# The Aachen Protocol for Deep Learning Histopathology: A hands-on guide for data preprocessing

Hannah Sophie Muti (1), Chiara Loeffler (1), Amelie Echle (1), Lara R. Heij (2, 3, 4), Roman D. Buelow (4), Jeremias Krause (1), Laura Broderius (1), Jan Niehues (1), Georgia Liapi (1), Peter Boor (4), Heike Grabsch (5, 6), Sara Kochanny (7), Alexander T. Pearson (7), Jakob Nikolas Kather (1)

1) Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany
2) Visceral and Transplant Surgery, University Hospital RWTH Aachen, Aachen, Germany
3) NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, the Netherlands
4) Institute of Pathology, University Hospital RWTH Aachen, Aachen, Germany
5) Pathology and GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands
6) Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom
7) Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, Illinois, USA

## Abstract

**Background:** Deep learning can predict clinically relevant features such as genetic alterations directly from H&E stained histology images.[1,2] In practice, many clinically relevant questions are limited by availability of clinical data and by the lack of standardized preprocessing pipelines. In our research projects, we strive to keep a consistent data format across projects to facilitate downstream analysis.

**Workflow:** We analyze cohorts of cancer patients and try to predict clinically relevant labels directly from whole slide images (WSI). To achieve this, we manually or automatically detect tumor tissue in the WSI, tessellate the tumor into smaller image tiles and store these tiles in a *cohort directory* (Figure 1). We prepare a *Slide Master Table*, specifying which WSI belongs to which patient and a *Patient Master Table*, specifying the labels (target categories) for each patient. Our publicly available scripts automate the remaining workflow:[3] Tiles are loaded, are matched to WSIs, which are matched to patients, which are matched to labels. Deep neural networks are trained to predict the labels and are evaluated on external cohorts.

**Target audience:** This is a best practice manual focused on practical aspects such as file names, ground truth data tables and ROI annotation. This document is intended for onboarding new team members and for our academic collaborators. We hope that beyond our teams, this consensus document might be useful for other groups in the deep learning histopathology community. Our data standards are is inspired by The Cancer Genome Atlas (TCGA) standards (http://portal.gdc.cancer.gov). Please give your feedback on http://kather.ai.

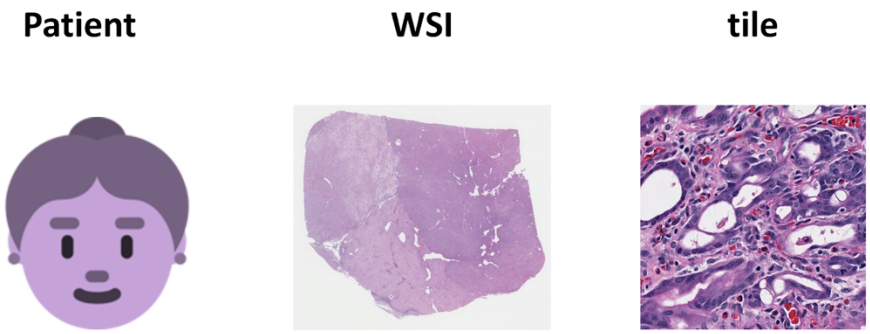


**Figure 1: A patient, a whole slide image (WSI) and a tile.** Typically, one or multiple WSIs are available for one patient. Each WSI generates many tiles. Clinically relevant labels are available on the patient level only, and all WSIs and tiles inherit the label of their parent patient. Image credit for patient: Twitter Twemoji, Image credit for WSI and tile: TCGA (https://portal.gdc.cancer.gov)

# Key terms and definitions

| | |
|---|---|
| Slide | A glass slide with a tumor section for microscopy |
| Whole slide image (WSI) | A scanned glass slide used for histological examination |
| Tile, patch or block | A small area cropped from a whole slide image |
| Patient label | A property of each patient, to be predicted from a WSI |
| Tile label | Tiles inherit the label from their parent patient |
| Cohort | A group of patients which is analyzed together |
| Region of Interest (ROI) | A region in a WSI, e.g. the tumor tissue |
| Annotation | A manually or automatically drawn ROI on a WSI |
| Conv. neural network (CNN) | An artificial neural network being trained on histology images |

# Preparing cohort metadata

## General

- For each cohort, we require a **Slide Master Table** and a **Patient Master Table**.
- The **Slide Master Table** assigns WSI to patients while the **Patient Master Table** assigns clinically relevant labels to patients. Examples for clinically relevant labels are genetic features (such as microsatellite instability, MSI) or survival.

## Preparing the Slide Master Table

- Each patient can have multiple slides (whole slide images, WSI). In the “Slide Master Table”, each WSI is linked to exactly one patient. There cannot be duplicate WSI names in this table. We require a CSV table with ‘,’ separator (Table 1)
- The column headers must be ‘PATIENT’ and ‘FILENAME’.
- Although one patient can have multiple slides, usually one slide per patient works well.

```
PATIENT,FILENAME
Aachen-HCC-0001,Aachen-HCC-0001-Slide001
Aachen-HCC-0001,Aachen-HCC-0001-Slide002
Aachen-HCC-0001,Aachen-HCC-0001-Slide003
Aachen-HCC-0002,Aachen-HCC-0002-Slide001
Aachen-HCC-0003,Aachen-HCC-0003-Slide001
```

**Table 1**: Example of a Slide Master Table

### Preparing the Patient Master Table

- Each patient can have several target variables which are later being used as ground truth labels for deep learning histopathology. We assume that all tiles inherit the target variable (label) of their parent WSI, which inherits the label of the patient. An example for a commonly used label is microsatellite status (MSI), as published in our previous work (Kather et al., Nature Medicine, 2019). Target variables can be continuous numeric data (e.g. `0.13, 0.54, 0.32`), categorical numeric (e.g. `1` and `2`) or categorical text (e.g. `MSIH` and `nonMSIH`). For categorical variables, text is preferred to numeric values. This data is stored in the 'Patient Master Table' which should be provided as a CSV file (with ',' separators) or in XLSX format, as shown in Table 2
- In the Patient Master Table, each patient is one row. There cannot be duplicate patient names in this table.

```
PATIENT,Hypermutated,MutationCount,AliveAfter1Year
Aachen-HCC-0001,0,15,NO
Aachen-HCC-0002,1,1200,YES
Aachen-HCC-0003,0,34,YES
```

**Table 2**: Example of a Patient Master Table

### Specifics for survival data

- Survival data (e.g. overall survival or progression-free survival) should contain the duration of follow-up period and whether an event (e.g. death) occurred.
- To be easily usable in downstream workflows, this type of survival data can be converted to categorical data, e.g. 'aliveAfter1Year', 'aliveAfter2Years' with values 'YES', 'NO', and 'NA' (or similar)

```
PATIENT,followUpDays,death
Aachen-HCC-0001,115,YES
Aachen-HCC-0003,410,NO
Aachen-HCC-0003,380,YES
```

**Table 3**: Example of survival data in a Patient Master Table

### Missing values

- missing values in the Patient Master Table can be indicated empty fields or by any of these strings: *'NA', 'NaN', 'N/A', 'na', 'n.a.', 'N.A.', 'NotAvailable', 'undef', 'unknown', 'x', 'NotApplicable', 'notperformed', 'NotPerformed', 'Notassigned', 'excluded', 'exclude'.*

## Preparing an image dataset

### General considerations: What makes a good image dataset?

1. **Heterogeneity**: Deep learning networks recognize patterns in images and learn from these patterns. Variations in coloring, tumor subtype, tissue shape, artifacts etc. can be advantageous for training a stable classifier. Of course, especially parameters such as tumor subtype or tissue shape can vary across different medical centers, especially in different countries. That is why establishing international academic collaborations is paramount to create high quality classifiers.
2. **Size**: The more images a classifier is trained upon, the better it gets. This rule applies to a certain number of images: once you have built up a dataset containing several thousand slides, you may stop collecting and start training. For cohort sizes, biological

variability is key: WSI from many different patients yield a better classifier than multiple WSI from a single patient.

3. **Class balance**: Once you have chosen a parameter of interest to train your classifier for, take a minute and check the number of patients in each group. Before training, tiles from the more abundant class are undersampled so that the classifier is trained on balanced classes. This may negatively affect training, so your dataset should be as balanced as possible to mitigate this problem. On the other hand, the deep learning system should be trained on a cohort which is as similar as possible to real-world cohorts to which the system will be applied. In that case, achieving real-world distribution of the target label is more important than class balance.
4. **Image quality**: In order to train a stable classifier, the visual quality of tumor tissue should not be impaired by technical artifacts such as air bubbles or pen marks on the glass slide. This begins with the scanning process: drops, dust, air inclusions, hair etc. should not be visible on WSI. Old scans cannot be changed anymore, but please keep that in mind when scanning new images.

## Scanning, naming files, naming patients

- H&E stained tissue slides should be scanned at 20x (0.5 µm/px) or 40x (0.25 µm/px). For processing, we currently resample all tiles to 0.5 µm/px.
- If possible, avoid pen marks on the slide (but if marks are present, often we can still work with the slides). Whole slide images are required to have at least 1 mm² of contiguous tumor tissue.
- Preferred scanner vendors are Aperio (SVS format) and Hamamatsu (NDPI format), other scanners might work as well as long as the files are compatible with QuPath (https://github.com/qupath/qupath/wiki/Supported-image-formats). (No conflicts of interest here). Avoid BigTIFF and TIFF formats because image metadata in these formats is often missing or causes downstream problems. Do not use slide scanners with exotic proprietary file formats.
- Files should have unique names and can be numerical or text. Preferably, use ‘001’, ‘010’, ‘100’ instead of ‘1’, ‘10’, ‘100’. A best practice example is *‘Aachen-HCC-0001-Slide001.svs’*. The patient name does not have to be part of the slide file name as the Slide Master Table (see below) takes care of matching slides to patients.
- Patients should have unique non-numerical pseudonyms. Best practice examples are *‘Aachen-HCC-0001’*
- No personal information (such as patient name or date of birth) should be present in any WSI - please check your filenames, image metadata and images for presence of personal information and remove this information
- keep patient pseudonyms and slide names simple and avoid problematic characters such as ‘ ‘, ‘/’, ‘ü’.

## Exclusion criteria for WSI

For the analysis of invasive tumor tissue, slides with following characteristics are excluded from analysis in our standard workflows:

- no tumor on slide (often, slides will not contain any tumor tissue, even in reference databases such as TCGA)
- not H&E stain (sometimes different stains end up in the same cohort by accident)
- Artifacts in >50% of the region of interest or globally altered image quality
- tissue microarray (TMA) extraction holes taking up >50% of region of interest or altering image quality
- blurred (out-of-focus) images or regions (caution: sometimes this is only visible when zoomed in)
- Pen marks covering most of the region of interest

# Annotating the region of interest in whole slide images

## Tumor resections and biopsies

- We generally want to constrain our analysis to regions containing invasive cancer and not include normal tissue or precancerous lesions (Figure 2). This can be achieved either by manual or by automatic tumor detection. In manual workflows, we achieve this by circling the tumor region in QuPath. These annotations can be performed by trained observers in 10-30 seconds per slide. Alternatively, fully automatic tumor detection can be used, as previously published in colorectal cancer.[2] However, in our experience such a tumor detector has to be calibrated for specific tumor types to achieve high accuracy. Often, manually annotating images is faster than training a new tumor classifier.
- In case of manual annotations, we use QuPath v0.1.2[4] to annotate WSI and store the annotations as .qpdata files. Alternatively, annotations can be made using Aperio ImageScope and be stored as .xml files which can be imported into QuPath.
- We make annotations at low levels of magnification and we do not manually exclude smaller-scale non-tumor tissue between tumor glands (Figure 2).
- White background on the slide can be included in the annotations as these regions will be automatically removed later (by a hard brightness threshold).
- Pencil stains will not be automatically removed and should be excluded from annotations
- If you exclude images from the analysis, this has to be documented and the reason has to be recorded ('no tumor on slide', 'poor image quality', etc.).

**Figure 2:** Manual annotation of invasive carcinoma in a scanned whole slide image of a colorectal cancer endoscopic biopsy. More examples are listed below, in the Supplementary Data. Image source: Heike Grabsch, personal communication.

# Working with tiles

## Creating the tiles with QuPath

- To be made accessible to deep neural networks, whole slide images need to be tessellated into tiles. Usually, we only use tiles if their center is within the region of interest

and if their median brightness is less than 220/255 (i.e. at least half the tile area is not background).

- Tiles are stored on the disk in JPG format. The file name of each tile must contain the file name of the parent WSI and its location in the WSI, as shown in Table 4.

| PARENTNAME_(CENTER_X,CENTER_Y).jpg |
|---|

**Table 4**: Example of a tile filename.

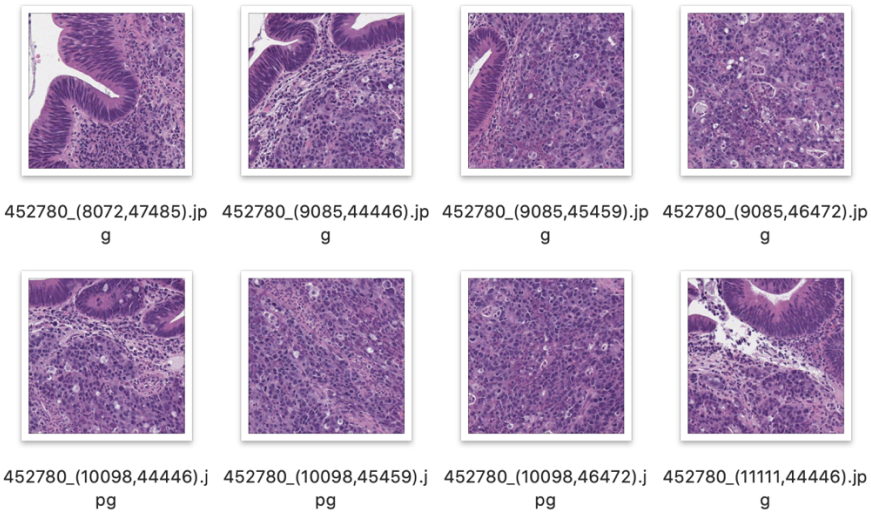


**Figure 3: Examples of tiles.** Tile size is 512x512 px at 0.5 µm/px.

## Storing the tiles

- Each cohort contains many WSI and each WSI generates many tiles
- The folder structure should be as shown in Table 5.

| COHORT_NAME/BLOCKS/WSI_NAME/TILE_NAME.jpg<br><br>*for example, in the cohort "TCGA-STAD-DX", a tile from the WSI "TCGA-ZA-A8F6-01Z-00-DX1.89510833-5BC8-4983-817E-0E9244824168" is:*<br><br>D:/TCGA-STAD-DX/BLOCKS/TCGA-ZA-A8F6-01Z-00-DX1.89510833-5BC8-4983-817E-0E9244824168/TCGA-ZA-A8F6-01Z-00-DX1.89510833-5BC8-4983-817E-0E9244824168_(9412,20271).jpg |
|---|

**Table 5**: Example of a folder name and a tile file name.

# Tile augmentation and preprocessing

- Before the training process, we rescale tiles to be the size of the input layer of the CNN. We usually use 512 px at 256 µm for each tile and use a CNN model with 512x512x3 input neurons.
- Data augmentation is performed on the fly during training to introduce more images for training the classifiers. We randomly flip tiles horizontally and vertically. This process also introduces rotational invariance, which is very important when working with histological slides. Apart from that, we do not consider any other type of image augmentation to be relevant if there are enough WSI available. By providing more data to the classifier overfitting is prevented. Empirically, shearing and other augmentation methods do not improve the performance of the final classifier.

## Conclusion and further steps

This protocol describes all steps from a patient of cohort to a set of image tiles that are ready to use for downstream deep learning analyses. All further steps usually involve training CNNs with algorithms implemented in MATLAB (such as Kather et al.[3]) or Python (such as Fu et al.[5]). However, clean, standardized input data are a prerequisite for such downstream algorithms to achieve high performance.

## References


1. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
2. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
3. Kather, J. N., Heij, L. R., Grabsch, H. I. & Kooreman, L. F. S. Pan-cancer image-based detection of clinically actionable genetic alterations. *bioRxiv* (2019).
4. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
5. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *bioRxiv* 813543 (2019) doi:10.1101/813543.