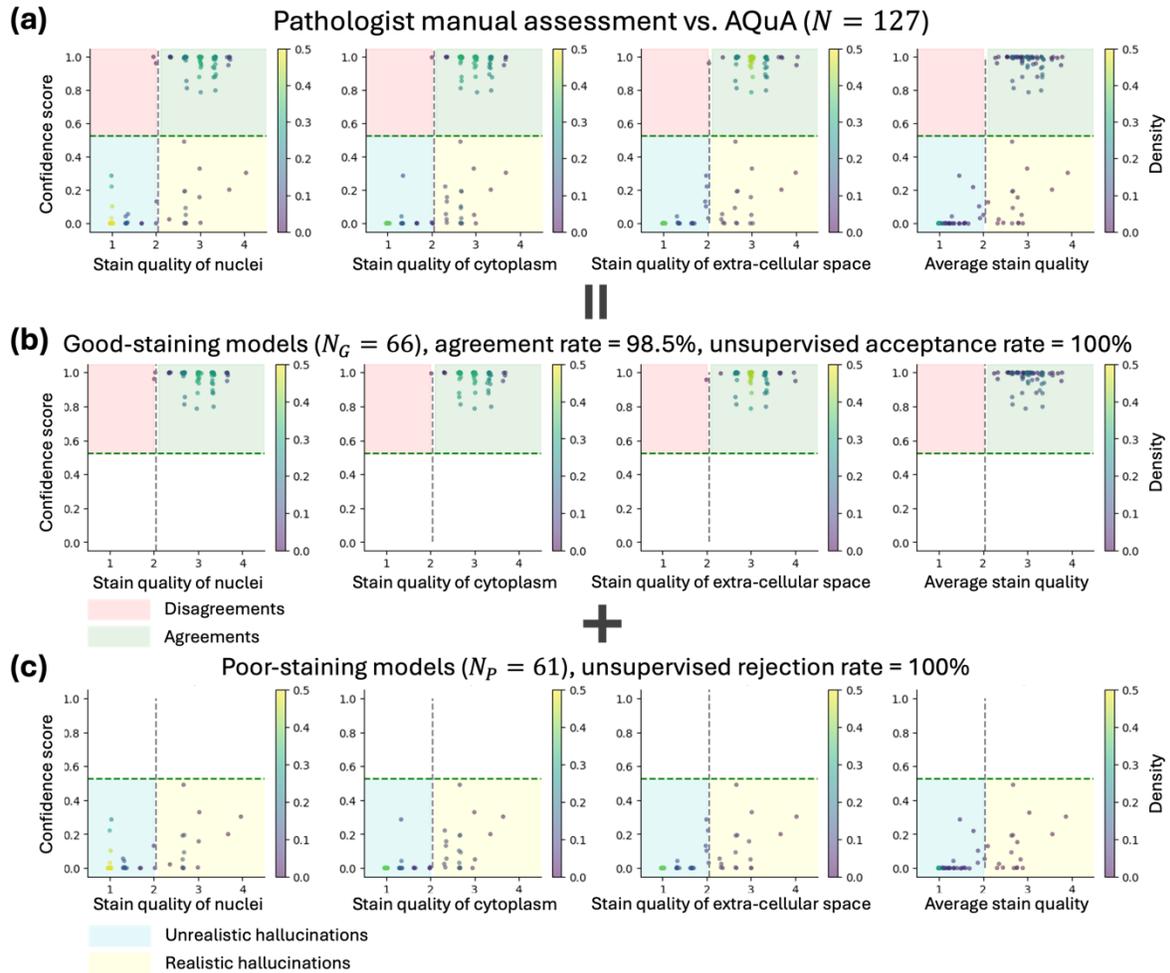


Figure 3 Comparison between AQuA and a group of board-certified pathologists on VS image quality and hallucination assessment. (a) Overall agreement on a test set of $N = 127$ VS images of human kidney samples. Each scatter plot is further broken down into the case of good-staining and poor-staining VS models. (b) Agreement on a test set of $N_G = 66$ VS images generated by good-staining VS models. AQuA achieves 100% accuracy on negative VS images while the pathologists agree on 98.5% of AQuA's results on average. (c) Agreement on a test set of $N_P = 61$ VS images generated by poor-staining VS models. AQuA rejects 100% of the positive VS images generated by poor-staining models while the pathologists failed to distinguish realistic hallucinations in these positive images. Some of

these realistic hallucination cases are further explored in Extended Data Fig. 1 and in the Discussion section.

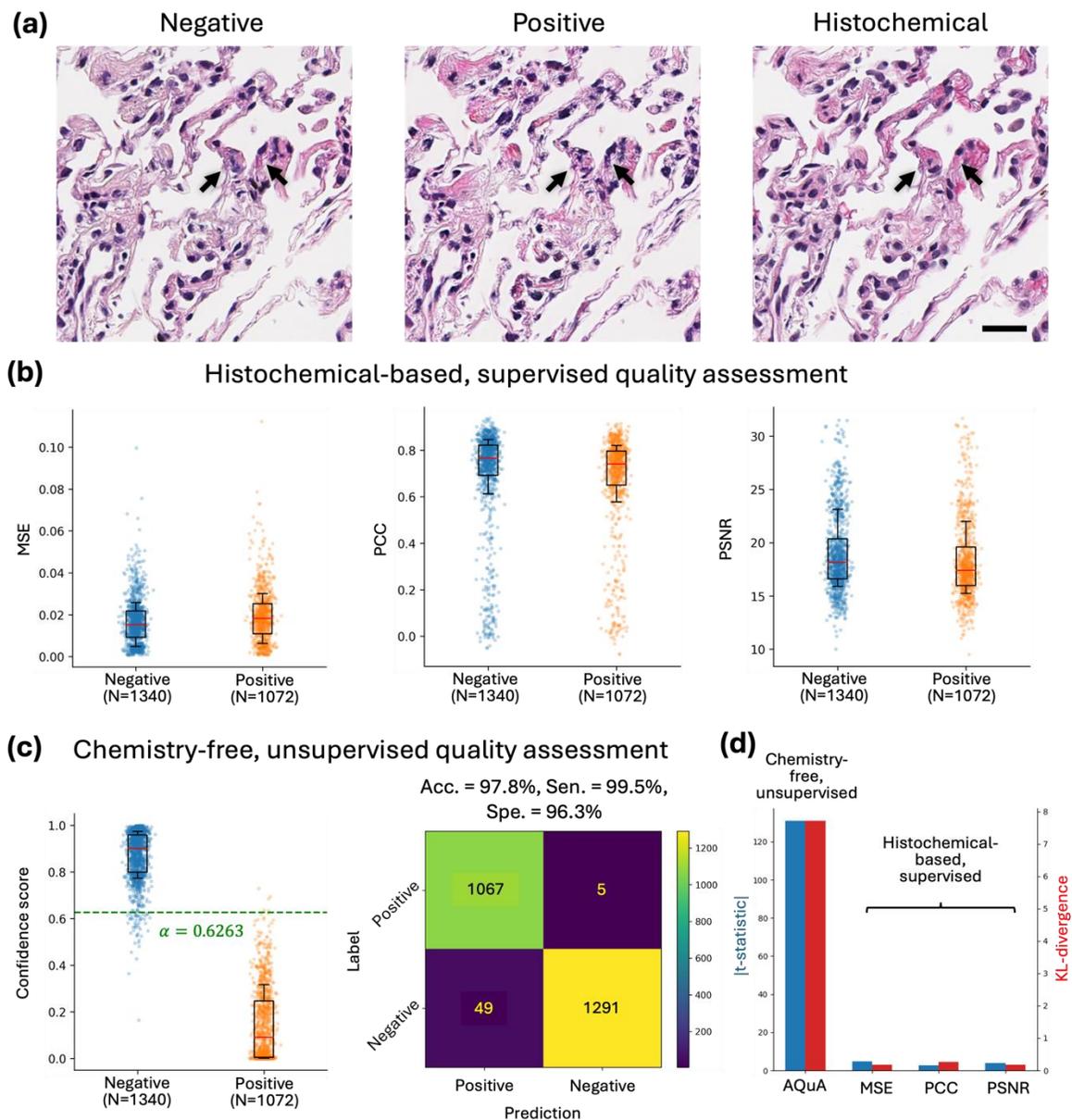


Figure 4 Image-level autonomous quality and hallucination assessment of lung tissue virtual staining using AQuA. (a) Examples of a negative VS image generated by a good-staining model, a positive VS image generated by a poor-staining model, and the

corresponding histochemically stained HE image for reference (ground truth). (b) Performance of histochemical-based, supervised quality assessment metrics on the test set of VS images of human lung samples. MSE, PCC, PSNR are calculated between the VS images and the corresponding HS ground truth images. (c) Chemistry-free, unsupervised quality assessment by AQuA on the same test set. Confidence score averaged from C=5 ensemble voters with T=5 virtual staining-autofluorescence cycles. The green dashed line illustrates the 100% sensitivity threshold determined on the validation set. (d) Despite not using HS ground truth images in its evaluation, AQuA provides better discrimination between the positive and negative VS image populations compared to histochemical-based, supervised image quality metrics. Scale bar: 20 μ m.

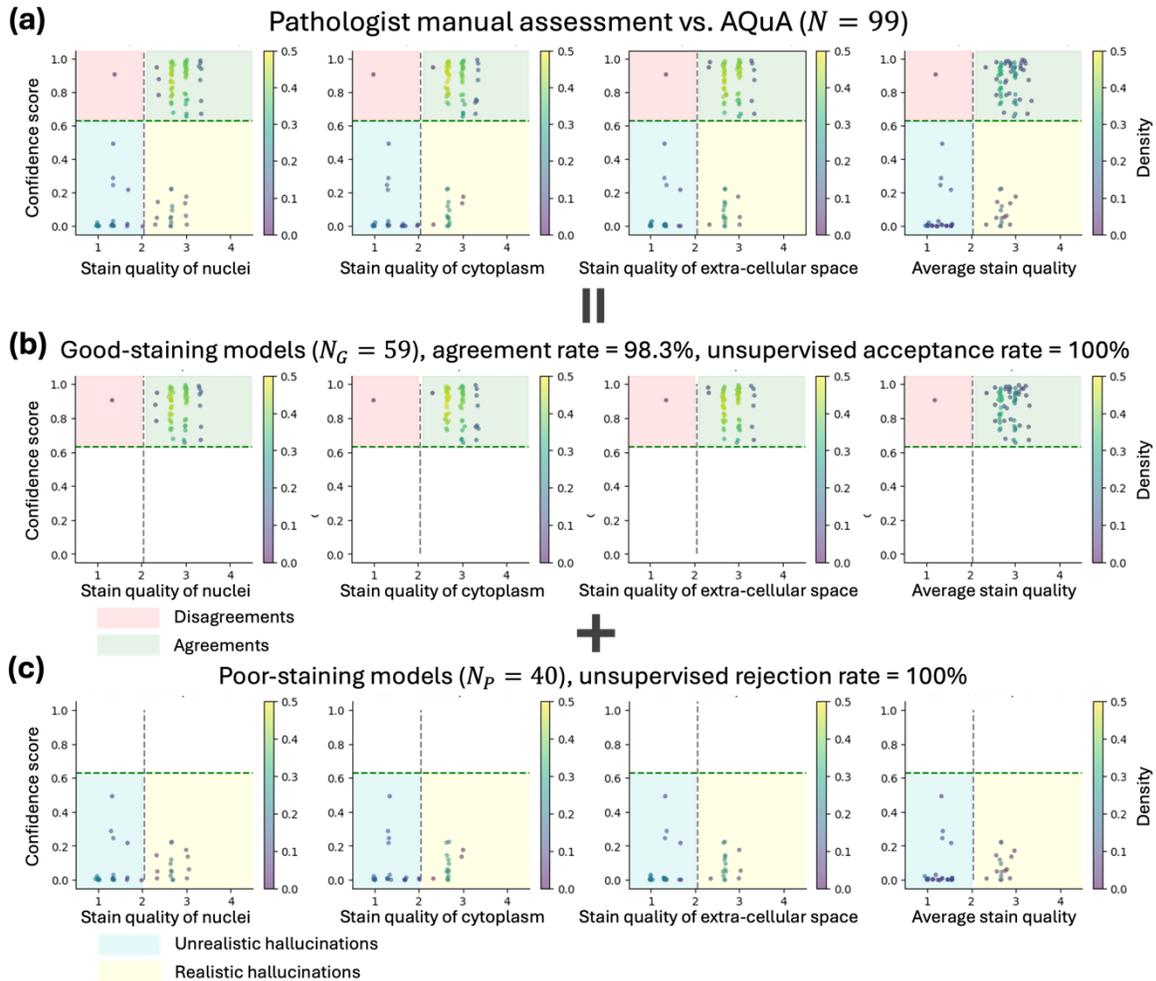


Figure 5 Comparison between AQuA and a group of board-certified pathologists on VS image quality and hallucination assessment. (a) Overall agreement on a test set of $N = 99$ VS images of human lung samples. Each scatter plot is further broken down into the case on good-staining and poor-staining models. (b) Agreement on a test set of $N_G = 66$ VS images generated by good-staining VS models. AQuA achieves 100% accuracy on negative images while the diagnosticians agree on 98.3% of AQuA's results on average. (c) Agreement on a test set of $N_P = 40$ VS images generated by poor-staining VS models. AQuA rejects 100% of the positive VS images generated by poor-staining models while the pathologists failed to distinguish realistic hallucinations in these positive images. Some of

these disagreement and realistic hallucination cases are further explored in Extended Data Fig. 1 and in the Discussion section.

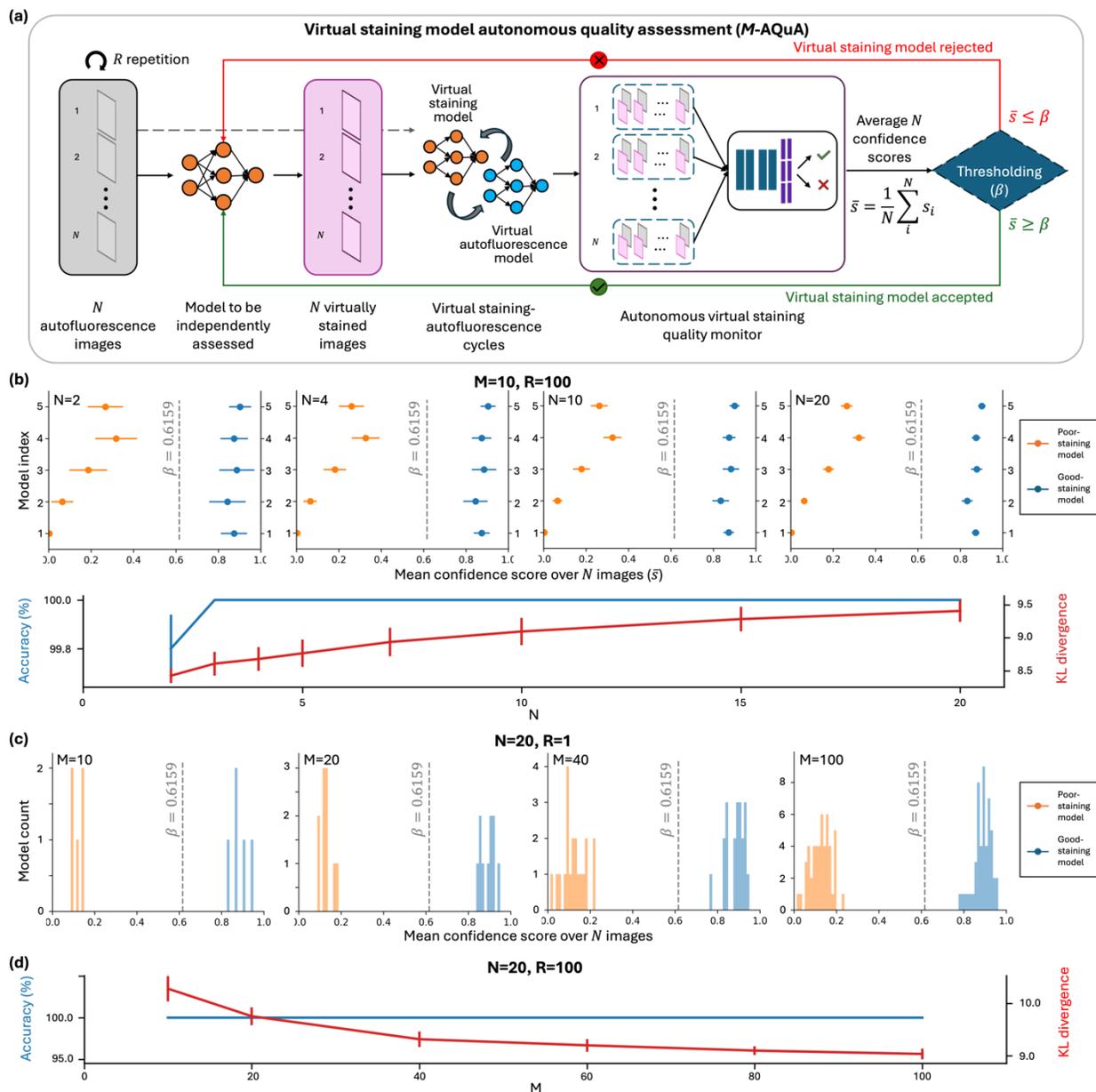


Figure 6 Virtual staining model autonomous quality assessment (M-AQuA). (a) The average confidence score over a set of N test images from the same VS model is compared against the threshold β to accept or reject the VS model under investigation. β is determined by

LDA on the training set of VS images of human lung tissue samples. (b) M -AQuA using varying numbers (N) of random testing images on $M = 10$ VS models of human lung tissue. The mean confidence score of each virtual staining model \bar{s} , the overall model classification accuracy and the KL divergence between two model groups (poor vs. good staining models) are shown. Mean and standard deviation values are reported on $R = 100$ repetitions. (c) M -AQuA using $N = 20$ random testing images on varying numbers (M) of VS models of human lung tissue. Histograms of the mean confidence scores of M models are shown for $R = 1$. (d) Accuracy and KL divergence values are reported as a function of M . Mean and standard deviation values are reported on $R = 100$ repetitions. In these analyses, the ratio of good- to poor-staining models is 1:1.

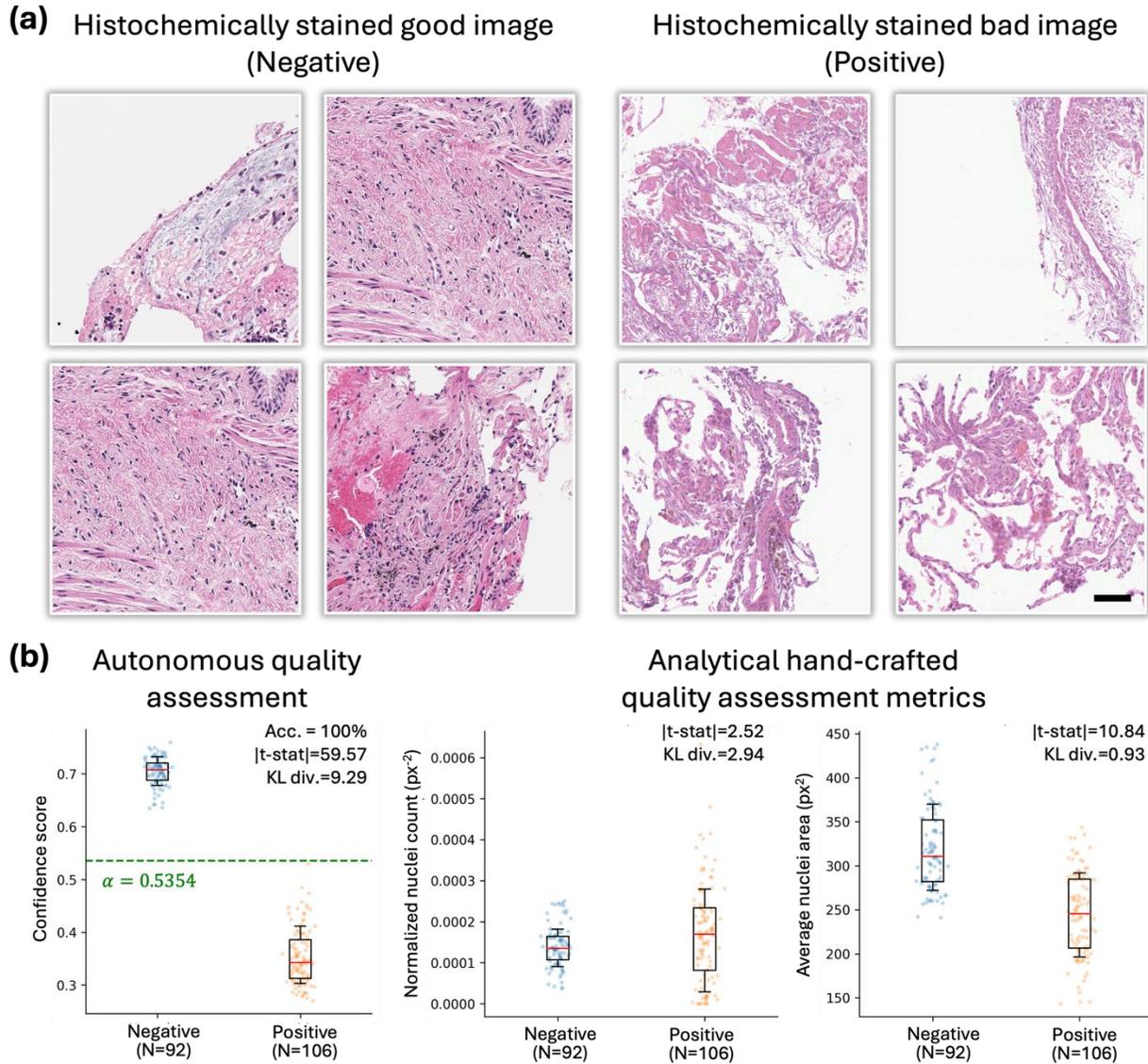


Figure 7 AQuA performance on histochemically stained (HS) images. (a) Examples of good (negative) and bad (positive) HS images. (b) Compared to conventional hand-crafted image quality metrics, AQuA provides better discrimination between the positive and negative HS image populations. Scale bar: 50 μ m.

Supplementary Information includes Methods on:

- Sample preparation and standard histochemical H&E staining
- Image acquisition
- Image co-registration between paired AF and H&E images
- Dataset preparation for training VS and VAF networks
- Network architecture and training schedule of VS networks
- Network architecture and training schedule of VAF networks
- Definition of good-staining and poor-staining models for virtual staining
- Dataset preparation and splitting of AQuA
- Architecture and training schedule of AQuA-Net
- Evaluation metrics
- Linear discriminant analysis
- Hand-crafted analytical metrics evaluating histochemically stained H&E slides