Medical Image Analysis 97 (2024) 103257

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

The ACROBAT 2022 challenge: Automatic registration of breast cancer tissue

Philippe Weitz [1,a,*], Masi Valkonen [1,b], Leslie Solorzano [1,a], Circe Carr [b], Kimmo Kartasalo [a], Constance Boissin [a], Sonja Koivukoski [c], Aino Kuusela [b], Dusan Rasic [d], Yanbo Feng [a], Sandra Sinius Pouplier [d], Abhinav Sharma [a], Kajsa Ledesma Eriksson [a], Stephanie Robertson [e], Christian Marzahl [f], Chandler D. Gatenbee [g], Alexander R.A. Anderson [g], Marek Wodzinski [h,i], Artur Jurgas [h,i], Niccolò Marini [h,j], Manfredo Atzori [h,k], Henning Müller [h,l], Daniel Budelmann [m], Nick Weiss [m], Stefan Heldmann [m], Johannes Lotz [m], Jelmer M. Wolterink [n], Bruno De Santi [o], Abhijeet Patil [p], Amit Sethi [p], Satoshi Kondo [q], Satoshi Kasai [r], Kousuke Hirasawa [s], Mahtab Farrokh [t], Neeraj Kumar [t], Russell Greiner [t,u], Leena Latonen [c], Anne-Vibeke Laenkholm [d], Johan Hartman [e,v], Pekka Ruusuvuori [2,b,w], Mattias Rantalainen [2,a,v]

[a] *Department of Medical Epidemiology and Biostatistics, Karolinska Insitutet, Stockholm, Sweden*
[b] *Institute of Biomedicine, University of Turku, Turku, Finland*
[c] *Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland*
[d] *Department of Surgical Pathology, Zealand University Hospital, Roskilde, Denmark*
[e] *Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden*
[f] *Gestalt Diagnostics, Spokane, USA*
[g] *Department of Integrated Mathematical Oncology, Moffitt Cancer Center, Tampa, USA*
[h] *Informatics Institute, University of Applied Sciences Western Switzerland, Switzerland*
[i] *Department of Measurement and Electronics, AGH University of Kraków, Poland*
[j] *Department of Computer Science, University of Geneva, Geneva, Switzerland*
[k] *Department of Neuroscience, University of Padova, Italy*
[l] *Medical Faculty, University of Geneva, Switzerland*
[m] *Fraunhofer Institute for Digital Medicine MEVIS, Lübeck, Germany*
[n] *Department of Applied Mathematics, Technical Medical Centre, University of Twente, Enschede, The Netherlands*
[o] *Multimodality Medical Imaging, Technical Medical Centre, University of Twente, Enschede, The Netherlands*
[p] *Department of Electrical Engineering, Indian Institute of Technology, Bombay, India*
[q] *Graduate School of Engineering, Muroran Institute of Technology, Hokkaido, Japan*
[r] *Faculty of Medical Technology, Niigata University of Health and Welfare, Niigata, Japan*
[s] *FORXAI Business Operations, Konica Minolta, Inc., Osaka, Japan*
[t] *Department of Computing Science, University of Alberta, Edmonton, Alberta*
[u] *Alberta Machine Intelligence Institute, Edmonton, Canada*
[v] *MedTechLabs, BioClinicum, Karolinska University Hospital, Stockholm, Sweden*
[w] *Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland*

ARTICLE INFO

Keywords:
Whole-slide-image registration
Computational pathology
Breast cancer
Immunohistochemistry

ABSTRACT

The alignment of tissue between histopathological whole-slide-images (WSI) is crucial for research and clinical applications. Advances in computing, deep learning, and availability of large WSI datasets have revolutionised WSI analysis. Therefore, the current state-of-the-art in WSI registration is unclear. To address this, we conducted the ACROBAT challenge, based on the largest WSI registration dataset to date, including 4,212 WSIs from 1,152 breast cancer patients. The challenge objective was to align WSIs of tissue that was stained with routine diagnostic immunohistochemistry to its H&E-stained counterpart. We compare the performance of eight WSI

* Corresponding author.
E-mail addresses: philippe.weitz@ki.se (P. Weitz), mattias.rantalainen@ki.se (M. Rantalainen).
[1] Contributed equally.
[2] Contributed equally.

https://doi.org/10.1016/j.media.2024.103257
Received 5 December 2023; Received in revised form 17 May 2024; Accepted 24 June 2024
Available online 1 July 2024

registration algorithms, including an investigation of the impact of different WSI properties and clinical covariates. We find that conceptually distinct WSI registration methods can lead to highly accurate registration performances and identify covariates that impact performances across methods. These results provide a comparison of the performance of current WSI registration methods and guide researchers in selecting and developing methods.

1. Introduction

Computational pathology is likely to significantly impact current routine clinical workflows in pathology labs. Applications range from the automation of routine procedures such as cancer detection and Gleason grading of prostate biopsies (Bulten et al., 2022, 2020; Ström et al., 2020) to the prediction of information that pathologists cannot obtain from visually inspecting tissue, including prognosis (Foersch et al., 2023; Wang et al., 2022), treatment response (Foersch et al., 2023) molecular subtypes (Couture et al., 2018; Kather et al., 2020) or gene expression (Fu et al., 2020; Schmauch et al., 2020; Wang et al., 2021; Weitz et al., 2022b). The rapid advancements of these methods in recent years have been enabled by progress in computer vision and the advent of digital pathology. In digital pathology, glass slides with tissue samples are digitised using whole-slide scanners, often at a magnification of 400, referred to as 40X, resulting in whole-slide-images (WSI) with a gigapixel scale. Pathologists then assess WSIs on a screen instead of physical glass slides with a microscope. Currently, the vast majority of methods in computational pathology are limited to WSIs of tissue stained with haematoxylin and eosin (H&E) (Bera et al., 2019),(Baxi et al., 2022). However, the analysis of immunohistochemically (IHC) stained tissue, e.g. for biomarker scoring, is an essential component of the diagnostic workflow. The combination of information from multiple stains has the potential to unlock both novel research and clinical applications. Examples in research are stain-guided learning (Su et al., 2022; Turkki et al., 2016; Valkonen et al., 2020), virtual staining (Burlingame et al., 2020; de Haan et al., 2021; Khan et al., 2023; Wieslander et al., 2021), the analysis of multiplex stained histology (Lin et al., 2023; Schapiro et al., 2022), 3D reconstruction (Kartasalo et al., 2018; Song et al., 2013) and the transfer of annotations or segmentations between WSIs (Duanmu et al., 2022; Huang et al., 2023; Weitz et al., 2023a). In the clinical setting, multi-stain information can aid in the identification of regions of interest (such as invasive cancer) during biomarker scoring, or the investigation of suspicious lesions at resection margins.

The combination of information from multiple WSIs requires the spatial alignment of corresponding tissue areas between WSIs, which is referred to as WSI registration. WSI registration is a particularly challenging registration task due to the gigapixel scale, differences between the appearances of differently stained tissue, changes in appearance, structure and morphology between tissue regions in non-consecutive sections and the introduction of artifacts, tears and deformations during processing of the micrometer-thin tissue sections. The main components of non-rigid multi-modal WSI registration methods are image pre-processing, a method that quantifies similarity and an optimization technique. Current WSI registration methods can be broadly categorized into feature-based and intensity-based methods. Feature-based methods extract local descriptors in both images of an image pair, attempt to find corresponding descriptors and minimize the distance between those. Intensity-based method aim to maximize a similarity metric in typically pre-processed image pairs. Common examples of these metrics are cross-correlation or convolution, mutual information and normalized gradient fields. Furthermore, it is common to split the registration into a coarse initial alignment, followed by a deformable registration step. There are many possible combinations and variations of these techniques, without a consensus on optimal choices. Furthermore, emerging deep learning methods are now also applied in WSI registration. While there has been research in this area for many years, non-rigid multi-modal WSI registration is therefore an active field of research.

To establish a comparison of the performance of current WSI registration methods in data originating from clinical workflows, we organized the ACROBAT (Automatic Registration of Breast Cancer Tissue, acrobat.grand-challenge.org) challenge. For this challenge, we published the currently largest publicly available data set of matched H&E and IHC WSIs, consisting of 4212 WSIs in total (Weitz et al., 2022a) and generated over 54,000 landmark points with 13 annotators. All WSIs in the data set originate from tissue sections that were generated during routine diagnostic workflows at the time of initial diagnosis. An example of an H&E WSI and corresponding IHC WSIs is depicted in Fig. 1. The objective of the challenge was to align tissue in the IHC WSIs to corresponding tissue in the H&E WSIs. WSI registration has previously been addressed in the ANHIR challenge (Borovec et al., 2020). While the ANHIR challenge made valuable contributions to the field of WSI registration, it was limited by the high quality of sections and WSIs, which is not representative of clinical material, as well as by the availability of both training and test data with only 355 WSIs in total, albeit originating from a wide variety of organs and stains.

Here, we describe the results of the ACROBAT registration challenge, in which we assess the performances and limitations of eight WSI registration algorithms. We evaluated accuracy and robustness of each method and performed a detailed analysis of the impact of different clinical covariates, such as cancer grade and biomarker statuses, and the registered tissue types, which to our knowledge is the first analysis of this kind in the histopathology domain. We expect that the results of this challenge will clarify the performance of current methods for multi-stain WSI registration and provide evidence for the integration of the analyzed methods into future research studies and clinical applications. The findings of this study can furthermore guide the development of WSI registration solutions that generalize between stains and tissue types and that can be applied to WSIs that originate from routine diagnostics.

2. Methods

2.1. Challenge design

The ACROBAT challenge took place between April and September 2022. The objective of the challenge was the fully automatic registration of test set landmarks that were provided for the IHC WSIs to their H&E counterparts. For the training data, no landmarks were available, methods that were optimized with the training data therefore needed to be optimized in an unsupervised manner. The use of external training data was permitted. Both for the validation and test data, IHC landmarks were published. The data set is available under a CC-BY license from the Swedish National Data Service (SND). Registered validation set landmarks could be submitted up to two times daily on the challenge website (acrobat.grand-challenge.org) to receive automated feedback on the algorithm performance. This submission system will remain open indefinitely. Test set performance was computed by the challenge organizers after the end of the challenge timeframe and no feedback on algorithm performance in the test set was available before the challenge workshop, which was held in conjunction with the MICCAI (Medical Image Computing and Computer Assisted Intervention) 2022 conference in Singapore. Details on the challenge timeline are available in Supplementary Table 1. To prevent information leakage, members of the organizers' institutions but not departments were allowed to participate in the challenge. During the challenge, there were 221 submissions by

16 teams for the validation data. Eight methods qualified to be evaluated in the test set by submitting test set landmarks and an algorithm description before the challenge deadline. These eight methods are assessed in this publication.

2.2. Data set

The ACROBAT data set consists of 4212 WSIs from 1153 female patients with primary breast cancer. Data was collected from the SöS study and patients originate from Sjödersjukthuset in Stockholm. All WSIs of a case contain tissue from the same tumor block, but sections are not necessarily consecutive. The cases were divided into a training set consisting of 750 cases (3406 WSIs), a validation set consisting of 100 cases (200 WSIs) and a test set consisting of 303 cases (606 WSIs). The number of cases included in the challenge was decided based on data availability for the training data and annotation feasibility for the validation and test data. Clinical characteristics of the test set cases are available in Supplementary Table 2. For each case in the training set, one H&E WSI and up to four IHC WSIs from the routine diagnostic stains, which are the nuclear stains ER, PGR, and KI67 and the membrane stain HER2. In the validation and test set, there is one H&E WSI and one randomly selected IHC WSI out of the four IHC stains available. All WSIs were digitized on either one of two NanoZoomer XRs or a NanoZoomer S360 at ca. 0.23 μm/pixel. The ACROBAT data set was published as pyramidal TIFF WSIs with a resolution of 0.92 μm/pixel at the highest magnification. This reduces the data set size from 10.13 TB to 482 GB, which facilitates data transfer, likely without any impact on the challenge, as current WSI registration methods typically operate at much lower magnifications. Further details of the data generation and processing workflows are available in the ACROBAT data set descriptors (Rantalainen and Hartman, 2023; Weitz et al., 2023b). Fig. 1 shows an example of an H&E WSI from the data set with corresponding IHC WSIs.

2.3. Landmark annotations

Registration performance in the ACROBAT challenge was quantified based on landmark annotations, which is generally the standard approach to evaluate registration performances. Annotations were generated by 13 members of the ABCAP research consortium (abcap.org), all of which have received histopathology education and who have previously worked with WSIs in research projects. Two of the annotators have pathologist training. Annotations were generated using a version of TissUUMaps (Solorzano et al., 2020) that was customized for the ACROBAT challenge. All annotations were generated using WSIs with 40X magnification as the highest resolution. For the validation data, annotations for each image pair were generated by a single annotator, whereas in the test data, each image pair was annotated by two annotators. Annotations were generated in two phases, the first of which was applied to both the validation and test data, whereas the second was only applied to the test data. Annotation protocols for both phases, including an example of what constitutes corresponding landmark locations, are available online (github.com/rantalainenGroup/ACROBAT).

During the first phase, annotators were asked to place 50 corresponding landmarks in an H&E-IHC image pair that was displayed side-by-side in TissUUMaps, placing first the IHC and then the H&E landmark. For the second phase, landmarks in the IHC WSIs were fixed in place and displayed, whereas landmarks in the H&E WSIs were randomly moved by ± 500 pixels, which corresponds to ±115 μm at 40X magnification. A second annotator was then selected randomly such that the first and second annotator were always different, with 87 combinations between annotators present in the data. The second annotator was then tasked with moving the H&E landmark to the locations that they considered to match the displayed IHC landmark. This results in 10,040 landmarks in the validation and 44,760 landmarks in the test set. In total, 13 annotators annotated 54,800 points with an average of 49.24 landmarks per image. Finding corresponding landmark locations from the images was challenging for some image pairs due to the drastic difference in visual appearance and the number of annotations for those was often less than the set target of 50 point pairs per image pair. Together, from the test and validation sets 83 % of the image pairs had exactly 50 landmarks, 6 % less than 50 landmarks, and 11 % above 50 landmarks. The majority of landmarks were placed in close proximity to each other by the first and second annotator, with 50 % of the landmarks less than 20 μm apart and 80 % less than 60 μm.

We then proceeded to exclude landmarks in the test set for which the distance between the location selected by the first and second annotator exceeds 115 μm to filter out annotations with low inter-annotator agreement. The distribution of distances between annotators is depicted in Fig. 2d). We furthermore excluded WSIs from evaluation for which fewer than 10 landmarks remained after this exclusion to ensure sufficient landmark density. This results in 13,130 landmarks and 297 WSIs included for final performance evaluation.

For 290 out of these 297 H&E WSIs in the test set, semantic

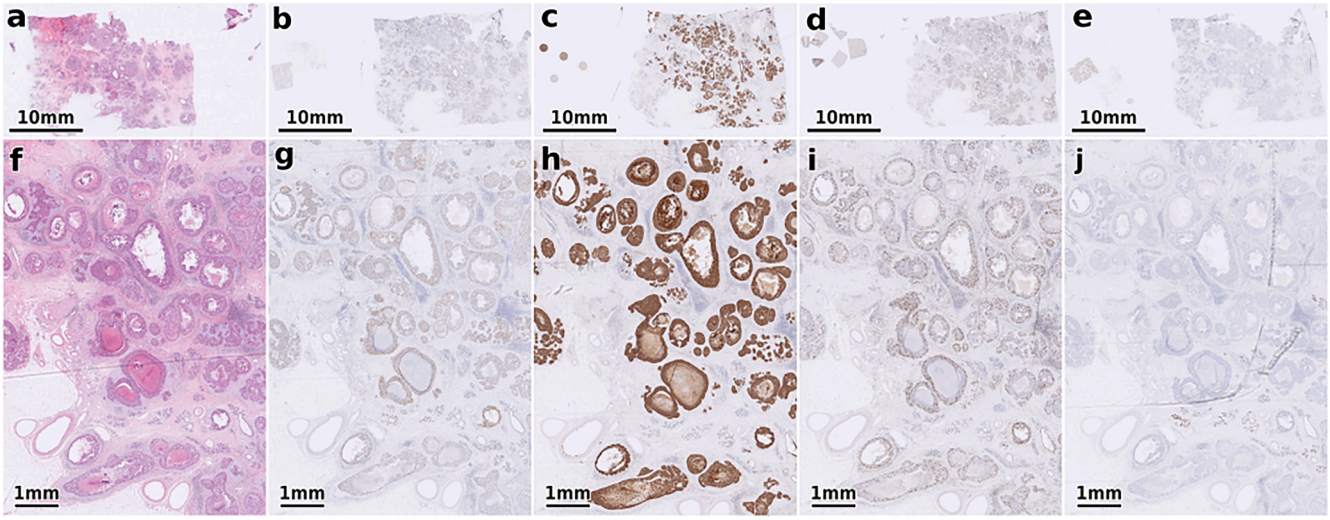

Fig. 1. Example of an H&E stained tissue section and corresponding IHC stained tissue. The first row (a-e) shows an overview over the entire WSIs, whereas the second row (f-j) shows corresponding tissue regions at a higher magnification. a) and f) show the H&E WSI, b) and g) tissue stained with ER IHC, c) and h) with HER2 IHC, d) and i) with KI67 IHC and e) and j) with PGR IHC.

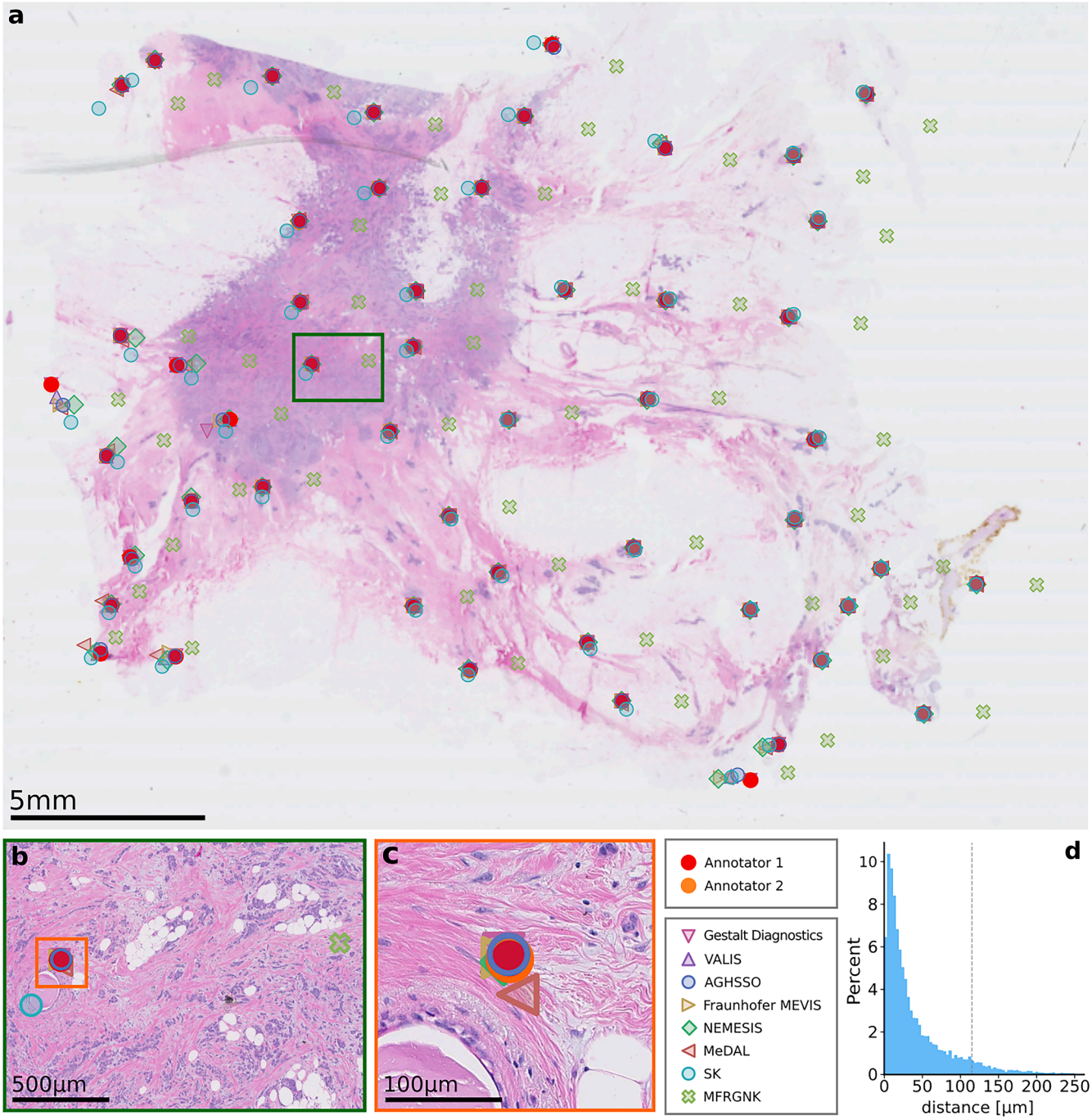

Fig. 2. Example of a target H&E WSI and distribution of DBAs. a) shows an overview of a H&E WSIs with annotated and registered landmarks. b) and c) depict a closer view of a specific landmark. d) shows a histogram of the DBAs, where the dashed line indicates the DBA exclusion threshold for performance metric computation. The histogram is capped at 250 µm.

annotations were generated by a trained pathologist who specializes in breast cancer. Annotations include the classes invasive cancer (IC), ductal carcinoma in situ (DCIS), lobular carcinoma in situ (LCIS), non-malignant changes (NMC), artifacts, lymphovascular invasion (LI) and normal tissue. We can therefore assign one of these classes to the majority of landmarks in the test set for analysis. Supplementary Table 3 lists the percentages of landmarks in each class, as well as the area percentage of each class in the total tissue area. Normal tissue and artifacts were under-annotated, with factors of 0.79 and 0.53. This could be explained by a lack of structures in normal tissue. Artifacts may only exist in one of the two images of an image pair. The proportion of IC landmarks corresponds to the proportion of IC area in the data set, with a factor of 1.03. DCIS, LCIS, NMC and LI were over-annotated, with factors of 4.73, 7.77, 2.88 and 9 respectively. For 3.64 % of landmarks, the class assignment differs, which might e.g. be common for landmarks that were placed on edges of structures.

2.4. Performance evaluation & ranking

Performance evaluation was based on the target registration error (TRE). For each registered landmark, there are two target landmarks by two different annotators. The H&E WSIs are the target images for the registration, whereas the IHC WSIs are the source images. The transformation that is found during the registration is therefore applied to the

IHC landmarks, to transform their coordinates to the H&E coordinate system. For each registered IHC landmark, we computed the distance in micrometers to each of these two H&E target points and used the mean distance as the error distance, which we will refer to as TRE. Within each WSI, we aggregated these error distances into a WSI-level score by taking the 90th percentile of the distances. We chose the 90th percentile to emphasize robustness. In the case of missing landmarks in the submissions, we used the coordinates of the unregistered landmarks in the source image, capped at the image borders, to compute error distances. Submissions were then ranked on the median of the 90th percentiles. Submissions were ranked in two leaderboards, one including all eligible test set submissions, the other one only those submissions for which the code was made publicly available. Monetary prizes were then allocated to the first three teams in each leaderboard. All team members of all teams ranked in the test set leaderboard were invited to contribute to the publication.

Beyond the median 90th percentile, we also computed the 90th percentiles of 90th percentiles, the mean 90th percentiles and the mean and median error distances across all landmarks without slide-level aggregation. Furthermore, we computed the mean reduction in the TRE in percent from unregistered landmark locations to the target locations.

2.5. Linear mixed effects model analysis

In order to investigate the impact of different properties of individual landmarks on the resulting error distances, we fitted Linear Mixed Effects (LME) Models with the *R* package *lme4(Bates et al., 2015)*. One LME was fitted for each team and one for the annotators. A LME is a linear model of the form $y = X\beta + Zu + \epsilon$. Here, y is a vector of the log10-transformed TREs for the teams and log10-transformed distances between annotators (DBAs) for the annotators for all landmarks in micrometers. The log-transform was necessary to ensure that residuals were approximately Gaussian. For the DBA LME model, we chose 1 mm as the exclusion threshold. β represents the fixed effects coefficients and u the random effects coefficients, with X and Z as matrices that contain the values of observations of covariates in their rows. ϵ denotes a random error term. Fixed effects covariates are independent of each other, whereas random effects are sampled from the same statistical units. Here, we consider each WSI as a statistical unit containing multiple landmarks, with 272 units for which all required information is available for inclusion in the LME analysis. Furthermore, the combination of first and second annotator is considered as a statistical unit, with 86 units in total. Both of these are therefore modeled as random effects in the LMEs. We included 16 fixed effects into the analysis. There are two continuous covariates, the distance of landmarks from the center of mass of the respective tissue mask in mm and the slide age. The slide age ranges from 0 to 5 years and indicates the time between sample preparation and scanning. For slides with a high slide age, the staining may have faded to some degree. HER2, PGR, KI67 indicate the IHC antibody used in the IHC WSI of an image pair, with ER as the reference category. With respect to semantic segmentation classes that landmarks are positioned in, landmarks can be assigned IC, artifact, DCIS, LCIS and NMC, with normal tissue as the reference category. We excluded landmarks with disagreeing tissue class between first and second annotator and LI landmarks, since there are too few of these to model. IC:NHG2 and IC:NHG3 indicate the grading of the invasive cancer region for the landmarks within the cancer region for an image pair, with IC:NHG1 as the reference category. IC:BS:KI67, IC:BS:HER2, IC:BS:PGR and IC:BS:ER indicate the clinical biomarker status (BS) of the respective antibody for landmarks within the invasive cancer region of a WSI. The biomarker status is considered as positive for ER and PGR above a threshold of 10 % and 20 % for KI67. HER2 was considered as negative based on IHC scores 0 to 1+, positive for IHC scores 3+ and negative or positive for 2+ depending on additional in situ hybridisation that assesses gene amplification. Biomarker statuses were assigned according to clinical guidelines at the time of diagnosis. Coefficient values whose 95 % confidence interval does not include 0 will be considered as different from 0 and therefore as associated with the error distance. Descriptions, ranges and rerferences of LME covariates are available in Supplementary Table 4. As shown in Supplementary Figure 1, the slide age, scanner model and presence of control tissue in the IHC WSI of an image pair are highly correlated. We therefore only included the slide age into the LME analysis to avoid collinearity. A scatterplot of the LME model residuals, a quantile-quantile plot and a plot of the density of residuals are available in the section of the Supplement that describes the algorithm of the respective team.

3. Results

3.1. Deep learning has become an ubiquitously used tool for WSI registration that complements classical image analysis techniques

A wide range of conceptually different WSI registration approaches were used in the challenge. Some methods relied on traditional image processing while others were based on deep learning. A common design pattern was to split the registration task into multiple steps, starting with image preprocessing, followed by an initial alignment and subsequently a deformable registration. The preprocessing step aims at normalizing or discarding information e.g. by color space transformation, contrast normalization and downscaling to focus registration on essential information and simplify the problem for the subsequent steps. Teams VALIS, Fraunhofer MEVIS, and MeDAL target computations to meaningful areas through an initial tissue segmentation and then focus registration only on the detected tissue area. The initial alignment step roughly aligns images using translation, rotation, reflection, scaling or affine transformations in order to simplify the task for the remaining deformable step. The deformable step involves elastic or curvature-controlled transformations that attempt to completely align image contents and can account for complex deformations of the tissue. Registration steps were typically performed iteratively starting at low resolutions with increasing resolutions during subsequent steps. A common division of registration algorithms is into intensity-based and feature-based methods. Intensity-based methods rely on finding correspondences between image intensities based on a similarity function, whereas feature-based methods extract features, such as points, and extract a descriptor from their local neighborhood to establish matching feature pairs between images.

Six out of the eight analyzed teams used feature-based registration, with some relying on more recent approaches such as SuperPoint (DeTone et al., 2017) and SuperGlue (Sarlin et al., 2020) while others used more classical approaches like SIFT , BRISK (Leutenegger et al., 2011) and RANSAC (Fischler and Bolles, 1981). For the intensity-based parts of participants' algorithms, the most common similarity criterion was cross-correlation (CC) and its variants normalized cross-correlation (NCC) and convolution. Only team Fraunhofer MEVIS employed normalized gradient fields (NGF) (Haber and Modersitzki, 2007), which bases its matching on intensity gradient orientations. No participant proposed a solution utilizing mutual information. Seven out of the eight top performing teams applied deep learning techniques in parts of their workflows. This shows deep learning has further permeated multi-stain WSI registration since the ANHIR challenge in 2019, where only one method applied deep learning. Table 1 summarizes all the aforementioned aspects and underlines which parts of the groups' workflows involve deep learning, either during preprocessing or registration. For instance, Gestalt Diagnostics and AGHSSO used the graph neural network-based SuperGlue (Sarlin et al., 2020) for feature matching, and SK used the convolutional neural network-based registration framework Voxelmorph (Balakrishnan et al., 2019).

The use of external data was allowed in this challenge. The three teams Fraunhofer MEVIS, Gestalt Diagnostics and AGHSSO used external data for their algorithm development. Fraunhofer MEVIS

Table 1
Summary of methods. A summary of each team's method with descriptions of their main workflow. More detailed descriptions are available in the Supplementary Materials. Underlined elements are where deep learning is used.

Team	Code available	Uses DL	Initial alignment	Initial alignment criterion	Deformable transformation	Deformable criterion	Optimization	Preprocessing
Gestalt Diagnostics		✓	Rotation search with Superglue/OpenGlue/LoFTR & RANSAC	No. of keypoint matches	Local affine for triangular partitions	RANSAC inliers	–	Contrast-normalising, random equalize, random sharpness, non-maximum suppression
VALIS	✓	✓	BRISK keypoints VGG descriptors & RANSAC	RANSAC inliers, Tukey inliers	DeepFlow(Weinzaepfel et al., 2013)	DeepFlow's DeepMatch	–	Downsample, foreground segmentation, grayscale, invert intensities, intensity normalisation with Akima interpolation.
AGHSSO	✓	✓	SIFT/SuperPoint & RANSAC/ SuperGlue	Sparse descriptor error	Multi-level and weighted local NCC optimization	NCC	Adam	Downsample, grayscale, invert intensities and equalize with CLAHE
Fraunhofer MEVIS		✓	Align centers of mass rotational search with NGF	NGF	Optimise control point grid with linear interpolation	NGF	L-BFGS	Downsample, grayscale, foreground segmentation
NEMESIS	✓	✓	SIFT & RANSAC tissue mask overlap	RANSAC inliers + Dice score	Implicit neural representation of transform with MLP(Wolterink et al., 06–08 Jul 2022)	NCC	Adam	Downsample, grayscale, histogram equalisation, gaussian smoothing
MeDAL	✓		Template matching (mask convolution, 1° intervals)	Convolution maximum	RBF in local keypoints found from tissue masks	convolution	–	Downsample, foreground segmentation, grayscale
SK		✓	Template matching (NCC, 0° and 180° rotations)	NCC	Voxelmorph(Balakrishnan et al., 2019)	MSE	Adam	Downsample, grayscale, WSI border removal
MFRGNK	✓		ORB & RANSAC with projective transform	RANSAC inliers	–	–	–	Downsample, Macenko colour normalization(Macenko et al., 2009)

trained a tissue segmentation model with external data, Gestalt Diagnostics used external data to optimize their final parametrization and AGHSSO further validated their model with data not provided by the organizers. More detailed summaries and additional information for each of the methods is available in the Supplementary Materials.

3.2. *Diverse registration methodologies result in several performance clusters, one method approaches human annotator accuracy*

The primary ranking metric for the challenge was the median 90th percentile of the target registration error (TRE) within each WSI pair. To reduce the effect of human error, TREs were computed by averaging the distance in μm of registered landmarks to two target landmarks placed by two different annotators. An example of a target H&E WSI, with annotator and registered landmarks is depicted in Fig. 2a-c. As a quality control step, we excluded landmarks with poor agreement between annotators from performance metric computations based on distance between annotators (DBA) >115 μm, as indicated in the histogram of DBAs in Fig. 2d. In total, 13,130 pairs of landmarks in 297 image pairs were included for metric computation. The distributions of the 90th percentiles in the validation and test data are depicted in Fig. 3a. Fig. 3b depicts scatterplots of 90th percentiles against the 90th percentiles of DBAs. Based on the primary ranking metric, the algorithm developed by Gestalt Diagnostics achieved the best score, with a median 90th percentile of 60.1 [55.8, 68.6] μm. This is approximately half the median 90th percentile of the methods that follow in the ranking, starting with VALIS with 123.3 [98.5, 144.1] μm, AGHSSO with 137.6 [120.3, 176.7] μm and Fraunhofer MEVIS with 155.3 [123.1, 184.7] μm. The solutions of NEMESIS and MeDAL are in the range of three to four times the lowest median 90th percentile, with 200.5 [176.7, 257.1] μm and 262.5 [225.4, 322.5] μm respectively. The median 90th percentile of SK of 1230.0 [1141.0, 1341.5] μm is one order of magnitude higher and the solution of MFRGNK one order of magnitude higher compared to SK with 15,938.0 [15,117.0, 16,598.6] μm. A comparison of the distributions of the 90th percentiles in Fig. 3a with two-sided Mann-Whitney U rank tests indicates that for Gestalt Diagnostics, 90th percentiles of TRE differ between the validation and test set, with Benjamini-Hochberg (BH) adjusted p-value < 0.01, with lower TREs in the test set. For the other methods, this comparison reveals no differences. Both Fig. 3a and 3b indicate that for all methods except Gestalt Diagnostic, there are outlier image pairs with considerably higher 90th percentiles of TRE. E.g. for VALIS, there is a higher number of outlier image pairs with poor registration quality compared to AGHSSO, which the median 90th percentile is robust against, but not the mean. Correspondingly, Fig. 3c shows the ranking for each metric that is available in Table 2. The rankings are mostly stable across metrics. Only the algorithm proposed by VALIS is ranked lower compared to AGHSSO regarding the mean 90th percentile and the median error distance across all landmarks, as well as the mean distance reduction and AGHSSO and Fraunhofer MEVIS regarding the mean error distance across all landmarks. Supplementary Figure 2 shows the stability of the ranking with the median 90th percentile for varying exclusion thresholds in μm for the DBA. Only for exclusion thresholds below 70 μm, the ranking between AGHSSO and Fraunhofer MEVIS begins to depend on the threshold. Paired two-sided Wilcoxon signed rank tests indicate that the distributions of 90th percentiles are different between all submissions (compare Supplementary Figure 3) with BH-adjusted p-values < 0.01 for each comparison. Supplementary Figure 4a shows the Spearman correlations of median 90th percentiles in the test set, which reveals a cluster of correlations for the six top-performing methods with correlations ranging from 0.61 to 0.93.

All median 90th percentiles, along with the 90th percentiles of 90th percentiles, mean 90th percentiles and the median and mean across all landmarks without WSI-wise aggregation are listed in Table 2. It also contains the slide-wise aggregated mean reduction in distance between source and target landmarks in percent, which may guide intuition on algorithm performances.

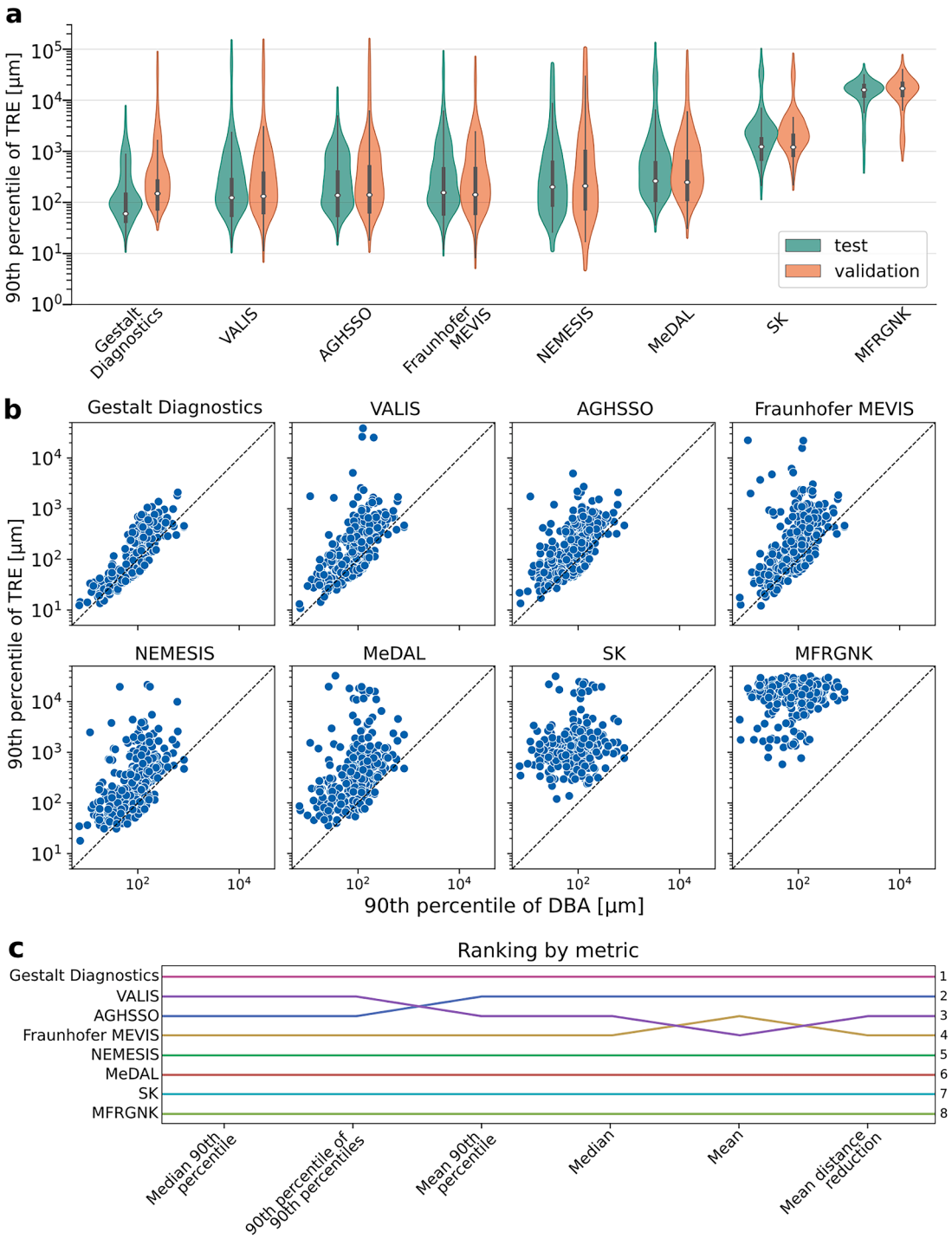

Fig. 3. Overview of 90th percentiles of TRE and rankings. a) shows violin plots of the distributions of 90th percentiles of TREs in the validation and test data for all eight investigated methods, while b) shows scatterplots of the WSI-wise 90th percentiles of DBAs compared to the corresponding 90th percentiles of TREs, excluding landmarks with a DBA > 1 mm when computing 90th percentiles for annotators. c) shows the ranking of methods for each of the metrics that are listed in Table 2.

To contextualize algorithm performances, we also computed all metrics for the DBA. The median 90th percentile of DBAs is 67.0 [62.2, 72.4] µm and therefore slightly higher compared to the value of Gestalt Diagnostics of 60.1 [55.8, 68.6] µm. It is possible to achieve a lower TRE than DBA by registering a landmark to a location between the two landmark positions chosen by the two annotators, as shown in Supplementary Figure 5. The mean 90th percentile of DBAs of 63.5 [60.3, 66.6] µm however is lower than the lowest corresponding TRE by Gestalt Diagnostics of 160.0 [134.0, 189.4], which is also the case for the mean across all landmarks of 31.1 [30.6, 31.6] µm, compared to 63.3 [60.1, 66.6] µm for the best-performing algorithm.

We also investigated failure cases to identify how algorithms can be improved further. The image pair with the worst mean 90th percentile across methods is depicted in Supplementary Figure 6. The high degree of cropping does not allow for the detection of a tissue outline and therefore a reliable initial alignment for some methods. While this impacts algorithms significantly, with 90th percentiles of 440.90 µm for Gestalt Diagnostics, closely followed by MeDAL with 442.61 µm and NEMESIS with 586.73 µm, it is not challenging for human annotators to find corresponding landmarks, with a 90th percentile of distances

Table 2
Metric values for the primary challenge metric, the median 90th percentile of error distances across WSIs, alongside further metrics that could be used to rank algorithm performances. Median and mean were computed on the landmark-level, without previous aggregation on the WSI-level. The mean distance reduction indicates the mean reduction in distance between source and target landmark position due to the registration. Confidence intervals were obtained by bootstrapping with 10,000 bootstrap samples.

Team	Median 90th percentile [μm]	90th percentile of 90th percentiles [μm]	Mean 90th percentile [μm]	Median [μm]	Mean [μm]	Mean distance reduction [%]
Gestalt Diagnostics	60.1 [55.8, 68.55]	449.64 [345.97, 535.96]	159.99 [132.78, 188.43]	22.29 [21.8, 22.72]	63.29 [60.05, 66.59]	98.97 [98.73, 99.18]
Annotators	66.99 [62.19, 72.37]	97.72 [95.26, 101.44]	63.47 [60.34, 66.66]	21.27 [20.79, 21.8]	31.09 [30.61, 31.56]	n.a.
VALIS	123.32 [98.49, 144.12]	694.45 [580.41, 857.44]	578.93 [274.38, 966.9]	37.98 [37.01, 38.99]	313.36 [275.43, 353.76]	97.41 [96.21, 98.33]
AGHSSO	137.63 [120.9, 175.65]	713.4 [604.16, 838.94]	303.16 [256.04, 358.43]	34.64 [33.75, 35.46]	122.21 [117.03, 127.63]	98.2 [97.86, 98.5]
Fraunhofer MEVIS	155.29 [123.1, 184.65]	1019.5 [856.2, 1343.29]	604.35 [393.19, 870.49]	40.88 [39.93, 41.93]	294.14 [269.65, 319.55]	96.56 [95.37, 97.59]
NEMESIS	200.47 [176.69, 257.13]	1308.64 [1013.13, 1834.37]	733.04 [510.86, 1010.42]	62.72 [60.96, 64.26]	349.89 [325.62, 375.91]	95.79 [94.37, 97.0]
MeDAL	262.49 [225.44, 322.47]	1607.82 [1177.53, 2558.87]	1221.55 [838.07, 1673.09]	81.9 [79.63, 83.92]	721.24 [674.45, 768.83]	93.24 [91.18, 95.13]
SK	1230.01 [1141.98, 1341.52]	3292.55 [2539.85, 4722.35]	2438.45 [1956.03, 2981.57]	628.44 [612.02, 644.17]	1524.35 [1466.42, 1582.3]	84.3 [81.63, 86.83]
MFRGNK	15,938.02 [15,117.95, 16,576.21]	22,946.95 [21,964.99, 23,844.39]	15,342.27 [14,576.8, 16,107.18]	9224.71 [9064.66, 9400.27]	9988.07 [9876.3, 10,101.04]	29.23 [26.38, 32.07]

between first and second annotator of 98.47 μm.

3.3. Linear mixed effects model analysis reveals covariates that consistently impact TREs across algorithms

In order to identify which properties of the landmarks and image pairs impact algorithm performances, we conducted a linear mixed effects (LME) model analysis for each team, with the log10-transformed landmark-wise TREs or DBAs as the endogenous variable. The analysis is adjusted for the slide ID and combination of first and second annotator as random effects. Percentage changes for one unit increase of the fixed effects for the LMEs for the TREs of the six best performing teams and the DBA are depicted in Fig. 4a. Supplementary Figure 7 shows the corresponding fixed effects coefficients, whose values are available in Supplementary Table 5. Due to the log-transform, percentage changes of the respective reference TRE accumulate multiplicatively across effects. For covariates for which the 95 % confidence intervals of most or all fixed effects include zero, it is nevertheless possible to observe and interpret trends.

As depicted in Fig. 4a, the antibody of the stain of the IHC WSIs in the image pairs compared to ER as the reference category appears to not impact algorithm performances across methods, potentially with the exception of HER2, where the point estimates of the effect sizes are consistently below zero. For each landmark, a semantic segmentation of the surrounding tissue is available, including invasive cancer (IC), non-malignant changes (NMC), artifacts, ductal carcinoma in situ (DCIS), lobular carcinoma in situ (LCIS) and normal tissue as the reference class. With the exception of artifacts, landmarks in all segmentation classes are associated with a lower TRE compared to normal tissue. This effect is particularly pronounced for NMC, DCIS and IC. For landmarks within the IC regions, we also included the Nottingham histological grade (NHG) as an interaction. Both for NHG 2 and 3, which have less clearly defined growth patterns than NHG 1, the registration error increases across methods. Besides the NHG, we also modeled an interaction between the IC region, the biomarker status (BS) as assigned at time of diagnosis and the IHC stain. It appears that there is a trend towards higher TREs for landmarks within HER2-positive IC regions and potentially a weaker trend towards higher TREs in PGR-positive IC regions. KI67 and ER BS within IC regions are not associated with TRE.

Besides categorical fixed effects, we also analyzed two continuous effects, the slide age and the distance of a landmark to the center of tissue mass. The increase in error in percent with increasing units is depicted in Supplementary Figure 8. The slide age is strongly associated with the TRE. With the exception of Gestalt Diagnostics, all teams have an increase of TRE compared to the respective reference of approximately 100 % at four years. For NEMESIS and MeDAL, TRE is also relevantly associated with the distance to the center of tissue mass with an increase of 60 % at 10 mm, at which there is an increase of ca. 20 % for VALIS, AGHSSO and Fraunhofer MEVIS. In contrast, there is a weak negative association for Gestalt Diagnostics and the annotators.

Fig. 4b shows the distributions of changes in percent for the estimated conditional means for the random effects. The interquartile range for the annotator combination is highest for the DBA, followed by Gestalt Diagnostics and decreasing with decreasing ranking, whereas the interquartile range for the slide ID is lowest for the annotators, followed by Gestalt Diagnostics and roughly increasing with decreased ranking. The correlations between conditional means for the slide ID are shown in Supplementary Figure 4b and closely resemble the correlations of the 90th percentiles, with a cluster for the six highest ranked methods but weaker correlations with the annotators.

Across fixed effects, the direction of statistically significant coefficients is the same throughout teams, with the exception of the distance to center for Gestalt Diagnostics. The DBA is associated with fewer fixed effects than the registration methods, and effect sizes are generally smaller.

4. Discussion

We organized the ACROBAT challenge to compare the performance of current multi-stain WSI registration algorithms and to test the applicability of current solutions for a real-world data set with slides from clinical routine. We published the largest-to-date data set for histology image registration, placed over 54,000 landmark points with 13 annotators for performance quantification and conducted an in-depth analysis of registration methods including clinical information and tissue segmentations.

Out of 16 teams that submitted registered landmarks for the validation data on the challenge website, 8 qualified to be ranked in the test data based on submitted registered test set landmarks and an algorithm description. We attribute this to the requirement for publishing an algorithm description, which might not have been considered worthwhile by some teams that did not accomplish a high rank in the validation data. Furthermore, publicizing details of the respective registration method might not have been desirable for some teams for IP reasons. We see the automated evaluation of registration performance in the validation data as a service to the WSI registration research community that

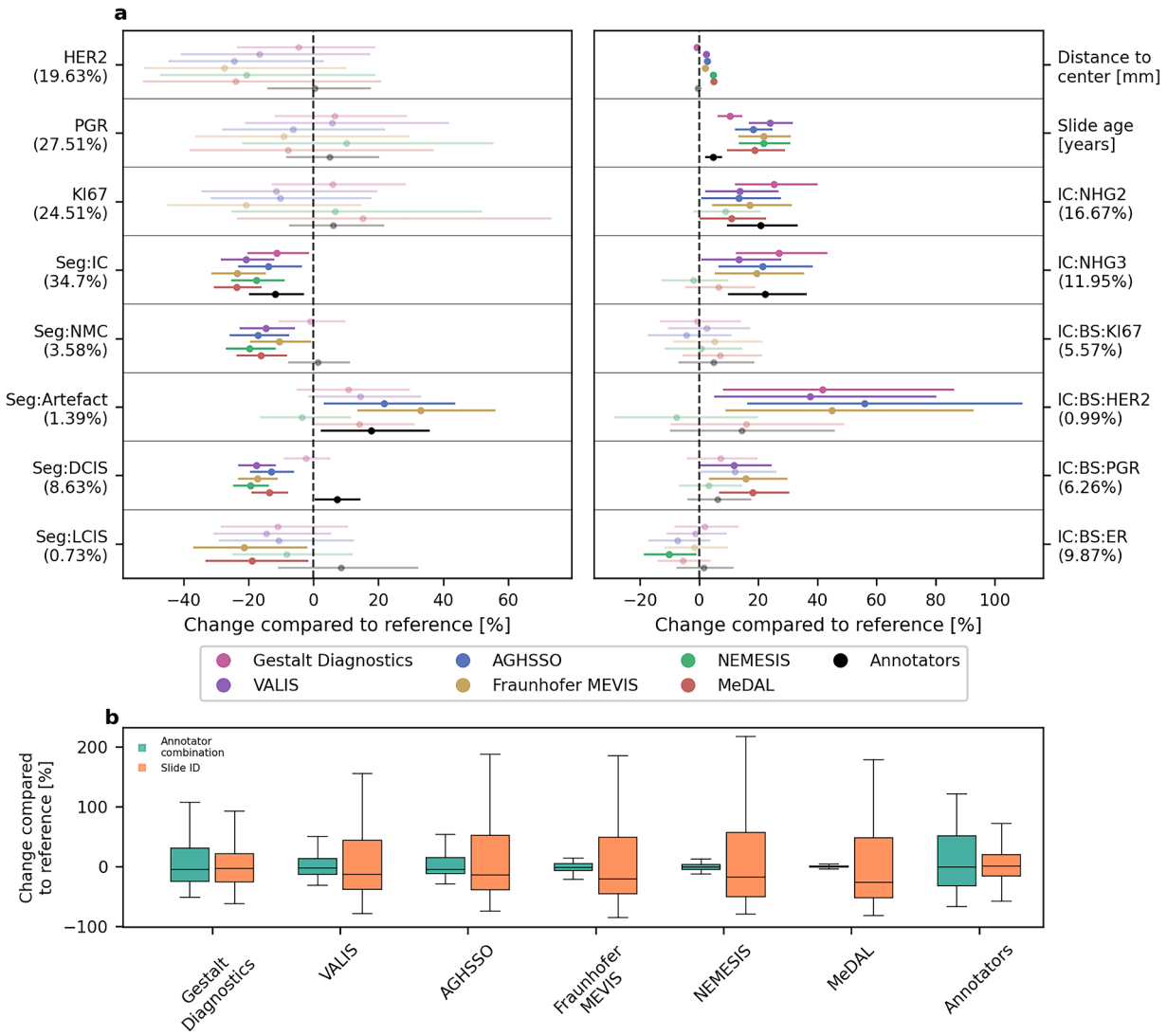

Fig. 4. Coefficients and conditional means of random effects of the LME analysis for the TREs of the six highest ranked teams and the DBA for annotators. a) shows the percentage changes in TRE or DBA for one unit increase of fixed effects coefficients with 95 % confidence intervals. Transparency of the respective marker is increased if the confidence interval includes zero. For categorical fixed effects, there are indications of the percentage of landmarks that are part of the respective category. For continuous effects, the unit is indicated. Effects starting with *Seg* indicate landmark tissue classes, with normal tissue as the reference class. b) shows boxplots that represent the distributions of the percentage changes of estimated conditional means of the random effects for the annotator combination and slide IDs. Boxes include the lower to upper quartile of data. Whiskers extend 1.5 times the interquartile range from the box outlines or the minimum or maximum value. Outliers outside of this range are not shown but are available in Supplementary Figure 7.

is independent from the challenge. Some teams might have submitted validation set landmarks without the intention to participate in the competition in the test data. We therefore do not consider this drop-out rate as a short-coming of the challenge design.

A wide range of conceptually different approaches were evaluated in the challenge, showing that since the ANHIR challenge, deep learning has become an impactful tool also in the WSI registration domain. The best performing method was by Gestalt Diagnostics in terms of accuracy and robustness. In this method, a tree structure of triangular partitions is constructed using DL-based feature matching. Most top performing methods relied upon feature-based registration, which shows its effectiveness with challenging data sets. Also the use of external data may have played a role in the outcome as its utilization increased at the better ranks. The obtained results suggest that feature-based approaches are able to achieve higher robustness than alternative methods, especially those using modern DL-based methods such as SuperPoint and SuperGlue. Feature-based methods may be less impacted by certain imperfections in the data such as tears. They have the advantage of being independent of foreground tissue segmentation and therefore make no assumptions of the presence of background. Intensity-based registration with cross-correlation is able to produce good results with grayscale converted H&E and IHC images when compared to more advanced multi-modal similarity metrics such as the NGF. One possible explanation for the high performance of the method by Gestalt Diagnostics is its fully feature-based approach. Adjacent histological sections have structures that are shared between them but also smaller structures that are not shared such as some of the cells, which can appear only on one of the sections. Feature-based registration may be able to focus on those shared structures between the images through the feature matching step. Another possible explanation is the lack of regularization in the applied transformation. Other teams among the four best performing methods regularized the transformation to produce smooth deformations, which however comes at the cost of accuracy. While registration without regularization can produce smaller errors in a setting with landmark based evaluation, there can be practical reasons to trade off accuracy for smoother transformations, e.g. if downstream analyses

of the registered images require less distorted tissue.

While a wide variety of image pre-processing and registration methods has been deployed in this challenge, it does not cover all methods that have been used in WSI registration during the last years. For example, no team used mutual information as a similarity metric in an intensity-based registration approach. Furthermore, while a wide variety of methods for image pre-processing, similarity quantification and optimization has been used in this challenge, there are many combinations between the methods used by individual teams that have not been explored. This might partially be explained by the relatively low number of teams that participated in the challenge. Furthermore, some methods that were common during the last years may simply have proven inferior and were therefore not deployed. Nevertheless, we conclude that the challenge covers a sufficiently large variety of methods to provide a useful performance comparison and to guide method selection and development.

Given the challenging nature of the data set with slides originating from routine clinical workflows, top-performing methods should be considered to have a high performance, both with regards to accuracy, as well as to robustness. Breast cancer cell diameters extend approximately up to 20 μm. The lowest mean TREs across all landmarks of 63.29 μm and 122.21 μm therefore cannot be assumed to allow a cell-level registration, but neighborhoods of cells can be assumed to be registered correctly. Furthermore, depending on the section spacing, actual cell-level correspondence between the sections is impossible to determine. Therefore, the performance level achieved by the best-performing methods may already have reached the limit set by the technical setup using non-consecutive sections. The high section spacing in some image pairs requires registration methods to be robust against image regions that cannot be aligned well. It appears like the top-performing methods have achieved this robustness. However, there is no guarantee that the ranking of algorithms would be the same for a data set consisting of only consecutive or re-stained sections. While there are differences in performance based on the computed metrics, the mean distance reductions in percentage provide an intuition that all top-performing methods are well suited to significantly reduce the initial TRE. The ranking of methods is mostly stable both across metrics and annotator disagreement exclusion thresholds for landmarks. Statistical testing indicates that the ranking based on the 90th percentiles is unlikely to arise by chance. Nevertheless, methods with a lower ranking could be shown to be capable of similar performances through better algorithm optimization in future work. The rankings in this study however provide a clear indication of which methods are currently preferable. A direct comparison of the registration performances to the ANHIR challenge proves difficult, since TREs were normalized with image diagonals in the ANHIR challenge, rather than provided in μm.

Correlations among 90th percentiles of TREs are notably higher among algorithms than between algorithms and corresponding 90th percentiles of DBAs, with the exception of Gestalt Diagnostic. This finding is also supported by the almost identical correlations of the conditional means of the random effect that captures the slide ID. This indicates that image pairs that were difficult to annotate were not necessarily the same as those that were difficult to register for the proposed methods, while generally the same image pairs were difficult to register for the six top-performing methods. The image pair with the worst registration performance based on the 90th percentiles emphasizes the importance of the initial alignment step, which fails in this case due to the precropping of the IHC WSI. This indicates that it would be worthwhile to focus future work in WSI registration on increased robustness against comparable failure cases.

In order to investigate which properties of the WSIs and tissue impact algorithm performances, we conducted LME analysis. The conditional means of the random effects of slide ID and annotator combination indicate that these effects have a higher impact on the TREs than the fixed effects. An analysis of the fixed effects shows that the antibody of the IHC stain within the image pair does not have a significant impact on the TREs, potentially with the exception of a trend towards lower TREs for HER2 WSIs compared to ER WSIs. This might be because the appearances of the routine diagnostic stains in breast cancer are relatively similar. Landmarks in HER2 BS positive IC regions are associated with higher TREs in IC regions, but BS seems to otherwise not impact TREs. It was not possible to investigate the BS in DCIS and LCIS regions, since BS is not routinely reported there, yet cells within these regions can be largely positively stained. Future research should investigate how visually more different stains impact TRE. Nevertheless, this means that the risk for biases due to IHC and BS in multi-stain studies that involve registration might not be high. Regions of DCIS, LCIS, and NMC both were over-annotated by annotators and appear to be associated with lower TREs compared to normal tissue. The reason for this might be the presence of more visually easily distinguishable structures, which could also explain the lower TRE in IC regions. Nevertheless, we find the reduced TRE in IC regions surprising, since IC is characterized through diffuse growth patterns. Potentially, increased nuclear density in IC regions leads to higher contrast reference points, which could be beneficial at lower resolutions. IC is likely to be located in the center of resections, but the negative association with the TRE remained when adjusting for the distance to the center of tissue mass. Higher NHGs are associated with increases of TRE and DBA in IC regions, which is likely due to increasingly poor differentiation of structures and cells. This indicates that there is some risk of bias from cancer grades in studies that deploy registration. The distance of landmarks to the center of tissue mass is positively associated with TREs for most methods, indicating room for improvements in deformable registration. The covariate with the highest impact on the TRE across algorithms is the slide age, which can more than double the baseline registration error for the oldest slides for some methods. In registration-based studies that focus on outcomes, the slide age could therefore confound analyses in multiple ways. However, this is only a concern for studies that focus on WSIs from archived tissue sections. During deployment, WSIs would be generated from recently sectioned tissue.

While it was not possible to investigate the section spacing in this study due to lack of recorded information, the LME analysis adjusts for this through the random effect that captures the slide IDs, alongside other slide-specific properties. The conclusions of the LME analysis regarding the fixed effects are therefore likely transferable between section spacings, nevertheless, future research that elucidates the effect of section spacing would be of high interest. Interestingly, the interquartile range of the conditional means of the random effect that captures the annotator combination are highest for Gestalt Diagnostics. This observation is in concordance with the lowest slope in the reduction of the median 90th percentile of TRE in Supplementary Figure 2a and the significant improvement between validation and test set performance for Gestalt Diagnostics, considering that it was possible to exclude low quality landmarks in the test data. This indicates that the performance estimation for Gestalt Diagnostics could be significantly limited through uncertainty in the annotations. While this also impacts the other teams, the relative impact on their performance metrics might not be as relevant and only the ranking between Fraunhofer MEVIS and AGHSSO could change depending on the DBA exclusion threshold. While the precision of landmarks might be a limiting factor for the quantification of algorithm performances, the LME analysis indicates that the DBA is only weakly associated with the covariates, which means that the TRE quantification is likely not biased within covariate categories.

Besides the precision of landmark placing, this study has several further limitations. It is important to consider that landmarks only allow the quantification of registration performances in sparsely sampled locations. While we hope that the performance in these points is a reliable proxy for the registration performance in all image regions, this is not guaranteed. Furthermore, there is no guarantee that the target landmarks identified by both annotators, even if in close proximity, actually correspond to the source landmark. This focus on landmarks might favor feature-based registration methods that rely on aligning key-points in

the image pairs. There is currently no consensus on how to best evaluate registration performance outside of landmark locations. Annotators over-annotated the classes DCIS, LCIS, NMC and LI and under-annotated normal tissue and artifacts. While this biases the evaluation of performance metrics towards the performance within the over-annotated classes, the correct registration of these tissue classes compared to normal tissue is likely of higher interest for future applications. We therefore think that this does not adversely impact the generalizability of quantitative results. Our evaluation is purely quantitative, no qualitative analysis was performed. Therefore, there is no guarantee that transformed images are feasible for possible downstream analyses following registration. The LME analysis is limited by the assumptions of additive relationships between covariates and a linear relationship between the covariates and the log10-transformed TRE. Furthermore, the high correlation between the WSI scanner, slide age and presence of control tissue does not allow to conclusively disentangle these effects. Another limitation of this challenge is the evaluation of registration algorithms as a whole. It could yield valuable insights to analyze different combinations of suggested pre-processing, initial alignment and deformable registration methods. Furthermore, we did not assess computational performances, since registration took place on the participants computing infrastructures. All tissue materials in this challenge originate from breast resections. Stains include three nuclear stains and one membrane stain. It is therefore not guaranteed that results are transferable to WSIs of tissue from other organs or other stains. However, since registration is typically performed at low resolutions where different tissues might not be distinguishable, we think that it is likely that the results would generalize to other tissues. With regards to the transferability of results between stains, the same considerations apply, although the generalizability might be lower since different stains can have a larger impact on tissue appearance at the macroscopic level.

Despite its limitations, we think that the ACROBAT 2022 challenge has elucidated the state of multi-stain WSI registration algorithms and their application to real-world data that originates from routine clinical workflows. While WSI registration is not yet a solved problem, the results in this study indicate that it has now become a sufficiently reliable technology to enable novel areas of research. Clinical applications are likely also possible with the observed registration performances, but would require further validation to ensure patient safety. Furthermore, this study has led to novel insights into specific strengths and weaknesses of current WSI registration methods and the mixed effects models analysis could be a model for future analyses of registration methods also outside of computational pathology. Five of the discussed methods are available under open-source licenses, and we believe that this study has generated sufficient evidence of algorithm performance and robustness to warrant the proliferation of top-performing methods into a wide range of future applications. The ACROBAT 2023 challenge will build on the results of the ACROBAT 2022 challenge and investigate computational performances and algorithm performance under domain shifts.

CRediT authorship contribution statement

Philippe Weitz: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Masi Valkonen:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Leslie Solorzano:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Data curation. **Circe Carr:** Data curation. **Kimmo Kartasalo:** Writing – review & editing, Data curation. **Constance Boissin:** Writing – review & editing, Data curation. **Sonja Koivukoski:** Data curation. **Aino Kuusela:** Data curation. **Dusan Rasic:** Data curation. **Yanbo Feng:** Data curation. **Sandra Sinius Pouplier:** Data curation. **Abhinav Sharma:** Data curation. **Kajsa Ledesma Eriksson:** Data curation. **Stephanie Robertson:** Writing – review & editing, Data curation. **Christian Marzahl:** Software, Methodology. **Chandler D. Gatenbee:** Software, Methodology. **Alexander R.A. Anderson:** Software, Methodology. **Marek Wodzinski:** Software, Methodology. **Artur Jurgas:** Software, Methodology. **Niccolò Marini:** Software, Methodology. **Manfredo Atzori:** Software, Methodology. **Henning Müller:** Software, Methodology. **Daniel Budelmann:** Software, Methodology. **Nick Weiss:** Software, Methodology. **Stefan Heldmann:** Software, Methodology. **Johannes Lotz:** Software, Methodology. **Jelmer M. Wolterink:** Software, Methodology. **Bruno De Santi:** Software, Methodology. **Abhijeet Patil:** Software, Methodology. **Amit Sethi:** Software, Methodology. **Satoshi Kondo:** Software, Methodology. **Satoshi Kasai:** Software, Methodology. **Kousuke Hirasawa:** Software, Methodology. **Mahtab Farrokh:** Software, Methodology. **Neeraj Kumar:** Software, Methodology. **Russell Greiner:** Software, Methodology. **Leena Latonen:** Funding acquisition. **Anne-Vibeke Laenkholm:** Funding acquisition. **Johan Hartman:** Funding acquisition. **Pekka Ruusuvuori:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Mattias Rantalainen:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Philippe Weitz reports a relationship with Stratipath AB that includes: employment. Mattias Rantalainen reports a relationship with Stratipath AB that includes: equity or stocks. Johan Hartman reports a relationship with Stratipath AB that includes: equity or stocks. Kimmo Kartasalo reports a relationship with Clinsight AB that includes: equity or stocks. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The ACROBAT data set is available at https://snd.gu.se/en/catalogue/study/2022-190.

Code Availability

Code used for displaying landmarks in a surrounding tissue region, code for computing registration performance metrics, as well as the annotator protocols are available from github.com/rantalainenGroup/ACROBAT.

Inclusion & Ethics

The study in whose terms the WSI data was generated has approval by the regional ethics review board (Etiksprövningsmyndigheten, Stockholm, Sweden, ref. 2017/2106–31 and amendments 2018/1462–32, 2019–02336). Due to the retrospective nature of the study, consent was not required.

Acknowledgements

We acknowledge support from Stratipath and Karolinska Institutet sponsoring the ACROBAT challenge prizes; MICCAI society for hosting the ACROBAT challenge, and Nguyen Thuy Duong Tran for support with digitizing histopathology slides.

We acknowledge funding from:

Vetenskapsrådet (Swedish Research Council)

Cancerfonden (Swedish Cancer Society)

ERA PerMed (ERAPERMED2019–224-ABCAP)

MedTechLabs

Swedish e-science Research Centre (SeRC)
VINNOVA
SweLife
Academy of Finland (#341967, #334782, #335976, #334774)
Cancer Foundation Finland
University of Turku Graduate School
Turku University Foundation
Oskar Huttunen Foundation
David and Astrid Hägelén Foundation
Orion Research Foundation
KI Research Foundation

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 945358. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA (www.imi.europe.eu).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2024.103257.

References

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imag. https://doi.org/10.1109/TMI.2019.2897538.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw., Articles 67, 1–48.

Baxi, V., Edwards, R., Montalto, M., Saha, S., 2022. Digital pathology and artificial intelligence in translational medicine and clinical practice. Mod. Pathol. 35, 23–32.

Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., Madabhushi, A., 2019. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat. Rev. Clin. Oncol. 16, 703–715.

Borovec, J., Kybic, J., Arganda-Carreras, I., Sorokin, D.V., Bueno, G., Khvostikov, A.V., Bakas, S., Chang, E.I.-C., Heldmann, S., Kartasalo, K., Latonen, L., Lotz, J., Noga, M., Pati, S., Punithakumar, K., Ruusuvuori, P., Skalski, A., Tahmasebi, N., Valkonen, M., Venet, L., Wang, Y., Weiss, N., Wodzinski, M., Xiang, Y., Xu, Y., Yan, Y., Yushkevich, P., Zhao, S., Munoz-Barrutia, A., 2020. ANHIR: automatic non-rigid histological image registration challenge. IEEe Trans. Med. ImAging 39, 3042–3052.

Bulten, W., Kartasalo, K., Chen, P.-H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C., van der Laak, J., Amin, M.B., Evans, A.J., van der Kwast, T., Allan, R., Humphrey, P.A., Grönberg, H., Samaratunga, H., Delahunt, B., Tsuzuki, T., Häkkinen, T., Egevad, L., Demkin, M., Dane, S., Tan, F., Valkonen, M., Corrado, G.S., Peng, L., Mermel, C.H., Ruusuvuori, P., Litjens, G., Eklund, M., PANDA challenge consortium, 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat. Med. 28, 154–163.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G., 2020. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 21, 233–241.

Burlingame, E.A., McDonnell, M., Schau, G.F., Thibault, G., Lanciault, C., Morgan, T., Johnson, B.E., Corless, C., Gray, J.W., Chang, Y.H., 2020. SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning. Sci. Rep. 10, 17507.

Couture, H.D., Williams, L.A., Geradts, J., Nyante, S.J., Butler, E.N., Marron, J.S., Perou, C.M., Troester, M.A., Niethammer, M., 2018. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ. Breast. Cancer 4, 30.

de Haan, K., Zhang, Y., Zuckerman, J.E., Liu, T., Sisk, A.E., Diaz, M.F.P., Jen, K.-Y., Nobori, A., Liou, S., Zhang, S., Riahi, R., Rivenson, Y., Wallace, W.D., Ozcan, A., 2021. Deep learning-based transformation of H&E stained tissues into special stains. Nat. Commun. 12, 4884.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2017. SuperPoint: self-supervised interest point detection and description. arXiv [cs.CV].

Duanmu, H., Bhattarai, S., Li, H., Shi, Z., Wang, F., Teodoro, G., Gogineni, K., Subhedar, P., Kiraz, U., Janssen, E.A.M., Aneja, R., Kong, J., 2022. A spatial attention guided deep learning system for prediction of pathological complete response using breast cancer histopathology images. Bioinformatics. 38, 4605–4612.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 381–395.

Foersch, S., Glasner, C., Woerl, A.-C., Eckstein, M., Wagner, D.-C., Schulz, S., Kellers, F., Fernandez, A., Tserea, K., Kloth, M., Hartmann, A., Heintz, A., Weichert, W., Roth, W., Geppert, C., Kather, J.N., Jesinghaus, M., 2023. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. Nat. Med. 29, 430–439.

Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., Gerstung, M., 2020. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat. Cancer 1, 800–810.

Haber, E., Modersitzki, J., 2007. Intensity gradient based registration and fusion of multi-modal images. Methods Inf. Med. 46, 292–299.

Huang, Z., Shao, W., Han, Z., Alkashash, A.M., De la Sancha, C., Parwani, A.V., Nitta, H., Hou, Y., Wang, T., Salama, P., Rizkalla, M., Zhang, J., Huang, K., Li, Z., 2023. Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. NPJ. Precis. Oncol. 7, 14.

Kartasalo, K., Latonen, L., Vihinen, J., Visakorpi, T., Nykter, M., Ruusuvuori, P., 2018. Comparative analysis of tissue reconstruction algorithms for 3D histology. Bioinformatics. 34, 3013–3021.

Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A.J., Bankhead, P., Kooreman, L.F.S., Schulte, J.J., Cipriani, N.A., Buelow, R.D., Boor, P., Ortiz-Brüchle, N., Hanby, A.M., Speirs, V., Kochanny, S., Patnaik, A., Srisuwananukorn, A., Brenner, H., Hoffmeister, M., van den Brandt, P.A., Jäger, D., Trautwein, C., Pearson, A.T., Luedde, T., 2020. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat. Cancer 1, 789–799.

Khan, U., Koivukoski, S., Valkonen, M., Latonen, L., Ruusuvuori, P., 2023. The effect of neural network architecture on virtual H&E staining: systematic assessment of histological feasibility. PATTER 0. doi:10.1016/j.patter.2023.100725.

Leutenegger, S., Chli, M., Siegwart, R.Y., 2011. BRISK: binary Robust invariant scalable keypoints. In: 2011 International Conference on Computer Vision. Presented at the 2011 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 2548–2555.

Lin, J.-R., Wang, S., Coy, S., Chen, Y.-A., Yapp, C., Tyler, M., Nariya, M.K., Heiser, C.N., Lau, K.S., Santagata, S., Sorger, P.K., 2023. Multiplexed 3D atlas of state transitions and immune interaction in colorectal cancer. CellCell 186, 363–381. .e19.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110.

Rantalainen, M., Hartman, J., 2023. ACROBAT - a multi-Stain Breast Cancer Histological Whole-Slide-Image Data Set from Routine Diagnostics For Computational Pathology. Swedish National Data Service (SND). https://doi.org/10.48723/w728-p041.

Sarlin, P.E., DeTone, D., Malisiewicz, T., 2020. Superglue: learning feature matching with graph neural networks. In: Proceedings of the.

Schapiro, D., Sokolov, A., Yapp, C., Chen, Y.-A., Muhlich, J.L., Hess, J., Creason, A.L., Nirmal, A.J., Baker, G.J., Nariya, M.K., Lin, J.-R., Maliga, Z., Jacobson, C.A., Hodgman, M.W., Ruokonen, J., Farhi, S.L., Abbondanza, D., McKinley, E.T., Persson, D., Betts, C., Sivagnanam, S., Regev, A., Goecks, J., Coffey, R.J., Coussens, L. M., Santagata, S., Sorger, P.K., 2022. MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. Nat. Methods 19, 311–315.

Schmauch, B., Romagnoni, A., Pronier, E., Saillard, C., Maillé, P., Calderaro, J., Kamoun, A., Sefta, M., Toldo, S., Zaslavskiy, M., Clozel, T., Moarii, M., Courtiol, P., Wainrib, G., 2020. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat. Commun. 11, 3877.

Solorzano, L., Partel, G., Wählby, C., 2020. TissUUmaps: interactive visualization of large-scale spatial gene expression and tissue morphology data. Bioinformatics. 36, 4363–4365.

Song, Y., Treanor, D., Bulpitt, A.J., Magee, D.R., 2013. 3D reconstruction of multiple stained histology images. J. Pathol. Inform. 4, S7.

Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D.M., Bostwick, D.G., Evans, A.J., Grignon, D.J., Humphrey, P.A., Iczkowski, K.A., Kench, J.G., Kristiansen, G., van der Kwast, T.H., Leite, K.R.M., McKenney, J.K., Oxley, J., Pan, C.-C., Samaratunga, H., Srigley, J.R., Takahashi, H., Tsuzuki, T., Varma, M., Zhou, M., Lindberg, J., Lindskog, C., Ruusuvuori, P., Wählby, C., Grönberg, H., Rantalainen, M., Egevad, L., Eklund, M., 2020. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. 21, 222–232.

Su, A., Lee, H., Tan, X., Suarez, C.J., Andor, N., Nguyen, Q., Ji, H.P., 2022. A deep learning model for molecular label transfer that enables cancer cell identification from histopathology images. NPJ. Precis. Oncol. 6, 14.

Turkki, R., Linder, N., Kovanen, P.E., Pellinen, T., Lundin, J., 2016. Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. J. Pathol. Inform. 7, 38.

Valkonen, M., Isola, J., Ylinen, O., Muhonen, V., Saxlin, A., Tolonen, T., Nykter, M., Ruusuvuori, P., 2020. Cytokeratin-supervised deep learning for automatic recognition of epithelial cells in breast cancers stained for ER, PR, and Ki-67. IEEe Trans. Med. ImAging 39, 534–542.

Wang, Y., Acs, B., Robertson, S., Liu, B., Solorzano, L., Wählby, C., Hartman, J., Rantalainen, M., 2022. Improved breast cancer histological grading using deep learning. Ann. Oncol. 33, 89–98.

Wang, Y., Kartasalo, K., Weitz, P., Ács, B., Valkonen, M., Larsson, C., Ruusuvuori, P., Hartman, J., Rantalainen, M., 2021. Predicting molecular phenotypes from histopathology images: a transcriptome-wide expression-morphology analysis in breast cancer. Cancer Res. 81, 5115–5126.

Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C., 2013. DeepFlow: large displacement optical flow with deep matching. In: 2013 IEEE International Conference on Computer Vision. Presented at the 2013 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 1385–1392.

Weitz, P., Sartor, V., Acs, B., Robertson, S., Budelmann, D., Hartman, J., Rantalainen, M., 2023a. Increasing the usefulness of already existing annotations through WSI registration. arXiv [cs.CV].

Weitz, P., Valkonen, M., Solorzano, L., Carr, C., Kartasalo, K., Boissin, C., Koivukoski, S., Kuusela, A., Rasic, D., Feng, Y., Pouplier, S.K.S., Sharma, A., Eriksson, K.L., Latonen, L., Laenkholm, A.-V., Hartman, J., Ruusuvuori, P., Rantalainen, M., 2022a. ACROBAT – a multi-stain breast cancer histological whole-slide-image data set from routine diagnostics for computational pathology. arXiv [eess.IV].

Weitz, P., Valkonen, M., Solorzano, L., Carr, C., Kartasalo, K., Boissin, C., Koivukoski, S., Kuusela, A., Rasic, D., Feng, Y., Sinius Pouplier, S., Sharma, A., Ledesma Eriksson, K., Latonen, L., Laenkholm, A.-V., Hartman, J., Ruusuvuori, P., Rantalainen, M., 2023b. A multi-stain breast cancer histological whole-slide-image data set from routine diagnostics. Sci. Data 10, 562.

Weitz, P., Wang, Y., Kartasalo, K., Egevad, L., Lindberg, J., Grönberg, H., Eklund, M., Rantalainen, M., 2022b. Transcriptome-wide prediction of prostate cancer gene expression from histopathology images using co-expression-based convolutional neural networks. Bioinformatics. 38, 3462–3469.

Wieslander, H., Gupta, A., Bergman, E., Hallström, E., Harrison, P.J., 2021. Learning to see colours: biologically relevant virtual staining for adipocyte cell images. PLoS ONE 16, e0258546.

06–08 Jul Wolterink, J.M., Zwienenberg, J.C., Brune, C., 2022. Implicit neural representations for deformable image registration. In: Konukoglu, E., Menze, B., Venkataraman, A., Baumgartner, C., Dou, Q., Albarqouni, S. (Eds.), Proceedings of The 5th International Conference on Medical Imaging with Deep Learning, Proceedings of Machine Learning Research. Presented at the Medical Imaging with Deep Learning, PMLR, pp. 1349–1359.