

Generalizable and Data-efficient Deep Learning for Computer-aided Cancer Diagnosis

Sophia Janine Wagner

Complete reprint of the dissertation approved by the TUM School of Computation, Information and Technology of the Technical University of Munich for the award of the

Doktorin der Naturwissenschaften (Dr. rer. nat.).

Chair: Prof. Dr. Nassir Navab

Examiners:

1. TUM Junior Fellow Dr. Tingying Peng
2. Ass. Prof. Dr. Lequan Yu
3. Prof. Dr. Peter Schüffler

The dissertation was submitted to the Technical University of Munich on 20 December 2024 and accepted by the TUM School of Computation, Information and Technology on 2 May 2025.

Abstract

Recent advances in computational pathology have positioned deep learning as a powerful tool to interpret digitized histology slides reaching human-level performance on well-defined tasks. By extracting representations directly from hematoxylin and eosin-stained tissue samples, methods aim to predict clinically relevant endpoints, such as genetic alterations, tumor grading, or patient prognosis. However, developing generalizable, data-efficient, and clinically relevant models for cancer diagnosis remains challenging for various reasons.

Deep learning models show low *generalization* performance across diverse clinical settings. Variations in tissue preparation, staining protocols, and scanner conditions create strong batch effects, causing models to perform poorly on unseen data. To address this, we propose generative style transfer methods for multi-domain stain augmentation in histopathology and microscopic imaging. A second challenge is *representation learning for whole-slide images*, which must distill gigapixel whole-slide images into compact embeddings. To this end, we propose a new training routine combining different training paradigms for patch encoding and a novel solely Transformer-based *feature aggregation network*. The Transformer-based aggregation model incorporates tissue context to predict multiple genetic alterations and demonstrates strong generalization capabilities for biomarker prediction in colorectal cancer. Notably, the resection-trained model predicts biomarkers from colorectal biopsies for the first time, advancing clinical utility. Furthermore, we show that low-resource fine-tuning of vision foundation models, originally trained on natural images, can outperform specialized histopathology encoders. Using this insight, we introduce a hematology foundation model that sets new benchmarks for classifying white blood cells and weakly supervised leukemia subtyping. Finally, the field of computational pathology suffers from low *reproducibility*, often due to inaccessible code, limited data availability, and inconsistent evaluation practices. We systematically assess the reproducibility and propose recommendations to enhance code release, data sharing, and statistical analysis.

In summary, this thesis’s objective is to develop clinically relevant deep learning models for cancer diagnosis to bring models a step closer to routine clinical workflows by improving generalization through stain augmentation, advancing representation learning with data-efficient methods, developing feature aggregation for clinical endpoint prediction, and raising reproducibility standards.

Zusammenfassung

Jüngste Fortschritte in der computergestützten Pathologie haben Deep Learning als ein leistungsstarkes Werkzeug zur Interpretation digitalisierter Histologieschnitte etabliert, wobei für klar definierte Aufgaben bereits eine dem Menschen vergleichbare Leistung erreicht werden kann. Durch das direkte Extrahieren von Repräsentationen aus Hämatoxylin- und Eosin-gefärbtem Gewebe zielen Methoden darauf ab, klinisch relevante Endpunkte, wie Mutationen, Tumorgrading oder Patientenprognosen vorherzusagen. Dennoch bleibt die Entwicklung generalisierbarer, daten-effizienter und klinisch relevanter Modelle für die Krebsdiagnostik aus verschiedenen Gründen eine Herausforderung.

Deep-Learning-Modelle weisen eine geringe Generalisierungsfähigkeit über verschiedene klinische Umgebungen hinweg auf. Unterschiede in der Gewebeaufbereitung, in Färbeprotokollen und bei den Scannereinstellungen führen zu deutlichen Batch-Effekten, sodass Modelle bei bisher ungesehenen Daten schlechter abschneiden. Um dieses Problem anzugehen, schlagen wir generative Methoden für Multi-Domain-Färbeaugmentierungen in der Histopathologie und der mikroskopischen Bildgebung vor.

Eine zweite Herausforderung ist das Repräsentationslernen für digitalisierte Histologiebilder, bei dem gigapixelgroße Bilder in kompakte Repräsentation überführt werden müssen. Zu diesem Zweck stellen wir eine neue Trainingsroutine vor, die verschiedene Trainingsparadigmen für die Patch-Repräsentationen kombiniert, sowie ein neuartiges reines Transformer-basiertes Feature-Aggregationsnetzwerk. Das auf Transformern basierende Aggregationsmodell integriert Gewebekontext, um mehrere genetische Mutationen vorherzusagen, und zeigt starke Generalisierungsfähigkeit bei der Biomarkervorhersage für Darmkrebs. Bemerkenswerterweise funktioniert das Modell erstmals erfolgreich Biomarker aus Kolonoskopen und steigert so den klinischen Nutzen. Darüber hinaus zeigen wir, dass ein ressourcen-effizientes Fine-Tuning von visuell vortrainierten Deep Learning Modellen (Vision Foundation Models), die ursprünglich auf natürlichen Bildern trainiert wurden, spezialisierte Histopathologie-Encoder übertrffen kann. Aufbauend auf diesen Erkenntnissen führen wir ein Hämatologie-Modell ein, das neue Maßstäbe bei der Klassifizierung von weißen Blutkörperchen und bei Leukämie-Klassifizierung setzt.

Abschließend leidet das Feld der computergestützten Pathologie unter einer geringen

Reproduzierbarkeit, häufig bedingt durch unzugänglichen Quellcode, begrenzte Datenverfügbarkeit und uneinheitliche Evaluationspraktiken. Wir bewerten die Reproduzierbarkeit systematisch und geben Empfehlungen, um Codefreigabe, Datenaustausch sowie statistische Analysen zu verbessern.

Zusammenfassend zielt diese Dissertation darauf ab, klinisch relevante Deep-Learning-Modelle für die Krebsdiagnostik zu entwickeln, um die Modelle einen Schritt näher an den klinischen Routineeinsatz heranzuführen. Dies geschieht durch die Verbesserung der Generalisierung mittels Färbeaugmentierung, den Fortschritt im Repräsentationlernen mittels daten-effizienter Methoden, die Entwicklung von Feature-Aggregation zur Vorhersage klinischer Endpunkte sowie die Anhebung von Reproduzierbarkeitsstandards.

List of Publications

This thesis is based on the following peer-reviewed publications, presented in the order of appearance in this thesis. The core publications of this thesis are marked with an asterisk () and the share of first authorship (1/n for n first authors).*

- P1. (*, 1) Wagner, S. J., Khalili, N., Sharma, R., Boxberg, M., Marr, C., De Back, W., & Peng, T. (2021). **Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations.** In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII* 24 (pp. 257-266). Springer International Publishing. (Section 3.1)
- P2. Liu, Y., Wagner, S. J., & Peng, T. (2022). **Multi-modality microscopy image style augmentation for nuclei segmentation.** *Journal of Imaging*, 8(3), 71. (Section 3.2)
- P3. Tran, M., Wagner, S. J., Boxberg, M., & Peng, T. (2022, September). **S5CL: unifying fully-supervised, self-supervised, and semi-supervised learning through hierarchical contrastive learning.** In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 99-108). Cham: Springer Nature Switzerland. (Section 3.3)
- P4. (*, 0.33) Roth, B.*, Koch, V.*, Wagner, S. J.*, Schnabel, J. A., Marr, C., & Peng, T. (2024). **Low-resource finetuning of foundation models beats state-of-the-art in histo-pathology.** *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. *Contributed equally (Section 3.4)
- P5. (*, 0.5) Koch, V.*, Wagner, S. J.*, Kazemini, S., Sancar, E., Hehr, M., Schnabel, J. A., ... & Marr, C. (2024, October). **DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology.** In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 520-530). Cham: Springer Nature Switzerland. *Contributed equally (Section 3.5)
- P6. Reisenbuechler, D., Wagner, S. J., Boxberg, M., & Peng, T. (2022, September). **Local attention graph-based transformer for multi-target genetic alteration prediction.** In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 377-386). Cham: Springer Nature Switzerland. (Section 3.6)

- P7. (*, 1) Wagner, S. J., Reisenbüchler, D., West, N. P., Niehues, J. M., Zhu, J., Foersch, S., ... & Kather, J. N. (2023). **Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study.** *Cancer Cell*, 41(9), 1650-1661. (Section 3.7)
- P8. (*, 0.5) Wagner, S. J.*, Matek, C.*, Boushehri, S. S., Boxberg, M., Lamm, L., Sadafi, A., ... & Peng, T. (2024). **Built to last? Reproducibility and reusability of deep learning algorithms in computational pathology.** *Modern Pathology*, 37(1), 100350.
*Contributed equally (Section 3.8)
- P9. Wagner, S. J.*, Matek, C.*, Shetab Boushehri, S., Boxberg, M., Lamm, L., Sadafi, A., ... & Peng, T. (2022). **Make deep learning algorithms in computational pathology more reproducible and reusable.** *Nature Medicine*, 28(9), 1744-1746.
*Contributed equally (Section 3.9)

Contents

Abstract	ii
Zusammenfassung	iii
List of Publications	v
1. Introduction	1
1.1. Use cases of deep learning in pathology	3
1.2. Deep learning for clinical endpoint prediction	4
1.3. Challenges on the path to the clinic and thesis contributions	6
2. Methods	8
2.1. Pre-processing and stain color adaptation	9
2.1.1. Color-matching methods	10
2.1.2. Generative adversarial networks	11
2.2. Representation learning for encoding histology patches	12
2.2.1. Proxy tasks for representation learning	13
2.2.2. Contrastive learning for self-supervision	14
2.2.3. Self-distillation for self-supervision	16
2.2.4. Multimodal alignment for self-supervision	17
2.2.5. Self-supervised learning for patch encoder in histopathology . .	18
2.3. Weakly-supervised learning for histopathology	20
2.3.1. Attention-based networks	21
2.3.2. Graph neural networks and multi-scale networks	22
2.3.3. Transformer-based networks	23
2.3.4. Representation learning on slide-level	24
3. Publications	26
3.1. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations	27
3.2. Multi-modality microscopy image style augmentation for nuclei segmentation	28

3.3. S5CL: unifying fully-supervised, self-supervised, and semi-supervised learning through hierarchical contrastive learning	29
3.4. Low-resource finetuning of foundation models beats state-of-the-art in histopathology	30
3.5. DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology	31
3.6. Local attention graph-based transformer for multi-target genetic alteration prediction	32
3.7. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study	33
3.8. Built to last? Reproducibility and reusability of deep learning algorithms in computational pathology	34
3.9. Make deep learning algorithms in computational pathology more reproducible and reusable	35
4. Discussion	36
List of Figures	39
List of Tables	40
Bibliography	41
A. Original Publications	59
A.1. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations	59
A.2. Multi-modality microscopy image style augmentation for nuclei segmentation	77
A.3. S5CL: unifying fully-supervised, self-supervised, and semi-supervised learning through hierarchical contrastive learning	96
A.4. Low-resource finetuning of foundation models beats state-of-the-art in histopathology	114
A.5. DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology	121
A.6. Local attention graph-based transformer for multi-target genetic alteration prediction	140
A.7. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study	158
A.8. Built to last? Reproducibility and reusability of deep learning algorithms in computational pathology	177

A.9. Make deep learning algorithms in computational pathology more reproducible and reusable	190
--	-----

1. Introduction

Cancer diagnosis and therapy have seen rapid progress in the past decades. Today, a state-of-the-art diagnosis involves the determination of a tumor’s phenotype via histopathological examination and in certain cases requires a tumor’s genotype via genetic profiling [72, 94, 56]. Histopathology describes the microscopic examination of tissue to detect and diagnose diseases and has been part of the diagnosis routine for over a century [137]. Recently, the advances in genetic profiling, which has become more cost-effective and widely available, and the subsequent new developments of molecular biomarkers for stratifying patient prognosis have cumulated in the foundation of the field of precision oncology, facilitating targeted therapies tailored to individual patients. Despite ongoing discussions over the importance of the phenotype versus the genotype [115], microscopic examination remains the gold standard of anatomic pathology as articulated by Rosai in 2007:

“The discussion of the role that microscopy plays and in all likelihood will continue to play in ‘the molecular age’ of medicine can be divided into two separate categories: diagnostic pathology and investigative pathology. Regarding the former, it is my opinion that there is no available technique that provides so much information so abundantly, so quickly and as inexpensively as conventional microscopy, in the form colloquially known as the H&E technique.” [124].

Hence, despite all advances, hematoxylin and eosin (H&E)-staining remains the baseline of current pathological workflows. From an information theory perspective, the field of pathology aims to find the most effective and at the same time most efficient representation of a disease and its related factors, which currently still is the microscopic examination of tissue samples. Similarly, computational pathology aims to find suitable representations of a patient’s disease from the available patient data. The emerging research question then becomes: **How can we find generalizable representations of disease or other aspects of interest in the pathological workflow and how can we integrate these effectively in the clinical setting?** In this thesis, I work towards finding generalizable representations in data-efficient settings to solve clinically relevant tasks such as diagnostic biomarker prediction from digitized whole slide images (WSIs).

Technical progress has been driving digitization in histopathology in the past decade. Coupled with advances deep learning (DL) more specifically, the field of computational

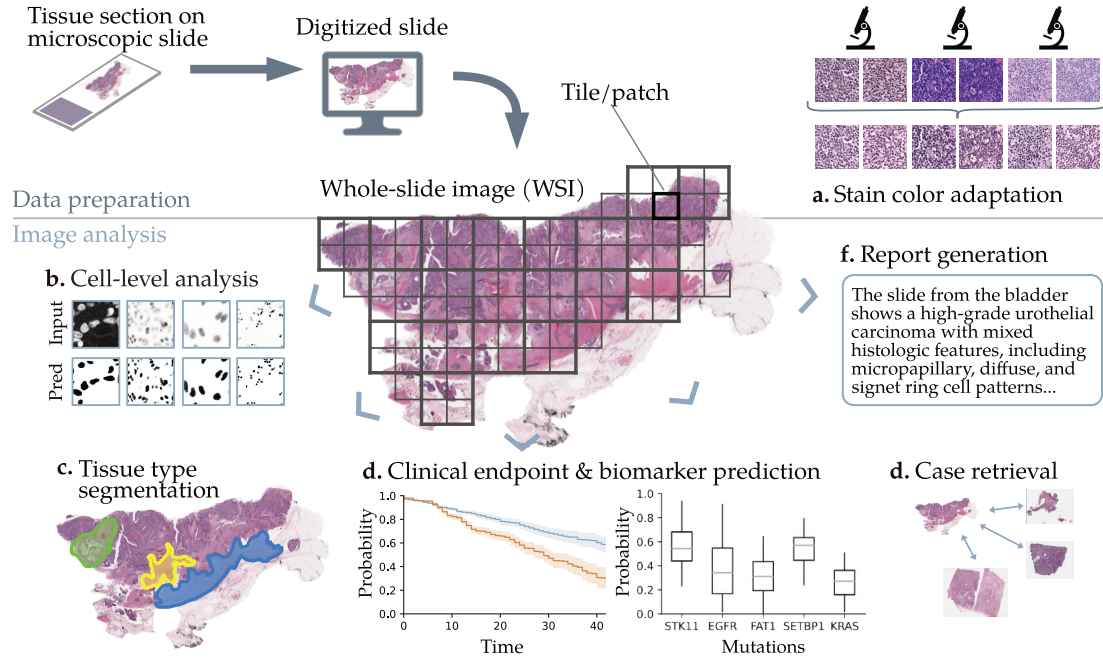


Figure 1.1.: **Use cases of deep learning in pathology.** After digitization of the whole slide image, computational methods can be divided into data pre-processing and image analysis methods. Among these (a) stain normalization, (b) evaluation of cell-level features, (c) tissue type segmentation, (d) clinical endpoint and biomarker prediction, (e) case retrieval, and (f) report generation are common tasks in the downstream workflow of computational pathology.

pathology has experienced an enormous increase in relevance. While computational means have always aimed to support pathologist, most directly in their diagnostic workflow [57], they can also be applied for therapy and drug development [141], such as for biomarker discovery. In particular, deep neural networks have shown early promise in reaching performance levels of medical experts on well-defined tasks such as skin cancer diagnosis [37], lung cancer subtype classification [28], or the recognition of malignant white blood cells [104]. Given the increasing number of cancer cases every year leading to an increasing caseload per pathology combined with a decline in the number of practicing pathologists [102, 107], computational pathology bears a huge potential to improve the clinical workflow by accelerating and aiding decision-making of pathologists.

1.1. Use cases of deep learning in pathology

Applications of artificial intelligence (AI) in pathology can be broadly divided along the clinical workflow into data preparation of the digitized WSI and subsequent image analysis tasks (Figure 1.1). To prepare tissue for microscopic examination, sections of formalin-fixed and paraffin-embedded (FFPE) tissues are cut into thin layers (3-5mm) using a microtome and placed onto a glass slide. The tissue is then stained to make the cellular structures visible. Routinely, the tissue slides are stained with H&E, where hematoxylin stains nuclei in purple-blue and eosin stains cytoplasm and extracellular matrix pink. Further stainings can highlight specific proteins, such as immunohistochemistry (IHC) stainings, or other cellular components. The stained tissue slides are then digitized using scanners that use objective powers with high magnifications, typically of $20\times$ or $40\times$, corresponding to resolutions of 0.5 to $0.25\mu\text{m}/\text{pixel}$, respectively. The resulting digitized WSIs can then have a size of up to $100,000 \times 200,000 \approx 20$ Giga pixels posing a challenge for computational models. For perspective, the ImageNet dataset [29], which enabled advances in DL, contains 1,281,167 images—roughly equivalent to the data in 128 WSIs at $20\times$ magnification assuming every WSI contains on average 10^4 images.

The resulting digitized WSIs can differ in appearance due to variations in tissue preparation, staining protocols, and scanning properties [153, 14]. These non-biological variations, commonly referred to as *batch effects*, can affect the downstream performance of computational models leading to poor generalization to external data cohorts, e.g., from other clinics [135]. To counteract this, computational methods adapt image patches to either have a uniform appearance or diversify the dataset to make the downstream model robust to stain variances (Figure 1.1a).

After data preparation, WSIs can be analyzed to extract clinically relevant information (Figure 1.1). The **evaluation of cell-level features** such as cell density, cell count [129], and nuclei segmentation [P2.] of particular cell types is a critical task in diagnostics, e.g., counting mitotic figures, where the number of dividing tumor cells is a prognostic factor for grading breast cancer [142]), or tumor-infiltrating lymphocytes (TILs), which are predictive for multiple solid tumors [61] (Figure 1.1b). Automated approaches can improve efficiency and reliability over manual counting when applied on the entire WSI [105, 24].

Understanding the composition and architecture of tissue types, such as epithelial, stroma, or tumor tissue, is integral to diagnosis based on histopathological images. Automatic **segmentation of tissue types** can subsequently be used to determine measurements, such as the tumor-stroma-ratio, a prognostic factor in breast cancer [89], colorectal cancer [45, 106], and lung cancer [155] (Figure 1.1c). Methods require either pixel-level semantic segmentation, patch-level tissue types to segment at WSI-level,

or WSI-level approaches where interpretability methods can be used to derive tissue segmentations [99].

Predicting **clinical endpoints**, such as grading, staging, or survival, or **biomarkers**, such as genetic alterations [43, 25, 126] or protein expression [12], is an important task in computational pathology, posing a challenge for computational approaches due to the large mismatch in input and output dimensionality, that is mapping from gigapixel WSIs as inputs to single numbers as outputs (Figure 1.1d). With the advent of precision oncology, molecular assays and genome sequencing were increasingly introduced in the diagnostic workflow of pathology. Directly **predicting genetic alterations** or specific molecular profiles from H&E-stained histology can serve as a time and cost-efficient alternative to current approaches and has invoked the developments of methods using weakly-supervised learning as detailed in the following section [134].

To further assist in the diagnostic process, large slide archives can support **case retrieval** to find visually similar cases for reference [148, 18, 127] (Figure 1.1d). This aids pathologists in recognizing rare diseases and refining their assessments through comparison and expert consultation of past cases.

Beyond direct prediction of clinical endpoints, AI models can also **generate clinical reports**, describing the microscopic findings [139, 131, 32] or giving advice on further testing to determine the disease [98] (Figure 1.1f). With recent developments in large language models and multi-modal vision-language models, models have been developed for patch and region captioning at several magnifications [97, 68, 70]. These models have also been extended to the slide level, where they can be aligned with clinical reports [158, 131] or synthetic captions and clinical reports [32].

1.2. Deep learning for clinical endpoint prediction

Discovering new biomarkers related to patient prognosis and predicting clinical endpoints of patients directly from H&E-stained images represent key goals for AI-driven histopathology for research and diagnostic purposes [134]. This can be *molecular* information from additional analysis or tests, such as genetic alterations, genetic assays, or molecular biomarkers, *morphological* information solely based on the WSI, such as tumor grading or tumor staging, or *clinical* information such as patient prognosis or related diagnostic measurements such as treatment response. These algorithms can be used for diagnostic workflows, e.g., to detect microsatellite instability (MSI) in colorectal cancer [83, 35, 146], but also for research purposes by developing new AI-based biomarkers or discovering patterns linked to molecular profiles [90].

From a technical perspective, mapping gigapixel WSIs to clinical endpoint prediction is inherently challenging due to the large imbalance of the input and output dimen-

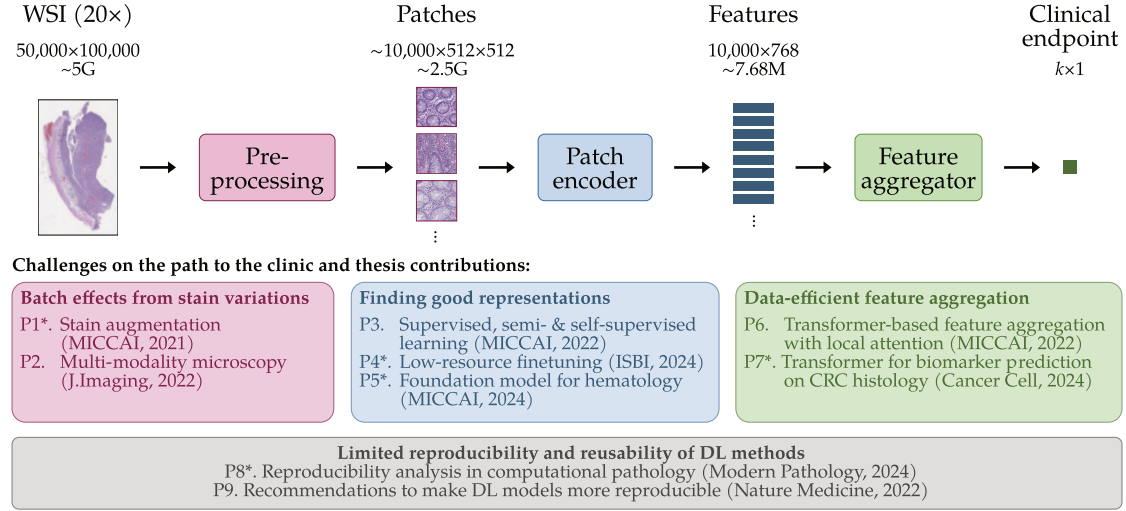


Figure 1.2.: **Computational workflow for clinical endpoint prediction.** Challenges in the computational workflow are detailed including the contributions of this dissertation-based thesis (core publications are marked by *).

sionality. Typically, the workflow is partitioned into three components (Figure 1.2): *pre-processing* of the WSI into smaller image patches, reducing the dimensionality of the image patches to one-dimensional vectors using a *patch encoder*, and aggregating all features per WSI, which can involve more than 10^4 features, with a *feature aggregation* network to a final prediction of a clinical endpoint.

Current pre-processing pipelines are rarely standardized. The typical workflow entails the semantic segmentation of tissue regions, which can be a challenging task when dealing with artifacts, regions with low staining absorption or low staining intensities as often present in IHC stainings, tiling of the tissue regions, typically into patches of size 256×256 or 512×512 , and other pre-processing steps such as stain color adaptation. To reduce batch effects present in histology settings [135], stain normalization methods aim to harmonize the visual appearance of WSIs to a chosen reference image [120, 100, 140]. Recently, color augmentation was proposed to make downstream DL networks robust to stain variances [135, 145, 101, 39].

Next, patch encoders extract the relevant information from histology image patches and encode them in low-dimensional embeddings. Advances in self-supervised learning in computer vision have led to vision foundation models that generalize well to unknown datasets and new tasks [163, 17, 117, 47]. Applying these training recipes on large, private repositories of WSIs, patch-level foundation models for histopathology [97, 20, 158, 144, 40, 125] have been developed and proven to yield versatile general-purpose

representations of histology patches forming a new generation of histopathology models for further downstream tasks [110, 11].

Finally, feature aggregation networks map the low-dimensional patch features to one or a few clinical endpoints. Early approaches such as simple pooling or clustering methods performed this in an unsupervised manner. This approach has been improved by attention-based networks, essentially computing a weighted average [71, 128, 99], or Transformer-based methods [132, 146], which also consider inter-patch relationships.

1.3. Challenges on the path to the clinic and thesis contributions

Developing AI models targeted for clinical applications in medical microscopy imaging faces several challenges. Below, I outline the major challenges in clinical endpoint prediction workflows, present the arising research questions, and discuss how the publications included in this thesis contribute to a solution (core publications are marked by *). When describing specific contributions of the publications, I use “we” to acknowledge the work of my co-authors.

Batch effects lead to poor generalization. Variations in staining protocols and scanner settings can cause strong batch effects leading to poor model generalization to unseen clinical settings even when stain normalization methods are used for testing. *How to build generalizable and robust DL models that perform well on unseen datasets?*

- (*) MICCAI, 2022: We propose the first generative method trained on data from multiple hospitals that can be applied for stain augmentation, able to synthesize a large variety of domain colors. The generative adversarial network disentangles the style from the content and can thereby preserve the morphology [P1].
- J.Imaging, 2022: We develop a style transfer method for microscopy images from diverse modalities. Using this structure-preserving approach, segmentation masks from less frequent microscopy domains can be augmented with synthetic data using segmentation masks from more frequent microscopy domains, leading to better downstream segmentation performance [P2].

Finding suitable representations as patch embeddings. Current patch-based foundation models are often trained on large internal WSI repositories, neglecting powerful general-purpose foundation models trained on natural images and existing annotations. Additionally, they do not generalize well to tasks in hematology. *How to find good representation in data-efficient settings?*

- MICCAI, 2022: We combine supervised, self-supervised, and semi-supervised learning to leverage available annotations and improve representation quality as evaluated on colorectal cancer histology and white blood cell classification [P3].
- (*) ISBI, 2024: We show that low-resource finetuning of foundation models trained on large-scale datasets of natural images outperforms state-of-the-art patch encoders in histopathology trained on millions of histopathology patches [P4].
- (*) MICCAI, 2024: We introduce a foundation model for hematology trained on the largest collection of 13 datasets of white blood cell images from peripheral blood smears and bone marrow smears. The model outperforms existing models on white blood cell classification and leukemia subtyping [P5].

Data-efficient feature aggregation. New feature aggregation models are often evaluated not using general-purpose Transformer-based patch encoders, non-relevant clinical tasks, single datasets, and only resections. Additionally, most methods evaluated on clinical tasks do not contextualize the patch features. *How to build DL models for clinical endpoint prediction and thoroughly evaluate them in diverse clinical settings?*

- MICCAI, 2022: We develop a Transformer-based aggregation network using localized self-attention to better reflect the underlying tissue biology. We evaluate our model on mutation prediction with a multi-task objective for ten mutations on colorectal cancer and assess the generalization capabilities [P6].
- (*) Cancer Cell, 2023: We develop a Transformer-based aggregation network for biomarker prediction on colorectal cancer histology. We assess the model performance and generalization on 16 patient cohorts of over 13,000 patients. The final model is trained on 13 patient cohorts and generalizes from resections to biopsy samples for the first time [P7].

Limited reproducibility. Limited code and data availability, combined with non-standardized evaluation practices lead to low reproducibility and reusability of DL methods in machine learning [4]. *Are the results of models in computational pathology reproducible and—if not—how can we develop reproducible and reusable models?*

- (*) Modern Pathology, 2024: We systematically assess the reproducibility and reusability of DL models in computational pathology, evaluating code availability, data accessibility, and statistical analysis based on a reproducibility checklist for code, data, and statistical analysis [P8].
- Nature Medicine, 2022: Building on our analysis in P8., we propose recommendations to improve reproducibility and reusability in computational pathology [P9].

2. Methods

From a technical perspective, the typical workflow in clinical endpoint prediction from hematoxylin- and eosin (H&E) stained whole slide images (WSIs) involves three main components (Figure 1.2): (i) *pre-processing* of the WSI into smaller images, so-called patches, (ii) reducing the dimensionality of the image patches to one-dimensional vectors using a *patch encoder*, and (iii) mapping all features of a patient sample or WSI with a *feature aggregation* network, which can involve more than 10^4 features, to the final prediction of a clinical endpoint. This can be *molecular* information from additional tests, such as molecular profiling via genetic assays, or protein expressions via immunohistochemistry stainings of additional WSIs, *morphological* information based on the pathologist’s assessment of the WSI, such as tumor subtype, grading, or staging, or *clinical* information, such as patient prognosis, or related therapeutic information, such as treatment response. Refer to Table 2.1 for a detailed overview of current publicly available datasets for clinical endpoint prediction including the available annotations. The most commonly used databases, The Cancer Genome Archiv (TCGA) [152] and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [36], are databases centrally collected from hospitals across the United States by the National Institute for Health providing an abundance of WSI with clinical information of patients linked to genomics (TCGA) and proteomics (CPTAC). Out of these pan-cancer datasets, diverse clinical endpoint tasks can be curated for morphological, molecular, and clinical endpoints. When using these two datasets, batch effects play an important role and site-preserving splits should be used [67]. The evaluations of the publications included in this thesis are mainly based on colorectal cancer cohorts from TCGA and CPTAC, the patch-level classification datasets of colorectal tissue types NCT-CRC-HE-100K [81], and the lymph node metastasis detection dataset Camelyon17 including data from five different hospitals [6].

The following chapter will present the current state of the field regarding the three components of the workflow in computational pathology for clinical endpoint prediction. For each component, the methods are the current state-of-the-art, and the publications in this thesis are presented from a technical perspective and then complemented with current applications in computational pathology. Following the applications in this thesis, the chapter focuses on the application of molecular biomarker prediction such as genetic alterations. Still, the underlying methods are not application-specific

and go beyond this subfield.

Datasets	Organ	#Samples	Targets (#classes)	Task category
BCNB [157]	Breast	1,058	ER (2), PR (2), HER2 (2)	Mol
BRACS [9]	Breast	547 (189)	subtype fine (7), coarse (3)	Morph
EBRAINS [62]	Brain	3,115	diagnosis (32), <i>IDH</i> (2)	Morph, mol
CPTAC [36]	Pan-cancer		Various	Morph, mol, clinical
TCGA [152]	Pan-cancer	29,763 (10,953)	Various	Morph, mol., clinical
PAIP [85]	Pan-cancer	2,457	Various	Morph, mol, clinical
Camelyon16 [95]	Lymph		Cancer	Morph
PANDA [10]	Prostate	11,000	ISUP grade (6)	Morph
IMP [111]	Colorectal	5,333	Grading (3)	Morph

Table 2.1.: **Publicly available datasets with labels for clinical endpoint prediction.**

Selection of tasks that can be used for clinical endpoint prediction, which can be categorized into molecular, morphological, and clinical labels.

2.1. Pre-processing and stain color adaptation

The process from raw tissue to digitized and patched WSIs involves multiple processing steps, each introducing potential variations to the outcome. These include tissue preparation factors such as fixation duration, the age and composition of H&E stains, digitization parameters such as scanner settings and objectives, and WSI patching components such as tissue segmentation and thresholding or artifact detection. These steps result in considerable variability in color appearances of WSIs both across different hospitals but also within the same hospital [67]. Such non-biological variations, commonly referred to as batch effects, are a major factor in the poor generalization of deep learning (DL) models to external datasets [135].

Although reliable tissue segmentation plays an important role in the downstream performance, and is still a challenging task, especially for immunohistochemistry stainings, which can be very bright and therefore difficult to detect, the largest batch effects arise from stain color variability [135]. Approaches addressing batch effects in histopathology fall into two categories: approaches for *stain normalization* that aim at aligning all images to a reference image from the training domain, mapping images from external dataset to the domain seen by the model during training (Figure 2.1a), and approaches for *stain color augmentation*, which introduces stain color variation during training to improve model robustness against batch effects (Figure 2.1b).

From a technical perspective, methods can be further partitioned into traditional, non-DL-based color-matching techniques and generative networks using neural networks

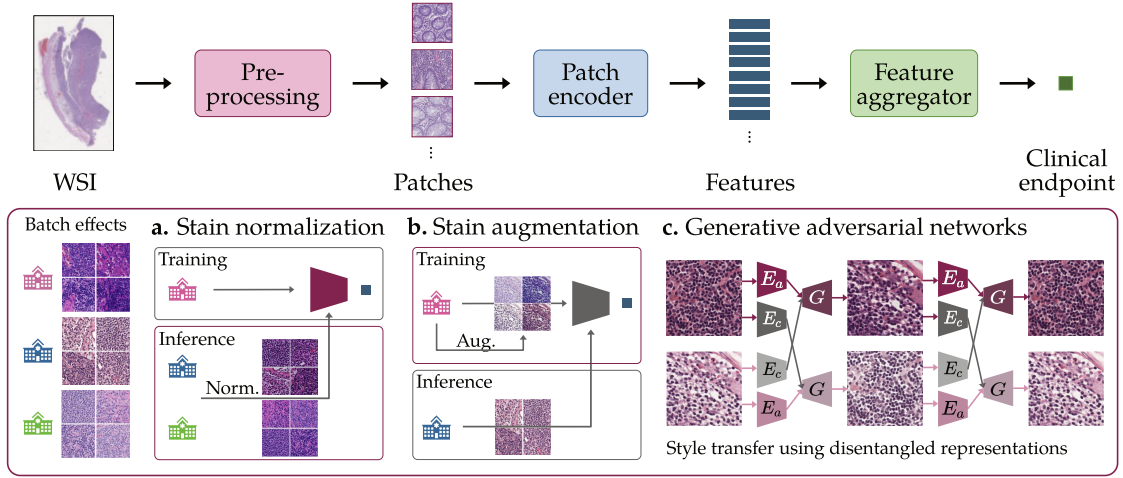


Figure 2.1.: **Pre-processing and stain color adaptation.** (a) Stain normalization methods map test images to the training domain. (b) Stain augmentation methods create diverse synthetic data augmentations to introduce stain variance during model training. (c) Exemplary GAN architecture for style transfer with unpaired image-to-image translation using disentangled representations (DRIT++ [93]).

for style transfer and image-to-image translation [8]. Both approaches can serve for stain normalization or augmentation, where traditionally stain normalization is the more common application.

2.1.1. Color-matching methods

Color-matching techniques generally assume that the proportions of the stain color components, i.e., the hematoxylin and the eosin components, are consistent across images. However, this assumption does not hold across different tissue types or even across images from the same tissue displaying different morphological structures, where the ratio and size of nuclei can vary greatly.

To address the limitations of simple histogram matching based on color space statistics [120], several methods have proposed to apply color deconvolution to identify and adjust H&E components in histology patches [100, 140, 84, 74]. These methods assume that H&E staining components stain distinct morphological structures and are therefore orthogonal to each other in the color space, i.e., can be detected using singular value decomposition. Under this assumption, the RGB color space is decomposed into two principal components—one aligned with hematoxylin, the other with eosin plus

a residual. This allows any source image to be transformed to match the stain color matrix of a target image chosen as a reference.

Despite their wide application potential, color deconvolution methods are computationally expensive, making them impractical for workflows that involve transforming thousands of image patches in a WSI as required for clinical endpoint prediction workflows or large-scale self-supervised training. Additionally, they may fail when the underlying assumption that the stain color proportions of the source and target image are similar is violated leading to artifacts in the transformed images.

2.1.2. Generative adversarial networks

In contrast to color-matching algorithms, Generative Adversarial Networks (GANs) learn the underlying distributions of entire datasets and are therefore not restricted to the color distribution of a single image. Concretely, GANs learn to generate synthetic images by optimizing a min-max objective where the generator G aims to produce realistic samples while the discriminator D is optimized to distinguish real from generated images [48]. Early works on image-to-image (I2I) translation, such as Pix2Pix [73], require paired data to learn style transformations on natural images. However, histopathological images are unavailable in a paired setting as these images are unique to their specific clinical setting. To employ style-transfer techniques from computer vision to histopathological images, GAN-based methods for unpaired style transfer are required.

J.-Y. Zhu et al. [164] introduced CycleGAN for unpaired I2I translation by employing a cycle consistency loss, i.e., mapping an input image through encoder and decoder to the second domain and back to the original domain. Let $M : \mathcal{X} \rightarrow \mathcal{Y}$ and $N : \mathcal{Y} \rightarrow \mathcal{X}$ be the mappings between two domains $\mathcal{X}$ and $\mathcal{Y}$. The training objective combines the classic adversarial loss given by

$$L_{\text{adv}}(M, D_{\mathcal{Y}}) = \mathbb{E}_y[\log D_{\mathcal{Y}}(y)] + \mathbb{E}_x[\log(1 - D_{\mathcal{Y}}(M(x)))] \quad (2.1)$$

as an example for an image $x \in \mathcal{X}$ (similarly for an image $y \in \mathcal{Y}$) with the cycle consistency loss defined as

$$L_{\text{cyc}}(M, N) = \mathbb{E}_x[\|N(M(x)) - x\|_1] + \mathbb{E}_y[\|M(N(y)) - y\|_1]. \quad (2.2)$$

The total loss ensures that the domain-specific mappings M and N preserve domain characteristics and generate realistic images. In this setting, the latent spaces are domain-specific, which can be adapted to a shared-latent space [96].

Despite its success, GANs like CycleGAN may suffer from hallucinations as the generator-discriminator interplay can introduce and ignore artifacts that resemble

network filters [27]. To address this, H.-Y. Lee et al. [93] propose Disentangled Representations for Image Translation (DRIT) to disentangle image content and style into separate representations, improving control over translation and thereby reducing artifacts (Figure 2.1c). In DRIT, images $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ are decomposed into content $z_x^c \in C$ and attribute $z_x^a \in A_x$ via encoders E_x^c and E_x^a :

$$\{z_x^c, z_x^a\} = \{E_{\mathcal{X}}^c(x), E_{\mathcal{X}}^a(x)\}, \quad \{z_y^c, z_y^a\} = \{E_{\mathcal{Y}}^c(y), E_{\mathcal{Y}}^a(y)\}. \quad (2.3)$$

To ensure that the content embedding is domain-invariant, i.g., style-invariant, a content adversarial loss defined as

$$\begin{aligned} L_{\text{adv}}^{\text{content}} = & \mathbb{E}_x \left[\frac{1}{2} \log D^c(E_{\mathcal{X}}^c(x)) + \frac{1}{2} \log(1 - D^c(E_{\mathcal{X}}^c(x))) \right] \\ & + \mathbb{E}_y \left[\frac{1}{2} \log D^c(E_{\mathcal{Y}}^c(y)) + \frac{1}{2} \log(1 - D^c(E_{\mathcal{Y}}^c(y))) \right], \end{aligned} \quad (2.4)$$

where D^c is the content discriminator, reinforces that the domain cannot be detected from the content latent embedding z^c . To obtain disentangled encoders and decoders for image content and style, DRIT employs a cross-cycle consistency loss. Given a non-paired images (x, y) , style representations z_x^a and z_y^a are swapped during forward and backward translations

$$u = G_{\mathcal{X}}(z_y^c, z_x^a), \quad v = G_{\mathcal{Y}}(z_x^c, z_y^a), \quad \hat{x} = G_{\mathcal{X}}(z_v^c, z_u^a), \quad \hat{y} = G_{\mathcal{Y}}(z_u^c, z_v^a). \quad (2.5)$$

The reconstruction loss is defined as

$$L_1^{\text{cc}} = \mathbb{E}_{x,y} [\|G_{\mathcal{X}}(z_v^c, z_u^a) - x\|_1 + \|G_{\mathcal{Y}}(z_u^c, z_v^a) - y\|_1], \quad (2.6)$$

enforces consistency by comparing the ℓ_1 -reconstruction error of the original images and their synthetic cross-cycle counterparts. DRIT can be extended to multiple modalities by removing the domain dependence of content and of domain encoders and decoders. The domain encoders and decoders are formulated as a variational autoencoder conditioned on the domain identifier.

These disentangled representations enable DRIT to achieve robust style transfer while preserving structural integrity, making DRIT particularly suited for histopathological applications. Recent developments of diffusion models as generative networks were also applied to the domain of histopathology for stain color adaptation [133] or proving to make models privacy-preserving [112].

2.2. Representation learning for encoding histology patches

“The success of machine learning algorithms generally depends on data representation, and we hypothesize that this is because different representations can entangle

and hide more or less the different explanatory factors of variation behind the data”, Bengio et al. [7].

Representation learning refers to training DL models to transform input data into above mentioned lower-dimensional feature representations. Finding general-purpose representations for histology patches is a key component for dimension reduction necessary when developing DL models for clinical endpoint prediction, where the representation quality of the patch embeddings directly impacts downstream performance (Figure 2.2). Initially, the field relied on transfer learning from models pretrained on large-scale, labeled datasets such as ImageNet [29], which enabled small histopathology datasets to benefit from pretrained feature representations that could detect distinct patterns [35]. However, representations from supervised learning were shown to be suboptimal, especially for domain shifts in the data typical for natural images to histopathology data transitions [20].

Self-supervised learning (SSL) approaches aim at learning representations from unlabeled data to produce task-agnostic yet semantically rich representations [21]. The idea is to learn representations with *multi-view invariance*. By comparing multiple augmentations (or “views”) of the same image, these methods aimed to learn representations invariant to transformations that preserve the underlying semantic content. Such invariances emerged naturally from proxy tasks designed around predicting missing information, repositioning patches, or grouping similar content, ultimately guiding the network towards representations that captured robust and task-agnostic features.

Following Balestriero et al. [5], I provide an overview of current methods in SSL divided into early approaches, that mostly use proxy tasks as training objectives, contrastive learning, self-distillation, and multimodal alignment. I conclude the section with an overview of the current methods in representation learning for patch encoders in histopathology and how the SSL recipes from computer vision translate to computational pathology.

2.2.1. Proxy tasks for representation learning

Early attempts at self-supervised representation learning primarily focused on designing so-called *proxy tasks* that replaced manual annotations from supervised tasks with artificial labels of the proxy task to learn features from unlabeled images and thereby extend the pretraining datasets to datasets orders of magnitudes larger than annotated datasets. By solving these auxiliary problems, models are encouraged to capture semantics and scene understandings that prove later useful when applied to downstream tasks.

A prominent line of work explored *information restoration* such as colorization tasks that require the model to predict the original RGB colors for grayscale images [161,

91, 143]. By doing so, the network has to understand object boundaries, textures, and spatial relationships to restore realistic colors. Similarly, *inpainting* tasks [118] involve reconstructing missing parts of an image requiring the model to form a coherent internal representation of scene structure and object semantics. Building on these ideas, more recent approaches such as masked autoencoder (MAE) [58], extended this concept by predicting masked patches in an image to learn representations similar to methods in large language models [30].

Spatial context and geometric properties can be formulated as proxy tasks through rotation prediction [46] pushing the model to identify the correct image orientation leading to rotation-invariant features, solving “Jigsaw” puzzles, i.e., arranging shuffled image patches into the correct spatial configuration, further encouraging the model to focus on object parts and their relative positions [34, 114], or counting objects within an image, pushing the model to discover recurring patterns that typically corresponded to distinct semantic entities [114].

Another form of indirect supervision is *clustering-based approaches*: These methods cluster unlabeled images in the latent space and thereby create pseudo-labels for semi-supervised objectives [15, 16]. Additionally, *generative modeling* can be seen as a precursor for SSL objectives, where classical autoencoders learn latent spaces by reconstructing the input image from a bottleneck layer. Although the reconstructions were often blurry and hence the representations less universal than later contrastive methods, autoencoders set the stage for using reconstruction as a learning signal [7, 5].

2.2.2. Contrastive learning for self-supervision

Contrastive learning is a self-supervised representation learning approach with the underlying principle that similar instances should have similar representations, i.e., map to nearby points in the embedding space, while dissimilar instances should have distinct representations, i.e., be placed far apart in the embedding space.

In other words, when two data points share the same class, their representations should be pulled closer together, and representations of data points from different classes should be pushed apart. This concept builds on ideas from metric learning, where the *triplet loss* [151] was originally proposed in a fully supervised context. Assume we have a set of embeddings $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots\}$ and a predefined set of *positive* pairs $\mathbb{P}$ indicating which samples belong to the same class. For each anchor embedding $\mathbf{z}_i$, we identify a positive embedding $\mathbf{z}_j$ that should be close to $\mathbf{z}_i$, and a negative embedding $\mathbf{z}_k$ that should be farther away. The triplet loss, with a margin $m > 0$ to enforce a

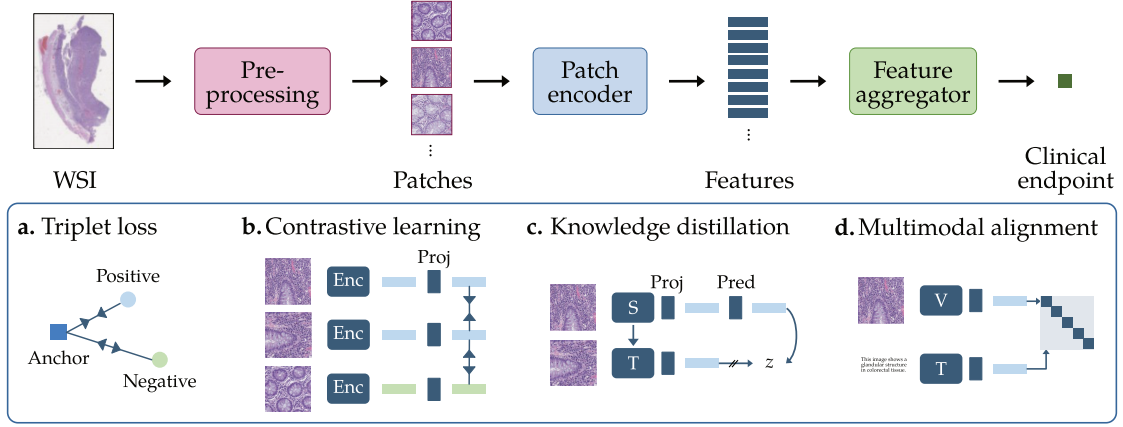


Figure 2.2.: **Representation learning for encoding histology patches.** Initial self-supervised learning objectives originated from metric learning objectives such as the triplet loss (a) and introduced contrastive learning (b). Current successful recipes involve mostly knowledge distillation (c) or multimodal alignment using image-caption pairs (d).

minimum gap between positive and negative distances, is formulated as

$$\mathcal{L}_{\text{triplet}}(\mathbf{Z}) = \sum_{(i,j) \in \mathbb{P}} \sum_{\substack{(k,l) \notin \mathbb{P} \\ k=i}} \text{ReLU}(\|\mathbf{z}_i - \mathbf{z}_j\|_2 - \|\mathbf{z}_i - \mathbf{z}_k\|_2 + m), \quad (2.7)$$

where $\|\cdot\|_2$ denotes the Euclidean distance and ReLU ensures that the loss only applies when the margin constraint is violated.

The classical triplet loss requires sampling techniques such that a batch contains all relevant examples, also called negative mining. Since unsupervised data does not include class labels, contrastive self-supervised learning introduces a new paradigm: images and transformed versions of it (“views”) are used as proxies for being of the same class (*positives*) while other images are treated as *negative* pairs (Figure 2.2b). As transformations, simple augmentations are applied to each image to generate views of the image, which preserve the underlying semantics. These augmentations can be transformations such as random crops, color jittering, or Gaussian blurring.

Oord et al. proposed the *InfoNCE loss* [116] for contrastive learning, which is closely related to Noise-Contrastive Estimation [154]. Suppose we have a set of representations $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N\}$ extracted from a batch of N samples. For a pair $(i, j) \in \mathbb{P}$, i.e., i and j are positive pairs corresponding to augmented views of the same image, the InfoNCE

loss is defined as

$$\begin{aligned}\mathcal{L}_{\text{infoNCE}} &= - \sum_{(i,j) \in \mathbb{P}} \log \left(\frac{e^{\cos(\mathbf{z}_i, \mathbf{z}_j)/\tau}}{\sum_{k=1}^N e^{\cos(\mathbf{z}_i, \mathbf{z}_k)/\tau}} \right) \\ &= - \sum_{(i,j) \in \mathbb{P}} \log \left(\text{softmax}_j \left(\frac{\cos(\mathbf{z}_i, \mathbf{z}_j)}{\tau} \right) \right),\end{aligned}\quad (2.8)$$

where the cosine similarity between two embeddings is defined as $\cos(\mathbf{z}_i, \mathbf{z}_j) = \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}$, and the temperature $\tau > 0$ is a hyperparameter that controls the sharpness of the distribution. Unlike the triplet loss, InfoNCE takes advantage of a non-parametric softmax normalization over all samples in the batch.

This direct, pairwise comparison approach avoids the need for explicit negative mining, i.e., the need to create representative batches that contain negative examples, and naturally encourages learning robust, transferable representations without requiring any labels. On the downside, however, contrastive learning objectives such as the InfoNCE loss require large batch sizes to have representative examples and cover many different images and variations.

2.2.3. Self-distillation for self-supervision

Knowledge or self-distillation methods employ a pair of encoders using the same architecture, commonly referred to as the *teacher* and *student* networks, to learn robust representations without negative samples. This paradigm was first introduced by BYOL (Bootstrap Your Own Latent) [51], which showed that SSL can bootstrap meaningful features by predicting the representation of one network from another for the same input image transformed by different augmentations (Figure 2.2c).

Concretely, consider an unlabeled dataset X and two distributions of semantic-preserving transformations T_1 and T_2 . For an input image $x \in X$, we draw two transformed views $t_1(x)$ and $t_2(x)$, with $t_1 \sim T_1$ and $t_2 \sim T_2$. Let f_{θ_s} and f_{θ_t} denote the student and teacher networks parameterized by θ_s and θ_t , respectively, and let p_γ be a predictor parameterized by γ that is added to the student’s output. The BYOL objective encourages the student’s representation of one view to predict the teacher’s representation of the other as follows

$$\mathcal{L}_{\text{BYOL}}(\theta_s, \gamma) = \mathbb{E}_{(x, t_1, t_2)} \left[\left\| \text{renorm}(p_\gamma(f_{\theta_s}(t_1(x)))) - \text{renorm}(f_{\theta_t}(t_2(x))) \right\|_2^2 \right], \quad (2.9)$$

where $\text{renorm}(\mathbf{v})$ normalizes the embedding vector $\mathbf{v}$ by

$$\text{renorm}(\mathbf{v}) = \frac{\mathbf{v}}{\max(\|\mathbf{v}\|_2, \epsilon)}, \quad (2.10)$$

$\|\cdot\|_2$ is the Euclidean norm, and $\epsilon > 0$ is a small constant to avoid division by zero. By preventing gradients from flowing back into the teacher encoder and instead updating its parameters by the exponential moving average (EMA) of the student’s parameters, BYOL avoids collapsing the network outputs to trivial solutions. X. Chen et al. [22] show that the stop-gradient operation is the critical factor to avoid collapse and EMA update additionally boosts performance.

Building on this framework, DINO [17] replaces the mean-squared error objective and ℓ_2 -normalization with a cross-entropy (CE) loss over a softmax distribution. Letting τ be the fixed temperature parameter of the softmax and $\text{sg}(\cdot)$ indicate the stop-gradient operation, the DINO objective is given by

$$\mathcal{L}_{\text{DINO}}(\theta_s, \gamma) = \mathbb{E}_{(x, t_1, t_2)} \left[\text{CE} \left(\frac{e^{f_{\theta_s}(t_1(x))/\tau}}{\sum_k e^{(f_{\theta_s}(t_1(x)))_k/\tau}}, \text{sg} \left(\frac{e^{\text{center}(f_{\theta_t}(t_2(x)))/\tau}}{\sum_k e^{(\text{center}(f_{\theta_t}(t_2(x))))_k/\tau}} \right) \right) \right]. \quad (2.11)$$

In this formulation, the teacher’s output is first centered to reduce sensitivity to large batch sizes. The teacher’s weights are also updated by the EMA of the student’s weights. The resulting framework can be interpreted as implicitly clustering representations into discrete modes. Further enhancements, such as iBoT [163], integrate masked image modeling inspired by masked token prediction in large language models [30]. Subsequent work, e.g., DINOv2 [117] refines these ideas and is trained on the largest collection of images dataset to date, demonstrating that self-distillation is a powerful method at scale, and yielding transferable and semantically meaningful visual representations.

2.2.4. Multimodal alignment for self-supervision

Self-supervised representation learning has evolved beyond single-modal vision encoders or large language models, extending to multimodal models combining modalities, e.g., visual data with text or audio [47]. By combining multiple modalities, models can learn a joint embedding space in which the modalities, e.g. image and text, are aligned, allowing them to interpret and leverage richer semantic cues than vision alone can offer [119, 77]. In analogy to the view creation through transformations in vision-only models, the core idea is to treat matched image-caption pairs as positive instances and unrelated image-text combinations as negatives, thus applying a contrastive objective that encourages semantically consistent alignment across modalities (Figure 2.2d).

Radford et al. propose Contrastive Language-Image Pre-training (CLIP) [119], which learns a shared embedding space for images and their corresponding captions. Let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ be a batch of N image-caption pairs, where $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the image

and text inputs, respectively. Two encoders, f_{θ_I} for images and f_{θ_T} for text, produce normalized feature embeddings $\mathbf{z}_i^I = f_{\theta_I}(\mathbf{x}_i)$ and $\mathbf{z}_i^T = f_{\theta_T}(\mathbf{y}_i)$. Defining the cosine similarity by $\cos(\mathbf{z}_a, \mathbf{z}_b) = \frac{\mathbf{z}_a^\top \mathbf{z}_b}{\|\mathbf{z}_a\|_2 \|\mathbf{z}_b\|_2}$ and employing a temperature parameter τ , the CLIP objective is defined as

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2} \left[-\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\cos(\mathbf{z}_i^I, \mathbf{z}_i^T)/\tau}}{\sum_{j=1}^N e^{\cos(\mathbf{z}_i^I, \mathbf{z}_j^T)/\tau}} - \frac{1}{N} \sum_{i=1}^N \log \frac{e^{\cos(\mathbf{z}_i^T, \mathbf{z}_i^I)/\tau}}{\sum_{j=1}^N e^{\cos(\mathbf{z}_i^T, \mathbf{z}_j^I)/\tau}} \right]. \quad (2.12)$$

The symmetric contrastive objective encourages each image to match closely with its paired caption and each caption with its paired image, pushing the two modalities to share the same semantics in a joint latent space. Models trained with vision-language objectives often yield representations that transfer robustly to a wide range of downstream tasks, including zero-shot classification and retrieval. Nevertheless, this language-based semantic alignment sometimes biases the model toward semantic over visual similarity, which may not be optimal for certain computer vision problems [108]. Subsequent work such as CoCa [160] extends the contrastive objective with generative text prediction enabling the models to also generate captions.

2.2.5. Self-supervised learning for patch encoder in histopathology

Learning representations for histopathology images faces unique challenges due to the considerable domain shift between pathology image patches and natural images, such that neural networks pre-trained on datasets of natural images like ImageNet often produce suboptimal representations for this domain. In addition, pathology image patches can be highly homogeneous for a human observer while still exhibiting strong batch effects. As we have seen above, vision representation learning is based on the assumption that semantically close inputs should have representations that are closely embedded in the latent space and vice versa for different inputs. This assumption is based on semantic preserving view creation through standard transforms such as cropping, color jitter, and Gaussian blurring. Defining such semantic-preserving augmentation in the medical domain, however, is not straightforward and standard transformation can violate underlying semantics [88]. Particularly for the domain of histopathology, transformations for SSL recipes need to be adapted as pathology images have no inherent orientation and color transformation can change the semantics.

To address this, early efforts leveraged large public WSI repositories, such as TCGA [152] to train models specifically on histopathology data. These domain-specific models improved sensitivity to the homogeneity and appearance of image patches. These models were trained applying early SSL recipes often using a proxy tasks (refer to Section 2.2.1 for more details) such as tissue site classification [122], a readily available

2. Methods

Model	Data			Method		Availability		
	WSIs	Patches	Datasets	Training	Backbone	D	C	M
KimiaNet [122]	7K	240K	TCGA	Supervised	DenseNet121	✓	✓	✓
CiGa [26]	25K	4M	57 public	SimCLR	-	✓	-	✓
CTransPath [149]	32.2K	15M	TCGA, PAIP	MoCov3	Swin-T	✓	✓	✓
RetCCL [148]	32K	15M	TCGA, PAIP	Custom	ResNet50	✓	-	✓
Remedis [3]	29K	50M	TCGA	SimCLR	ResNet50	✓	-	✓
Lunit [79]	37K	33M	TCGA, TULIP	Various	ViT-S	-	-	✓
Phikon [40]	6K	46M	TCGA	iBOT	ViT-B	-	-	✓
PLIP [68]	-	243K	Med. Twitter	CLIP (VL)	ViT-B	-	-	✓
CONCH [97]	-	1.17M	PubMed	CoCa (VL)	ViT-B	-	-	✓
UNI [20]	100K	100M	MASS-100K	DINOv2	ViT-S	-	-	✓
Quilt-1M [70]	-	1M	YouTube	CLIP (VL)	ViT-B	-	-	✓
Virchow [144]	1.5M	380M	MSKCC	DINOv2	ViT-H	-	-	✓
Campanella et al. [13]	400K	3B	Mount Sinai	MAE/DINO	ViT-S	-	-	✓
Rudolf-V [33]	103K	750M	TCGA, Prop.	DINOv2	ViT-L	-	-	-
Roth et al. P4.	632	2.5M	TCGA	DINOv2	ViT-S/g	✓	✓	✓
PLUTO [78]	158K	195M	Prop.	Various	ViT-S	-	-	✓
Prov-GigaPath [158]	170K	1.3B	Providence	DINOv2	ViT-g	-	-	✓
Hibou [109]	1.1M	1.2B	Prop.	DINOv2	ViT-L	-	-	✓
H-Optimus-0 [125]	500K	-	Prop.	DINOv2	ViT-g	-	-	✓
Virchow2 [165]	3.1M	1.9B	MSKCC	DINOv2	ViT-G	-	-	✓
Phikon v2 [41]	60K	450M	3 public	DINOv2	ViT-L	-	-	✓

Table 2.2.: **Overview of existing patch encoders** sorted by their release date. The following information was collected for every model if the information was available: information about the training datasets including the number of WSIs (patients) and the number of patches used during training, methodological details such as training framework, backbone architecture, and whether the latent space is trained with vision-language (VL) alignment, finally, the data (D), training code (C), and model (M) availability is noted.

annotation. Other approaches introduced clustering-based loss terms designed for WSIs, ensuring that patches from the same WSI were embedded closer together than those from different WSIs [148], or extended state-of-the-art SSL algorithms, such as MoCov3 [23], with pathology-specific adaptations [149].

Recent advances in SSL, coupled with the increasing availability of digitized WSIs, have pushed representation learning in computational pathology to the next level. Successful SSL recipes such as knowledge distillation with DINO [17], masked autoencoders (MAE) [58], the combination of masked image modeling with knowledge distillation in iBot [163] and its extensions in DINOv2 [117], improved representation learning in natural images (Section 2.2.3). These recipes were subsequently heav-

ily adopted in computational pathology leading to emerging patch-level foundation models trained on large repositories of WSIs exceeding the publicly available repositories (Table 2.2).

Parallel to the developments in SSL for general computer vision, where models such as CLIP [119] used image captions to contrast vision embeddings with orthogonal information from language descriptions, vision-language training recipes were adopted in computational pathology, e.g., using databases such as Twitter [68], YouTube [70], and medical textbooks together with PubMed [97] to generate paired pathology image-caption data opening up new use cases such as zero-shot classification.

2.3. Weakly-supervised learning for histopathology

In computational pathology, weakly-supervised learning uses slide-level annotations rather than pixel- or patch-level annotations in the training procedure and comes with a lower manual annotation effort. After dimensionality reduction via patch encoding, a WSI is represented as a collection of feature vectors, referred to as a *bag*. In the context of clinical endpoint prediction, each bag corresponds to one or more clinical labels. This framework, known as *multiple instance learning* (MIL), addresses the inherent ambiguity in weakly-supervised settings, where there is no explicit one-to-one correspondence between individual instances (patches) in the bag and the slide-level label. Instead, individual instances contribute to the predicted target to a variable extent, reflecting their varying relevance to the clinical outcome.

Let $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_K\}$ represent a bag of K embeddings extracted from K patches of a WSI. Weakly-supervised learning aims to aggregate the bag $\mathbf{x}$ into a single vector $\mathbf{z}$, which can further be mapped to a target $\mathbf{y}$ using a classifier. Two simple, parameter-free aggregation methods are *max pooling* and *mean pooling*. For max pooling, the maximum value of all embeddings along the embedding dimension is chosen, i.e.,

$$\mathbf{z} = \max_{k \in \{1, \dots, K\}} \mathbf{x}_k, \quad (2.13)$$

while for mean pooling, the patch embeddings are averaged along the embedding dimension, i.e.,

$$\mathbf{z} = \text{mean}_{k \in \{1, \dots, K\}} \mathbf{x}_k. \quad (2.14)$$

Both methods require no optimization and serve as a robust unsupervised baseline, outperforming other supervised trained aggregation methods when combined with powerful patch encoders [75, 76].

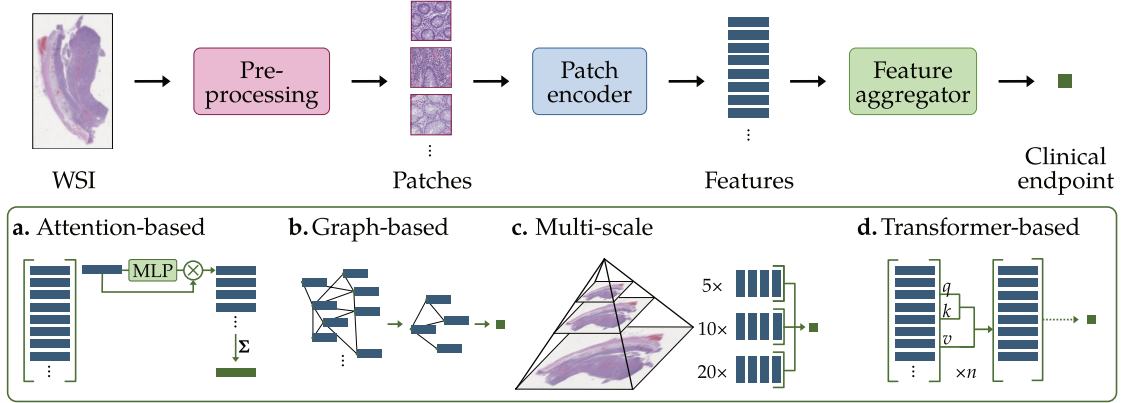


Figure 2.3.: **Weakly-supervised learning for histopathology.** Architectures for feature aggregation networks can be divided into (a) attention-based networks, (b) graph-based networks, (c) multi-scale networks, and (d) Transformer-based networks.

In the following, I will introduce MIL approaches that go beyond simple pooling mechanisms, including attention-based networks, graph convolutional networks, multi-scale approaches, and Transformer-based networks, which all improve aggregation by transforming and weighting the bag instances in $\mathbf{x}$.

2.3.1. Attention-based networks

Attention-based networks improve upon simple pooling aggregation methods by learning a weighted average of patch embeddings. Ilse et al. [71] propose to apply a two-layer multi-layer perceptron (MLP) to each feature vector, producing an *attention* score for each patch embedding. These attention scores are then used to calculate a weighted average of the individual patch embeddings to obtain a single WSI embedding (Figure 2.3a).

Given a bag of embeddings $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_K\}$, where $\mathbf{x}_k \in \mathbb{R}^d$ represents the d -dimensional embedding of the k -th patch, the final aggregated representation $\mathbf{z} \in \mathbb{R}^L$ is given by

$$\mathbf{z} = \sum_{k=1}^K a_k \mathbf{x}_k, \quad (2.15)$$

where a_k is the attention score for $\mathbf{x}_k$. The attention scores are calculated by

$$a_k = \frac{e^{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{x}_k^\top)}}{\sum_{j=1}^K e^{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{x}_j^\top)}}, \quad (2.16)$$

, where $\tanh$ introduces a non-linearity, $\mathbf{w} \in \mathbb{R}^L$ and $\mathbf{V} \in \mathbb{R}^{L \times d}$ are trainable parameters of the MLP, and the softmax function ensures the attention weights sum to 1.

This mechanism allows the network to prioritize regions of interest, such as tumor regions, within the WSI, producing a weighted representation $\mathbf{z}$ that is subsequently passed to a classification head for clinical endpoint prediction. Attention-based networks can vary in their size and architecture of their MLP by including pre-attention transformations, multi-head attention [75], and post-processing techniques like feature clustering [99].

However, a limitation of attention-based networks is that each patch is taken independently, ignoring relationships between patches by neglecting the spatial context of patches from the WSI. While patches may capture local contexts, such as cell interactions within regions approximately of size $256 \mu\text{m}^2$ at $20\times$ magnification (512×512 pixels at $0.5 \mu\text{m}/\text{pixel}$), broader contextual information at the region or slide level is lost. *Context-aware approaches* address this limitation by modeling interactions between feature vectors, enabling a more holistic understanding of the tissue microenvironment.

2.3.2. Graph neural networks and multi-scale networks

Graph neural networks (GNNs) offer a natural way to incorporate spatial and topological information into patch feature aggregating models. Rather than treating each patch independently, GNN-based approaches represent a WSI as a graph, where each node corresponds to a feature embedding of a patch, cell, or other relevant tissue structure, such as glands [162, 50]. The edges between nodes capture spatial adjacency, enabling the model to propagate contextual information across the graph through graph convolutional layers. Additionally, GNNs can handle unstructured or missing data more naturally than grid-based methods, a common challenge when handling irregularly shaped WSIs.

For instance, one can form a graph by connecting patches that are spatially close within the WSI, and then apply graph convolutions or related GNN operations to model local context and interactions [162]. Alternatively, the graph can be constructed at a finer granularity, representing cells or glands as nodes to capture distinct morphological or functional relationships [50]. Hierarchical or multilevel GNNs further enrich this representation by modeling the information from multiple resolutions of the WSI separately, capturing both fine-grained cellular details and higher-level structural organization, and fusing the information later in the network [65, 55, 64].

Similarly, multi-scale networks take inspiration from the pathologist’s workflow, where tissue is examined across varying magnifications, e.g., the tumor outlines can be detected at $2\text{-}5\times$ up to cell-level features at $40\times$. Such approaches process image features derived from multiple pyramid levels of a WSI and fuse them into one representation throughout aggregation. By combining low-magnification global context with high-magnification local detail, these networks can improve performance in tasks such as tumor classification and grading [136] (Figure 2.3c).

2.3.3. Transformer-based networks

The shift from recurrent architectures to *self-attention*-based networks, called *Transformers*, has expanded modeling capabilities in computational pathology following their success in natural language processing. With Transformer-based models, large sequences of patch embeddings from WSIs can be processed to produce a single output capturing the interdependencies of all patches among each other [146].

Transformers are based on the self-attention mechanism, also called *scaled dot-product attention*, which computes pairwise interactions between elements of an input sequence (Figure 2.3d). In particular, assuming that $\mathbf{x} \in \mathbb{R}^{N \times d}$ is the input sequence representing a WSI with N patch embeddings of dimension d , the self-attention layer computes a query-key product by

$$\text{SA}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V, \quad (2.17)$$

where queries $Q \in \mathbb{R}^{N \times d_k}$, keys $K \in \mathbb{R}^{N \times d_k}$, and values $V \in \mathbb{R}^{N \times d_v}$ such that $\text{SA}(Q, K, V) \in \mathbb{R}^{N \times d_v}$. These are computed from the input sequence $\mathbf{x}$ by

$$Q = \mathbf{x} \cdot W_Q, \quad K = \mathbf{x} \cdot W_K, \quad \text{and} \quad V = \mathbf{x} \cdot W_V, \quad (2.18)$$

where $W_Q \in \mathbb{R}^{d \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$, and $W_V \in \mathbb{R}^{d \times d_v}$ are learnable parameters and d_k and d_v are the dimensionality of the queries, keys, and values, respectively. Multi-headed self-attention applies self-attention in parallel and concatenates each head in a weighted manner

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W_O, \quad (2.19)$$

where

$$\text{head}_i = \text{SA} \left(Q^{(i)}, K^{(i)}, V^{(i)} \right) \quad \text{for } i \in \{1, \dots, h\}, \quad (2.20)$$

and $W_O \in \mathbb{R}^{hd_v \times d}$ the projection of the concatenated h heads.

While self-attention provides a powerful mechanism for modeling sequence dependencies, its quadratic complexity constitutes a computational limitation for long sequences as common for WSIs consisting of 10^4 patch embeddings or more. To address this, various approximations such as Nystroem-attention [132] or local-attention [121] have been proposed, allowing the model to handle long feature sequences more efficiently. More recent advances in handling long sequences from large language models have quickly informed histopathology modeling. For instance, state space models [54, 52, 53] provide an alternative to self-attention and have been adapted as MambaMIL [159] and similar works [42, 38]. Improvements in long-context self-attention mechanisms, such as dilated attention [31], have also translated to histopathology, as demonstrated by LongViT [147].

The above-mentioned methods ensure that patch features are not considered independently. However, contextualizing patch embeddings with plain multi-head self-attention in Transformers does not encode the position among the features yet to retain the global context. Unlike natural images, WSIs are irregularly shaped with greatly varying dimensions and hence are represented by a variable number of patch features. Additionally, WSIs lack an absolute orientation. While for natural images orientations are a crucial factor in visual reasoning, orientation can be neglected in pathology images. Designing suitable two-dimensional positional encodings that preserve the spatial context and handle variable-size inputs is therefore non-trivial. Approaches include rearranging features into 2D grids [132], restricting self-attention to local neighborhoods [121], defining a maximal WSI size and learning fixed embeddings for all smaller sizes [158], or using absolute positional encodings within fixed regions of interest (ROIs) [19]. Each of these strategies aims to retain global context and spatial structure without imposing assumptions about orientation or uniform size that are common in natural images, still, a universal solution for this challenge has not been proposed yet.

2.3.4. Representation learning on slide-level

Rather than training a feature aggregation network from scratch, whole slide representation models can be pretrained to distill pathology-specific knowledge from large WSI collections to simplify clinical endpoint prediction. The challenge then lies in developing a suitable pretraining strategy for the feature aggregation network that can encode tissue context from sequences of patch embeddings while handling arbitrarily large WSIs.

Although relatively underexplored, slide-level self-supervision can be performed with vision-only pretraining, either through masked image reconstruction [158], using tissue classification as proxy task [150], or intra-slide contrastive learning [92, 66, 87], or in a multimodal fashion involving pathology reports, bulk transcriptomics,

or immunohistochemistry [139, 131, 158, 75, 76, 2]. Furthermore, long-range context modeling can either be neglected, essentially treating a WSI as a bag of independent features[92, 75, 131, 158, 150], or explicitly modeled using Transformers[19, 139, 158, 66].

As research into general-purpose slide representations intensifies, adapting successful patch-level representation learning strategies to the entire WSIs holds the promise of powerful general-purpose slide representations that can be used off-the-shelf for downstream tasks. Inspired by knowledge distillation and masked-image modeling[163, 117] for patch encoder pretraining (Section 2.2.3), we developed a multimodal, vision-language pretraining paradigm that leverages millions of high-resolution regions-of-interest (ROIs at $8,192 \times 8,192$ pixels) for large-scale, resolution-agnostic pretraining and scalable WSI encoding. Trained using 336K WSIs across 20 organ types, our vision-only trained model produces general-purpose slide representations that can readily be applied to slide-level tasks such as cancer subtyping, biomarker prediction, outcome prognosis, and slide retrieval tasks, outperforming supervised baselines and existing multimodal slide foundation models [32]. This work is currently under review and hence not included in the scope of this thesis.

3. Publications

In the following, I present the publications included in this thesis by detailing for each publication, first, the motivation and background, second, the existing work and its shortcomings, then, the methods proposed in the publication, and finally, my contributions. The publications tackle challenges arising when developing deep learning (DL) for clinical endpoint prediction from digitized whole-slide images (WSIs) in computer-aided cancer diagnosis (Figure 3.1).

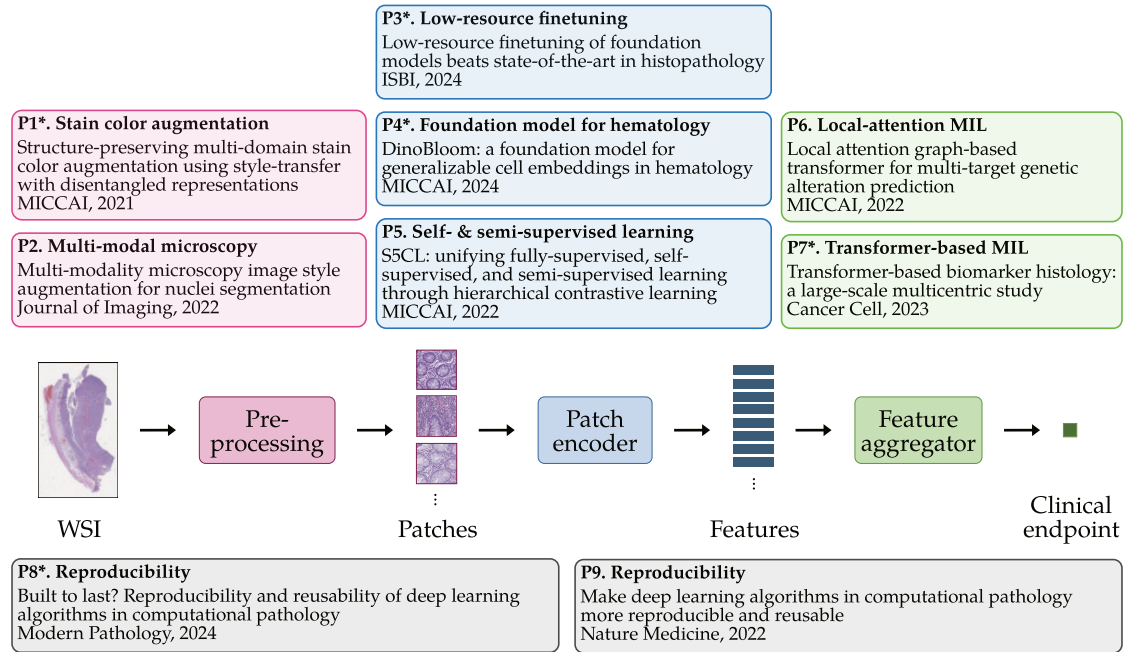


Figure 3.1.: **Overview of publications included in this thesis**, positioning the publications in the context of the computational workflow for clinical endpoint prediction in computer-aided cancer diagnosis.

3.1. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations

Variations in staining procedures and scanner settings across different laboratories introduce strong color differences in WSIs [153]. When applied to external datasets, these color shifts lead to poor generalization of DL models trained on WSIs [135].

Existing methods to counteract this can be divided into stain color normalization techniques and stain color augmentation methods. For stain color normalization, recent approaches have utilized style-transfer methods like cycle-consistent generative adversarial networks (CycleGANs) to transfer different styles between domains [130]. While effective in matching target distributions, these models can inadvertently alter the tissue's morphological structure, referred to as hallucinations. Alternatively, approaches for stain color augmentations have been proposed to make the downstream neural networks robust to stain color variations already during training [135] (Section 2.1). However, both approaches are simple color space transformations and are therefore restricted in handling a wide variability in staining appearances while preserving the morphological structures.

To overcome these limitations, I developed HistAuGAN, a **stain color augmentation technique based on a generative adversarial network** (GAN) for image-to-image translation [P1.]. Based on the multi-domain version of DRIT++ [93], HistAuGAN disentangles the morphological content of histological images from their stain color attributes. This separation allows the model to alter the color while preserving the tissue structure, thus preventing the GAN from hallucinating. By training on multiple domains, HistAuGAN learns to simulate a wide variety of realistic histology stain colors. When used as a stain augmentation technique during the training of a neural network for a downstream task, HistAuGAN renders the network stain-invariant. This means the trained models generalize better to external datasets without additional normalization steps during testing as shown on a binary tumor classification task using the publicly available dataset Camelyon17 including data from five hospitals [6]. Networks trained with HistAuGAN color augmentation outperform networks trained with conventional color augmentation techniques on external datasets. Through clustering analysis, we confirm that HistAuGAN disentangles morphological structures and stain colors and can therefore help to counteract present batch effects more effectively.

Contributions. I developed the concept of the study together with my supervisor Tingying Peng. I implemented the code for training and evaluating HistAuGAN. I drafted the manuscript and the figures except for the clustering analysis.

3.2. Multi-modality microscopy image style augmentation for nuclei segmentation

Nuclei segmentation in microscopy images is a critical task in histopathology and is essential for determining cell-level features. It facilitates correct cancer type identification, grading, and prognosis and strengthens clinical diagnoses. However, annotating images for nuclei segmentation is laborious and time-consuming, requiring medical experts to create accurate segmentation masks. The performance of DL models for automated nuclei segmentation heavily relies on the quality and quantity of these annotations. Additionally, the diversity of microscopy imaging modalities—such as bright-field and fluorescence microscopy—and the various cell assay types introduce challenges. Models often struggle with less-represented image types, and a decreasing accuracy and generalizability across different datasets.

The 2018 Data Science Bowl (DSB) [49] competition aimed at developing unified segmentation algorithms capable of handling various microscopy setups, cell lines, and imaging modalities. Top-performing methods predominantly utilized segmentation architectures like Mask R-CNN [59] and U-Net [123], with performance influenced by the amount of training data, complex pre- and post-processing techniques, and ensemble learning strategies. An early identified issue was dataset bias: models achieved high accuracy on fluorescent images, which comprised most of the training set, but performed poorly on minority imaging modalities like bright-field images of stained tissue sections. While some teams employed style transfer methods for data augmentation, these approaches trained one network per imaging domain [63].

We propose a GAN-based style augmentation technique [P2.]. The method handles images from multiple domains such as varying cell assay types, lighting conditions, and imaging modalities, including bright-field and fluorescence microscopy. Disentangled representations for content and style preserve the original image’s structure (nuclei locations and shapes) while altering its style during augmentation. Using this augmentation technique, **synthetic images of less-represented image types can be generated from well-represented ones preserving the segmentation mask, effectively increasing the number of minority modalities in the training set without resampling.** We evaluate the style augmentation technique on the DSB dataset, by incorporating the augmented images into the training process of two top-ranked Mask R-CNN-based nuclei segmentation algorithms from the competition, which increased segmentation accuracy.

Contributions. I conceptualized the study and provided the initial code base. I contributed to the writing of the manuscript and figures.

3.3. S5CL: unifying fully-supervised, self-supervised, and semi-supervised learning through hierarchical contrastive learning

In computational pathology, annotating histopathological images to obtain labels for semantic segmentation is very time-consuming due to the gigapixel scales of WSIs. This results in a scarcity of annotated WSIs for segmentation tasks but an abundance of unlabeled data, making it challenging to train supervised DL models. The imbalance of little labeled and mostly unlabeled data provides a perfect use case for semi-supervised learning, which leverages the information from unlabelled data to improve model performances.

Current semi-supervised methods typically involve a two-stage process: First, SSL is used to learn suitable representations from unlabeled data (Section 2.2), and, subsequently, the model is fine-tuned using labeled data. However, these methods have limitations: SSL recipes often require large batch sizes, long training times, and many parameters. Additionally, self-supervised representations may struggle with highly imbalanced datasets due to biases towards the majority classes.

To address these challenges, we developed S5CL, **a unified framework that combines fully supervised, self-supervised, and semi-supervised learning into a single-stage training process using hierarchical contrastive learning** [P3.]. S5CL employs three contrastive losses simultaneously for labeled, unlabeled, and pseudo-labeled images. This approach constructs an embedding space that reflects a hierarchy of distance relationships, where similar images are closest, different images from the same class further, and images from different classes furthest apart. The framework is flexible, allowing adaptation to different scenarios by adjusting the weights of the contrastive losses, and can function in fully supervised, self-supervised, or semi-supervised modes depending on the available data. We evaluate the method on two public histopathological datasets: the NCT-CRC-HE-100K colorectal cancer dataset [81] and the Munich AML Morphology dataset of single white blood cells from leukemia patients [104]. Our results show that S5CL improves performance, especially in sparse-label scenarios. For the colorectal cancer dataset, S5CL increases accuracy by up to 9% compared to a supervised cross-entropy loss. For the highly imbalanced leukemia dataset, it improves the F1-score by up to 6%. These findings demonstrate that S5CL effectively leverages both labeled and unlabeled data, handles class imbalances, and enhances feature representation learning in computational pathology.

Contributions. I contributed the idea of hierarchy loss, dataset evaluations, and supervised the project throughout. I further helped with writing the manuscript.

3.4. Low-resource finetuning of foundation models beats state-of-the-art in histopathology

Foundation models in computer vision have shown that distilling knowledge from vast amounts of data can substantially improve feature quality and lead to general-purpose representations that can be used across datasets and tasks such as classification, object detection, or semantic segmentation [117, 47, 138, 86].

These recipes have been adopted in computational pathology to train patch-level foundation models on millions of patches (Table 2.2). However, none of the existing models in computational pathology leverage the knowledge of these large-scale pre-trained models from computer vision, e.g., trained on curated datasets constructed from 1.2B unique images [117]. Efforts to scale up the pre-training for histopathology-specific feature extractors have been resource-intensive, requiring extensive computational power and time and large repositories of clinical data, which are not publicly available, limiting their accessibility and reproducibility.

To overcome these limitations, **we propose a shift toward low-resource fine-tuning of large foundation models from classical computer vision for specific histopathology tasks** [P4.]. We benchmark four prominent vision foundation models as feature extractors for histopathology data, SAM [86], ImabeBind [47], DINOv2 [117], DEit [138], and two feature extractors for histopathology, CTransPath [149], RetCCL [148], evaluating them in both slide-level and patch-level classification settings for colorectal cancer. Specifically, we fine-tune Vision Transformer (ViT) models repurposing the pretrained checkpoints of the effective knowledge distillation approach DINOv2, which is the best performing of the computer vision foundation models. We finetune the checkpoints on publicly available colorectal cancer patches and WSIs from NCT-CRC-100K and TCGA, respectively. Our experiments demonstrate that fine-tuning these foundation models on task-specific data can match or even outperform state-of-the-art domain-specific feature extractors like CTransPath and RetCCL. Remarkably, this performance is achieved using significantly fewer resources—a single GPU for only two hours to three days, depending on the dataset—compared to the large-scale resources required for training such models from scratch. This approach makes high-quality feature extraction accessible to institutions with limited computational capabilities and therefore represents a considerable shift from the current state, where only few institutions with substantial resources can develop such models.

Contributions. I conceptualized the study with Valentin Koch, supported Benedikt Roth in model implementation and training, selected downstream tasks, and wrote large parts of the manuscript.

3.5. DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology

Microscopic examination of peripheral blood and bone marrow smears are essential to make accurate diagnoses in Hematology. For these examinations, experts manually analyze hundreds of white blood cells (WBCs) per patient, making the process labor-intensive and susceptible to variability among observers.

Recent advancements in DL have shown promising attempts in automating manual work such as required for acute myeloid leukemia (AML) subtype classification from microscopic images [60] or human-level blast recognition in AML [104]. However, clinical adoption is limited because existing models rely on supervised learning with small, labeled datasets, restricting their ability to generalize across diverse samples. Challenges include large batch effects, varying cell phenotypes, and poor transfer learning performance from models pretrained on natural images.

To address these limitations, we introduce **DinoBloom, the first foundation model specifically designed for single-cell image analysis in hematology** [P5.]. DinoBloom leverages a customized DINOv2 [117] self-supervised learning (SSL) pipeline based on vision transformers to learn rich visual features from unlabeled images. We trained the model on the largest assembled hematology cohort to date, comprising 13 diverse, publicly available datasets of peripheral blood and bone marrow smears, totaling over 380,000 WBC images. This extensive multi-source dataset enhances the model’s ability to generalize across different domains and mitigates batch effects. These models enable applications in few-shot classification, MIL, and cell embeddings for disease profiling. We evaluate our trained models for WBC classification in peripheral blood smears on the external dataset Acevedo [1] and in bone marrow smears on a held-out test set of the dataset BMC [103], as well as in weakly supervised AML subtyping on a held-out test set of the dataset AML Hehr [60]. Our models outperform existing medical and non-medical vision models on all tasks, even on external datasets with challenging domain shifts. By publicly releasing all DinoBloom models, along with the source code and training parameters, we encourage the research community to build upon our work and advance the field collaboratively.

Contributions. I conceptualized the study together with Valentin Koch and supported Valentin Koch in the methods implementation and model training. I created the figures of the manuscript including the contained analysis and wrote the manuscript together Valentin Koch.

3.6. Local attention graph-based transformer for multi-target genetic alteration prediction

The gigapixel size of WSIs makes their direct analysis challenging, if not impossible. The analysis usually requires a preprocessing step in which WSIs are divided into smaller patches, but diagnoses are typically only available at the slide not at the patch level. Multiple instance learning (MIL) addresses this problem by treating a WSI as a bag of instances (patches). Traditional MIL methods often assume that these instances are independent and identically distributed, neglecting the rich contextual information and spatial relationships between tiles, which are crucial for accurate diagnosis and prediction [71, 128, 99].

Attention-based pooling layers have been used to weight tiles based on their relevance but do not model inter-tile dependencies. Transformers with global self-attention have been proposed to model these relationships comprehensively [132]. However, global self-attention is computationally expensive, especially for WSIs containing thousands of tiles, due to its quadratic complexity in the sequence length. Z. Shao et al. [132] propose to use the linear approximation of self-attention, Nyström-attention [156], to address the long sequence length. However, approximating self-attention always comes at the cost of performance degradation compared to normal self-attention.

Alternatively, we propose LA-MIL, a **local attention graph-based Transformer tailored for MIL in computational pathology** [P6.]. We hypothesize that cells and other cellular structures only affect the neighboring tissue. Therefore, LA-MIL restricts self-attention computations to local neighborhoods by constructing k -nearest neighbor (kNN) graphs among tiles within a WSI, effectively modeling local dependencies without the computational burden of global attention. This method introduces an inductive bias by explicitly contextualizing instances in adaptive local regimes, capturing both spatial and morphological relationships. An adapted loss function facilitates learning multi-task supervision for the joint analysis of multiple biomarkers. We evaluate the architecture for genetic alteration prediction on colorectal and gastric cancer using TCGA [152] and CPTAC [36] as external evaluation cohorts when possible. Our experimental results demonstrate that LA-MIL achieves state-of-the-art performance, notably outperforming existing models in predicting important biomarkers like microsatellite instability in colorectal cancer. These findings suggest that local self-attention mechanisms are sufficient for modeling dependencies in WSIs, providing a computationally efficient and effective alternative to global self-attention Transformers.

Contributions. I conceptualized the study, supervised the project, and helped to write the manuscript.

3.7. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study

Colorectal cancer (CRC) treatment increasingly depends on the identification of genetic biomarkers like microsatellite instability (MSI) and mutations in the *BRAF* and *NRAS/KRAS* genes. Traditional methods for detecting these biomarkers – such as polymerase chain reaction (PCR), sequencing, or immunohistochemical assays – are costly, time-consuming, and often inaccessible in low-resource settings.

Recent studies have shown that DL can predict these biomarkers directly from hematoxylin and eosin (H&E)-stained pathology slides, which are routinely available for all patients with CRC [43, 82]. However, existing DL approaches primarily use convolutional neural networks (CNNs) for feature extraction [83, 35] combined with simple aggregation methods [113]. Furthermore, they have mostly been validated only on small patient cohorts [121]. These methods exhibit limitations in performance and generalizability, especially to biopsy tissues, and fail to capture the contextual relationships between image patches crucial for accurate diagnosis.

To address these limitations, I propose a fully **transformer-based pipeline for end-to-end biomarker prediction from pathology slides** [P7.]. Transformers, originally developed for natural language processing tasks, have shown superior capabilities in modeling long-range dependencies and contextualizing information, making them promising for image analysis tasks that require an understanding of both local and global context. I combine a pre-trained Transformer-based patch encoder for feature extraction with a Transformer-based network for feature aggregation. In our evaluations, this Transformer-based approach improves performance, generalizability, data efficiency, and interpretability compared to current state-of-the-art algorithms that rely on CNNs with attention-based feature aggregation. I evaluate the proposed methods on a large, multi-centric cohort with over 13,000 patients from 16 colorectal cancer cohorts worldwide and on two cohorts of CRC biopsies, totaling over 1,500 patients. The results demonstrate that the Transformer-based approach substantially improves performance for biomarker prediction of microsatellite instability (MSI) and the genetic alterations *BRAF* and *KRAS*. Notably, **the model trained exclusively on resection specimens generalizes well to endoscopic biopsy tissues**, solving a long-standing diagnostic problem by enabling accurate MSI testing on initial biopsy samples.

Contributions. I developed the concept of the study together with Tingying Peng and Jakob Nikolas Kather. I implemented methods and evaluations, evaluated all experiments, and created the figures. I drafted the manuscript with all co-authors.

3.8. Built to last? Reproducibility and reusability of deep learning algorithms in computational pathology

“More than 70% of researchers have tried and failed to reproduce another scientist’s experiments, and more than half have failed to reproduce their own experiments.” states a provoking analysis about a *Nature* survey of 1,576 researchers [4]. These tendencies are further fortified in the fields of machine learning [69] or applications of AI [80].

In computational pathology, we can observe that despite a vastly increasing amount of publications only a few have reached clinical implementation. We hypothesize that this is due to challenges in reproducibility, reusability, and generalizability: Many studies lack available code or provide insufficient instructions, making it difficult for non-DL experts – especially pathologists without computational background – to reproduce results or reuse methods in a clinical setting for research purposes. Additionally, DL algorithms often fail to generalize across different datasets because of variations in staining or scanner settings, underscoring the need for external validation and robustness analysis.

In our analysis, we first define the terms *reproducibility*, *robustness*, *reusability*, and *repliable* in the context of computational pathology. Further, we compile a reproducibility checklist concerning methods, data, and statistical analysis. We systematically analyzed DL methods for the analysis of H&E-stained (WSIs) published between January 2019 and March 2021 [P8.]. We collected 160 studies for cancer diagnosis and treatment, categorizing them into five use cases: (1) stain normalization, (2) tissue type segmentation, (3) evaluation of cell-level features, (4) genetic alteration prediction, and (5) inference of grading, staging, and prognostic information. We introduce each use case, review the latest DL methods—focusing on those that provide accompanying code—and analyze their reproducibility. Additionally, we provide an overview of current data handling tools essential for reading, annotating, and sharing histopathological data. We found that only every fourth publication provides code with the methods, varying drastically between the use cases and can hence conclude that there is a reproducibility issue in methods for computational pathology.

Contributions. I compiled the reproducibility criteria and checklist in discussion with all authors. I analyzed one use case section and coordinated the analysis of the remaining use cases among the authors. I created the figures and the main part of the manuscript with the feedback of all authors. The use case sections were written by the respective authors who analyzed the use cases.

3.9. Make deep learning algorithms in computational pathology more reproducible and reusable

Computational pathology, enhanced by DL algorithms, has the potential to revolutionize diagnostic medicine through the assessment of digitized histological slides. These algorithms can match or surpass human experts in specific tasks and predict information not easily discernible from microscopic images. Despite thousands of publications in this field, only a small fraction of algorithms have been reused by other researchers or integrated into clinical workflows [P8.]. This gap between research advancements and clinical application highlights the urgent need for improved reproducibility and reusability to optimize clinical workflows and enhance patient health.

A review of 161 peer-reviewed articles published between January 2019 and March 2021 revealed significant issues with reproducibility and reusability in computational pathology e.g., only 26% of these publications made their code publicly available and met minimum requirements for reproducibility [P8.]. Based on these findings, in this comment, we compiled the **recommendations on how to address the gap in reusability, reproducibility, and generalizability of methods in computational pathology** [P9.].

First, we suggest to involve clinicians early in the development process. Engaging domain experts, e.g., pathologists, from the outset ensures that algorithms address clinically relevant tasks and incorporate necessary modifications. This collaboration helps define the required data and avoids biases during model training. Second, we recommend publishing the data including the splits used. Whenever possible, publicly shared multi-institutional datasets facilitate future research. Data splits should be performed at patient- and cohort levels to prevent data leakage and overestimation of model performance. Third, transparent data preprocessing and model hyperparameters are required. Detailed documentation of data handling tools, preprocessing steps, and parameters used need to be shared, as well as well-documented code along with final model weights to allow independent external validation. In the best case, ready-to-use scripts and information on hyperparameters, hardware specifications, and training times to enhance reusability.

Finally, comprehensive evaluation and publication practices are necessary. The use of external test sets for final evaluations to assess generalizability including appropriate statistical analyses should be standard practice. Overall, our recommendations summarize the different aspects contained in enhancing reproducibility, reusability, and generalizability in computational pathology.

Contributions. Together with Christian Matek, I compiled the suggestions out of the analysis and checklists compiled in P8..

4. Discussion

Learning good representations for whole slide images (WSIs) is challenging due to the images' gigapixel size and the limited availability of clinical data and annotations. This is the case in routine cancer diagnostics and even more so for rare diseases. The overall goal of computational pathology algorithms is to help pathologists in their clinical workflow by accelerating diagnostics for routine cases, automating tedious manual tasks such as cell counting, detecting rare cases, and supporting the overall decision-making process [44].

This thesis has dealt with the technical use case of clinical endpoint prediction of morphological, molecular, or clinical markers. Typically, pathologists assess such markers with the help of additional tests such as molecular profiling or immunohistochemistry or are collected from a patient's clinical history. Linking the histopathology image to clinical outcomes or treatment response prediction is the ultimate goal of pathological workflows whether the relevant information was obtained manually or in an automated way using AI. Ultimately, better AI models for pathological workflows will improve clinical applications and potentially new diagnostic insights [134].

As a first objective, batch effects and the resulting low generalization of DL models were tackled. We proposed a generative method to model different visual appearances from multiple clinics to introduce realistic staining variations during training even when the training data was only from a single data source. Our model HistAuGAN was the first generative network that implemented stain augmentation instead of stain normalization. HistAuGAN introduced more realistic color augmentations designed for histopathology images making downstream models more robust as proven by our downstream analysis for lymph node metastasis detection. However, the method is computationally less efficient than conventional color augmentations. Finding realistic H&E-specific augmentations that can be used for large-scale self-supervised training for patch-level foundation models in histopathology remains an open research question. Currently, this question is only addressed by training the models on as much and diverse data as possible to introduce enough variance during training.

Next, the research presented in this thesis aimed to improve the current state of representation learning for patch encoders in histopathology. Frameworks for improving self-supervised learning recipes by including available annotations with supervised contrastive losses and semi-supervised approaches and resource-efficient finetuning of

existing vision foundation models were proposed to overcome current limitations in resource-heavy foundation model training. Further, we collected the most extensive collection of publicly available datasets to build the first foundation model for hematology. These works were solely based on vision, i.e., histology or hematology microscopic data. In contrast, most recent works show that vision-based representations can be further improved using multimodal objectives including data from other modalities such as image captions or genomics [97].

Using the patch embeddings, feature aggregation networks map WSI inputs to clinical endpoints. In the scope of this thesis, a biology-inspired Transformer-based aggregation with local attention linking every patch only to other patches in a neighborhood was proposed for multi-task genetic alteration prediction on colorectal cancer histology. Similarly, a fully-Transformer-based pipeline including a SwinTransformer as patch encoder and Transformer-based aggregation was evaluated on a large-scale multicentric patient cohort of over 13,000 patients for the three biomarkers commonly tested for in colorectal cancer as recommended by the World Health Organization. This evaluation demonstrated the clinical utility of deep learning workflows, especially by proving for the first time that such algorithms could generalize from resections to biopsies. Our model therefore allows a much earlier application in the clinical workflow than models only working on resections. While recent developments of patch-level foundation models have proven that weakly-supervised learning pipelines reach unprecedented performances, no end-to-end training is currently possible: Patch encoder and feature aggregation networks are still trained separately.

Finally, our analysis has shown that the deep learning methods in computational pathology lack reproducibility and reusability. This holds especially true for code and data sharing but many studies also lack statistical analysis of results and external validations. With the recent developments of patch-level foundation models, we observe that the training code base is not shared as many corporate interests entered the field. While this impacts research around computational pathology, model weights are still shared and the documentation on how to use them improved, hinting towards a better clinical implementation in the future.

The thesis aimed to advance computer-aided cancer diagnosis by building generalizable and data-efficient DL models. With AI developments progressing at a fast pace together with an increasing availability of digitized data in the clinical workflow, representation learning for histopathology slides will be brought to a new level.

Outlook

The ultimate goal of pathology is to find the best description of a patient's disease in terms of the disease's biology and progression status to determine the most adequate diagnosis. In analogy, computational pathology aims to improve pathology by accelerating existing workflows and making diagnoses more accurate. In Chapter 1, we learned that pathologists seek to find the best and most efficient representations of disease in the diagnostic workflow, especially in the case of cancer. Rosai reflects on the achievements accomplished by the interplay of the fields of molecular biology and surgical pathology:

“Actually, if one thinks about the major advances that have been made in the field in recent years, one realizes that most of them [...] have occurred as a result of the symbiosis of pathologists and molecular biologists, rather than each working in isolation. [...] Prediction of therapeutic response by molecular profiling is the logical and natural extension of the work of the surgical pathologist. [...] it needs to be said that a key requirement for integration is a close cooperation among surgical pathologists and molecular biologists.” [124]

Hence, Rosai advocates for interdisciplinary collaboration between molecular biologists and pathologists to find the best disease representations. Similarly, computational pathology should aim to integrate molecular information into the currently morphology-based diagnostic workflow, thus enhancing the gold standard of microscopic examination.

Recent work in computational has already demonstrated that combining multiple data modalities such as clinical reports [98, 2] or transcriptomics [75] improve whole-slide image representations. Based on these findings, I am convinced that AI can help bridge the phenotype-genotype gap by integrating both molecular and morphological information into model training, ultimately providing better-informed representations that advance clinical diagnostic workflows. This holds especially for molecular information but additional data such as clinical reports or other imaging modalities also provide further knowledge.

To pursue this, new models need to be trained on a broader range of data requiring structured data access across departments. For such an endeavor, data sharing and interdisciplinary and international research collaborations are important to pave the way for successful clinical implementation.

List of Figures

1.1. Use cases of deep learning in pathology. After digitization of the whole slide image, computational methods can be divided into data pre-processing and image analysis methods. Among these (a) stain normalization, (b) evaluation of cell-level features, (c) tissue type segmentation, (d) clinical endpoint and biomarker prediction, (e) case retrieval, and (f) report generation are common tasks in the downstream workflow of computational pathology.	2
1.2. Computational workflow for clinical endpoint prediction. Challenges in the computational workflow are detailed including the contributions of this dissertation-based thesis (core publications are marked by *). . .	5
2.1. Pre-processing and stain color adaptation. (a) Stain normalization methods map test images to the training domain. (b) Stain augmentation methods create diverse synthetic data augmentations to introduce stain variance during model training. (c) Exemplary GAN architecture for style transfer with unpaired image-to-image translation using disentangled representations (DRIT++ [93]).	10
2.2. Representation learning for encoding histology patches. Initial self-supervised learning objectives originated from metric learning objectives such as the triplet loss (a) and introduced contrastive learning (b). Current successful recipes involve mostly knowledge distillation (c) or multimodal alignment using image-caption pairs (d).	15
2.3. Weakly-supervised learning for histopathology. Architectures for feature aggregation networks can be divided into (a) attention-based networks, (b) graph-based networks, (c) multi-scale networks, and (d) Transformer-based networks.	21
3.1. Overview of publications included in this thesis, positioning the publications in the context of the computational workflow for clinical endpoint prediction in computer-aided cancer diagnosis.	26

List of Tables

2.1. Publicly available datasets with labels for clinical endpoint prediction. Selection of tasks that can be used for clinical endpoint prediction, which can be categorized into molecular, morphological, and clinical labels.	9
2.2. Overview of existing patch encoders sorted by their release date. The following information was collected for every model if the information was available: information about the training datasets including the number of WSIs (patients) and the number of patches used during training, methodological details such as training framework, backbone architecture, and whether the latent space is trained with vision-language (VL) alignment, finally, the data (D), training code (C), and model (M) availability is noted.	19

Bibliography

- [1] A. Acevedo, S. Alf  rez, A. Merino, L. Puigvi, and J. Rodellar. “Recognition of peripheral blood cell images using convolutional neural networks.” In: *Computer methods and programs in biomedicine* 180 (2019), p. 105020 (cit. on p. 31).
- [2] F. Ahmed, A. Sellergren, L. Yang, S. Xu, B. Babenko, A. Ward, N. Olson, A. Moltashamian, Y. Matias, G. S. Corrado, et al. “Pathalign: A vision-language model for whole slide images in histopathology.” In: *arXiv preprint arXiv:2406.19578* (2024) (cit. on pp. 25, 38).
- [3] S. Azizi, L. Culp, J. Freyberg, B. Mustafa, S. Baur, S. Kornblith, T. Chen, N. Tomasev, J. Mitrovi  , P. Strachan, et al. “Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging.” In: *Nature Biomedical Engineering* 7.6 (2023), pp. 756–779 (cit. on p. 19).
- [4] M. Baker. “Reproducibility crisis.” In: *nature* 533.26 (2016), pp. 353–66 (cit. on pp. 7, 34).
- [5] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A. G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum. *A Cookbook of Self-Supervised Learning*. arXiv:2304.12210 [cs]. June 2023. doi: 10.48550/arXiv.2304.12210 (cit. on pp. 13, 14).
- [6] P. B  ndi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. Ehteshami Bejnordi, B. Lee, K. Paeng, A. Zhong, Q. Li, F. G. Zanjani, S. Zinger, K. Fukuta, D. Komura, V. Ovtcharov, S. Cheng, S. Zeng, J. Thagaard, A. B. Dahl, H. Lin, H. Chen, L. Jacobsson, M. Hedlund, M.   etin, E. Halıcı, H. Jackson, R. Chen, F. Both, J. Franke, H. K  sters-Vandeveld, W. Vreuls, P. Bult, B. van Ginneken, J. van der Laak, and G. Litjens. “From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge.” In: *IEEE Transactions on Medical Imaging* 38.2 (2019), pp. 550–560. doi: 10.1109/TMI.2018.2867350 (cit. on pp. 8, 27).
- [7] Y. Bengio, A. Courville, and P. Vincent. “Representation learning: A review and new perspectives.” In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828 (cit. on pp. 13, 14).

- [8] A. Bentaieb and G. Hamarneh. "Adversarial Stain Transfer for Histopathology Image Analysis." In: *IEEE Transactions on Medical Imaging* 37.3 (2018), pp. 792–802. doi: 10.1109/TMI.2017.2781228 (cit. on p. 10).
- [9] N. Brancati, A. M. Anniciello, P. Pati, D. Riccio, G. Scognamiglio, G. Jaume, G. De Pietro, M. Di Bonito, A. Foncubierta, G. Botti, et al. "Bracs: A dataset for breast carcinoma subtyping in h&e histology images." In: *Database* 2022 (2022), baac093 (cit. on p. 9).
- [10] W. Bulten, K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. Van Boven, R. Vink, et al. "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge." In: *Nature medicine* 28.1 (2022), pp. 154–163 (cit. on p. 9).
- [11] G. Campanella, S. Chen, R. Verma, J. Zeng, A. Stock, M. Croken, B. Veremis, A. Elmas, K.-l. Huang, R. Kwan, et al. "A clinical benchmark of public self-supervised pathology foundation models." In: *arXiv preprint arXiv:2407.06508* (2024) (cit. on p. 6).
- [12] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images." In: *Nature medicine* 25.8 (2019), pp. 1301–1309 (cit. on p. 4).
- [13] G. Campanella, R. Kwan, E. Fluder, J. Zeng, A. Stock, B. Veremis, A. D. Polydorides, C. Hedvat, A. Schoenfeld, C. Vanderbilt, et al. "Computational Pathology at Health System Scale—Self-Supervised Foundation Models from Three Billion Images." In: *arXiv preprint arXiv:2310.07033* (2023) (cit. on p. 19).
- [14] G. Campanella, A. R. Rajanna, L. Corsale, P. J. Schüffler, Y. Yagi, and T. J. Fuchs. "Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology." In: *Computerized medical imaging and graphics* 65 (2018), pp. 142–151 (cit. on p. 3).
- [15] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. "Deep clustering for unsupervised learning of visual features." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 132–149 (cit. on p. 14).
- [16] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. "Unsupervised learning of visual features by contrasting cluster assignments." In: *Advances in neural information processing systems* 33 (2020), pp. 9912–9924 (cit. on p. 14).

- [17] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. “Emerging properties in self-supervised vision transformers.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660 (cit. on pp. 5, 17, 19).
- [18] C. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, A. J. Schaumberg, and F. Mahmood. “Fast and scalable search of whole-slide images via self-supervised deep learning.” In: *Nature Biomedical Engineering* 6.12 (2022), pp. 1420–1434 (cit. on p. 4).
- [19] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood. “Scaling vision transformers to gigapixel images via hierarchical self-supervised learning.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16144–16155 (cit. on pp. 24, 25).
- [20] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, et al. “Towards a general-purpose foundation model for computational pathology.” In: *Nature Medicine* 30.3 (2024), pp. 850–862 (cit. on pp. 5, 13, 19).
- [21] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. “A simple framework for contrastive learning of visual representations.” In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607 (cit. on p. 13).
- [22] X. Chen and K. He. “Exploring simple siamese representation learning.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 15750–15758 (cit. on p. 17).
- [23] X. Chen, S. Xie, and K. He. “An empirical study of training self-supervised vision transformers.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9640–9649 (cit. on p. 19).
- [24] Z. Chen, T. Xie, S. Chen, Z. Li, S. Yao, X. Lu, W. He, C. Tang, D. Yang, S. Li, et al. “AI-based tumor-infiltrating lymphocytes scoring system for assessing HCC prognosis in patients undergoing liver resection.” In: *JHEP Reports* (2024), p. 101270 (cit. on p. 3).
- [25] D. Cifci, S. Foersch, and J. N. Kather. “Artificial intelligence to identify genetic alterations in conventional histopathology.” In: *The Journal of Pathology* 257.4 (2022), pp. 430–444 (cit. on p. 4).
- [26] O. Ciga, T. Xu, and A. L. Martel. “Self supervised contrastive learning for digital histopathology.” In: *Machine Learning with Applications* 7 (2022), p. 100198 (cit. on p. 19).

- [27] J. P. Cohen, M. Luck, and S. Honari. "Distribution matching losses can hallucinate features in medical image translation." In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer. 2018, pp. 529–536 (cit. on p. 12).
- [28] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyő, A. L. Moreira, N. Razavian, and A. Tsirigos. "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning." In: *Nature medicine* 24.10 (2018), pp. 1559–1567 (cit. on p. 2).
- [29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database." In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on pp. 3, 13).
- [30] J. Devlin. "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv preprint arXiv:1810.04805* (2018) (cit. on pp. 14, 17).
- [31] J. Ding, S. Ma, L. Dong, X. Zhang, S. Huang, W. Wang, N. Zheng, and F. Wei. "Longnet: Scaling transformers to 1,000,000,000 tokens." In: *arXiv preprint arXiv:2307.02486* (2023) (cit. on p. 24).
- [32] T. Ding, S. J. Wagner, A. H. Song, R. J. Chen, M. Y. Lu, A. Zhang, A. J. Vaidya, G. Jaume, M. Shaban, A. Kim, et al. "Multimodal Whole Slide Foundation Model for Pathology." In: *arXiv preprint arXiv:2411.19666* (2024) (cit. on pp. 4, 25).
- [33] J. Dippel, B. Feulner, T. Winterhoff, T. Milbich, S. Tietz, S. Schallenberg, G. Dernbach, A. Kunft, S. Heinke, M.-L. Eich, et al. "RudolfV: a foundation model by pathologists for pathologists." In: *arXiv preprint arXiv:2401.04079* (2024) (cit. on p. 19).
- [34] C. Doersch, A. Gupta, and A. A. Efros. "Unsupervised visual representation learning by context prediction." In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1422–1430 (cit. on p. 14).
- [35] A. Echle, H. I. Grabsch, P. Quirke, P. A. van den Brandt, N. P. West, G. G. Hutchins, L. R. Heij, X. Tan, S. D. Richman, J. Krause, et al. "Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning." In: *Gastroenterology* 159.4 (2020), pp. 1406–1416 (cit. on pp. 4, 13, 33).
- [36] N. J. Edwards, M. Oberti, R. R. Thangudu, S. Cai, P. B. McGarvey, S. Jacob, S. Madhavan, and K. A. Ketchum. "The CPTAC data portal: a resource for cancer proteomics research." In: *Journal of proteome research* 14.6 (2015), pp. 2707–2713 (cit. on pp. 8, 9, 32).

- [37] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. “Dermatologist-level classification of skin cancer with deep neural networks.” In: *nature* 542.7639 (2017), pp. 115–118 (cit. on p. 2).
- [38] J. Fan, T. Lv, Y. Di, L. Li, and X. Pan. “PathMamba: Weakly Supervised State Space Model for Multi-class Segmentation of Pathology Images.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 500–509 (cit. on p. 24).
- [39] K. Faryna, J. van der Laak, and G. Litjens. “Automatic data augmentation to improve generalization of deep learning in H&E stained histopathology.” In: *Computers in Biology and Medicine* 170 (2024), p. 108018 (cit. on p. 5).
- [40] A. Filiot, R. Ghermi, A. Olivier, P. Jacob, L. Fidon, A. M. Kain, C. Saillard, and J.-B. Schiratti. “Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling.” In: *medRxiv* (2023). doi: 10.1101/2023.07.21.23292757. eprint: <https://www.medrxiv.org/content/early/2023/07/26/2023.07.21.23292757.full.pdf> (cit. on pp. 5, 19).
- [41] A. Filiot, P. Jacob, A. Mac Kain, and C. Saillard. “Phikon-v2, A large and public feature extractor for biomarker prediction.” In: *arXiv preprint arXiv:2409.09173* (2024) (cit. on p. 19).
- [42] L. Fillioux, J. Boyd, M. Vakalopoulou, P.-H. Cournède, and S. Christodoulidis. “Structured state space models for multiple instance learning in digital pathology.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 594–604 (cit. on p. 24).
- [43] Y. Fu, A. W. Jung, R. V. Torne, S. Gonzalez, H. Vöhringer, A. Shmatko, L. R. Yates, M. Jimenez-Linan, L. Moore, and M. Gerstung. “Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis.” In: *Nature cancer* 1.8 (2020), pp. 800–810 (cit. on pp. 4, 33).
- [44] T. J. Fuchs and J. M. Buhmann. “Computational pathology: challenges and promises for tissue analysis.” In: *Computerized Medical Imaging and Graphics* 35.7-8 (2011), pp. 515–530 (cit. on p. 36).
- [45] J. Gao, Z. Shen, Z. Deng, and L. Mei. “Impact of tumor–stroma ratio on the prognosis of colorectal cancer: a systematic review.” In: *Frontiers in oncology* 11 (2021), p. 738080 (cit. on p. 3).
- [46] S. Gidaris, P. Singh, and N. Komodakis. “Unsupervised representation learning by predicting image rotations.” In: *arXiv preprint arXiv:1803.07728* (2018) (cit. on p. 14).

- [47] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. “Imagebind: One embedding space to bind them all.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15180–15190 (cit. on pp. 5, 17, 30).
- [48] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial networks.” In: *Communications of the ACM* 63.11 (2020), pp. 139–144 (cit. on p. 11).
- [49] A. Goodman, A. Carpenter, E. Park, jlefman-nvidia, Josette_{Booz}Allen, Kyle, Maggie, Nilofer, P. Sedivec, and W. Cukierski. *2018 Data Science Bowl*. <https://kaggle.com/competitions/data-science-bowl-2018>. Kaggle. 2018 (cit. on p. 28).
- [50] S. Graham, Q. D. Vu, M. Jahanifar, S. E. A. Raza, F. Minhas, D. Snead, and N. Rajpoot. “One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification.” In: *Medical Image Analysis* 83 (2023), p. 102685 (cit. on p. 22).
- [51] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. “Bootstrap your own latent-a new approach to self-supervised learning.” In: *Advances in neural information processing systems* 33 (2020), pp. 21271–21284 (cit. on p. 16).
- [52] A. Gu and T. Dao. “Mamba: Linear-time sequence modeling with selective state spaces.” In: *arXiv preprint arXiv:2312.00752* (2023) (cit. on p. 24).
- [53] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré. “Hippo: Recurrent memory with optimal polynomial projections.” In: *Advances in neural information processing systems* 33 (2020), pp. 1474–1487 (cit. on p. 24).
- [54] A. Gu, K. Goel, and C. Ré. “Efficiently Modeling Long Sequences with Structured State Spaces.” In: *The International Conference on Learning Representations (ICLR)*. 2022 (cit. on p. 24).
- [55] Z. Guo, W. Zhao, S. Wang, and L. Yu. “HIGT: Hierarchical Interaction Graph-Transformer for Whole Slide Image Analysis.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 755–764 (cit. on p. 22).
- [56] I. S. Hagemann. “Molecular testing in breast cancer: a guide to current practices.” In: *Archives of Pathology & Laboratory Medicine* 140.8 (2016), pp. 815–824 (cit. on p. 1).

- [57] M. G. Hanna, V. E. Reuter, M. R. Hameed, L. K. Tan, S. Chiang, C. Sigel, T. Hollmann, D. Giri, J. Samboy, C. Moradel, et al. "Whole slide imaging equivalency and efficiency study: experience at a large academic center." In: *Modern Pathology* 32.7 (2019), pp. 916–928 (cit. on p. 2).
- [58] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. "Masked autoencoders are scalable vision learners." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009 (cit. on p. 14, 19).
- [59] K. He, G. Gkioxari, P. Dollár, and R. Girshick. "Mask r-cnn." In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969 (cit. on p. 28).
- [60] M. Hehr, A. Sadafi, C. Matek, P. Lienemann, C. Pohlkamp, T. Haferlach, K. Spiekermann, and C. Marr. "Explainable AI identifies diagnostic cells of genetic AML subtypes." In: *PLOS Digital Health* 2.3 (2023), e0000187 (cit. on p. 31).
- [61] S. Hendry, R. Salgado, T. Gevaert, P. A. Russell, T. John, B. Thapa, M. Christie, K. Van De Vijver, M. V. Estrada, P. I. Gonzalez-Ericsson, et al. "Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group: Part 2: TILs in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors." In: *Advances in anatomic pathology* 24.6 (2017), pp. 311–335 (cit. on p. 3).
- [62] D.-T. Hoang, E. D. Shulman, R. Turakulov, Z. Abdullaev, O. Singh, E. M. Campagnolo, H. Lalchungnunga, E. A. Stone, M. P. Nasrallah, E. Ruppin, et al. "Prediction of DNA methylation-based tumor types from histopathology in central nervous system tumors with deep learning." In: *Nature Medicine* (2024), pp. 1–10 (cit. on p. 9).
- [63] R. Hollandi, A. Szkalisity, T. Toth, E. Tasnadi, C. Molnar, B. Mathe, I. Grexa, J. Molnar, A. Balind, M. Gorbe, et al. "nucleAIzer: a parameter-free deep learning framework for nucleus segmentation using image style transfer." In: *Cell systems* 10.5 (2020), pp. 453–458 (cit. on p. 28).
- [64] W. Hou, H. Huang, Q. Peng, R. Yu, L. Yu, and L. Wang. "Spatial-hierarchical graph neural network with dynamic structure learning for histological image classification." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 181–191 (cit. on p. 22).

- [65] W. Hou, L. Yu, C. Lin, H. Huang, R. Yu, J. Qin, and L. Wang. "H²-MIL: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 36. 1. 2022, pp. 933–941 (cit. on p. 22).
- [66] X. Hou, C. Jiang, A. Kondepudi, Y. Lyu, A. Chowdury, H. Lee, and T. C. Hollon. "A self-supervised framework for learning whole slide representations." In: *arXiv preprint arXiv:2402.06188* (2024) (cit. on pp. 24, 25).
- [67] F. M. Howard, J. Dolezal, S. Kochanny, J. Schulte, H. Chen, L. Heij, D. Huo, R. Nanda, O. I. Olopade, J. N. Kather, et al. "The impact of digital histopathology batch effect on deep learning model accuracy and bias." In: *BioRxiv* (2020), pp. 2020–12 (cit. on pp. 8, 9).
- [68] Z. Huang, F. Bianchi, M. Yuksekogonul, T. J. Montine, and J. Zou. "A visual-language foundation model for pathology image analysis using medical twitter." In: *Nature medicine* 29.9 (2023), pp. 2307–2316 (cit. on pp. 4, 19, 20).
- [69] M. Hutson. *Artificial intelligence faces reproducibility crisis*. 2018 (cit. on p. 34).
- [70] W. Ikezogwo, S. Seyfioglu, F. Ghezloo, D. Geva, F. Sheikh Mohammed, P. K. Anand, R. Krishna, and L. Shapiro. "Quilt-1m: One million image-text pairs for histopathology." In: *Advances in neural information processing systems* 36 (2024) (cit. on pp. 4, 19, 20).
- [71] M. Ilse, J. Tomczak, and M. Welling. "Attention-based deep multiple instance learning." In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136 (cit. on pp. 6, 21, 32).
- [72] E. Imyanitov and E. Kuligina. "Molecular testing for colorectal cancer: Clinical applications." In: *World Journal of Gastrointestinal Oncology* 13.10 (2021), p. 1288 (cit. on p. 1).
- [73] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks." In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. 2017 (cit. on p. 11).
- [74] A. Janowczyk, A. Basavanahally, and A. Madabhushi. "Stain normalization using sparse autoencoders (StaNoSA): application to digital pathology." In: *Computerized Medical Imaging and Graphics* 57 (2017), pp. 50–61 (cit. on p. 10).
- [75] G. Jaume, L. Oldenburg, A. Vaidya, R. J. Chen, D. F. Williamson, T. Peeters, A. H. Song, and F. Mahmood. "Transcriptomics-guided slide representation learning in computational pathology." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9632–9644 (cit. on pp. 20, 22, 25, 38).

- [76] G. Jaume, A. Vaidya, A. Zhang, A. H Song, R. J Chen, S. Sahai, D. Mo, E. Madrigal, L. Phi Le, and F. Mahmood. "Multistain Pretraining for Slide Representation Learning in Pathology." In: *European Conference on Computer Vision*. Springer. 2024, pp. 19–37 (cit. on pp. 20, 25).
- [77] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. "Scaling up visual and vision-language representation learning with noisy text supervision." In: *International conference on machine learning*. PMLR. 2021, pp. 4904–4916 (cit. on p. 17).
- [78] D. Juyal, H. Padigela, C. Shah, D. Shenker, N. Harguindeguy, Y. Liu, B. Martin, Y. Zhang, M. Nercessian, M. Markey, et al. "PLUTO: Pathology-Universal Transformer." In: *arXiv preprint arXiv:2405.07905* (2024) (cit. on p. 19).
- [79] M. Kang, H. Song, S. Park, D. Yoo, and S. Pereira. "Benchmarking Self-Supervised Learning on Diverse Pathology Datasets." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 3344–3354 (cit. on p. 19).
- [80] S. Kapoor and A. Narayanan. "Leakage and the reproducibility crisis in machine-learning-based science." In: *Patterns* 4.9 (2023) (cit. on p. 34).
- [81] J. N. Kather, N. Halama, and A. Marx. *100,000 histological images of human colorectal cancer and healthy tissue*. Apr. 2018. DOI: 10.5281/zenodo.1214456 (cit. on pp. 8, 29).
- [82] J. N. Kather, L. R. Heij, H. I. Grabsch, C. Loeffler, A. Echle, H. S. Muti, J. Krause, J. M. Niehues, K. A. Sommer, P. Bankhead, et al. "Pan-cancer image-based detection of clinically actionable genetic alterations." In: *Nature cancer* 1.8 (2020), pp. 789–799 (cit. on p. 33).
- [83] J. N. Kather, A. T. Pearson, N. Halama, D. Jäger, J. Krause, S. H. Loosen, A. Marx, P. Boor, F. Tacke, U. P. Neumann, et al. "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer." In: *Nature medicine* 25.7 (2019), pp. 1054–1056 (cit. on pp. 4, 33).
- [84] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee. "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution." In: *IEEE transactions on Biomedical Engineering* 61.6 (2014), pp. 1729–1738 (cit. on p. 10).
- [85] Y. J. Kim, H. Jang, K. Lee, S. Park, S.-G. Min, C. Hong, J. H. Park, K. Lee, J. Kim, W. Hong, et al. "PAIP 2019: Liver cancer segmentation challenge." In: *Medical image analysis* 67 (2021), p. 101854 (cit. on p. 9).

- [86] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. "Segment anything." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026 (cit. on p. 30).
- [87] A. Kondepudi, M. Pekmezci, X. Hou, K. Scotford, C. Jiang, A. Rao, E. S. Harake, A. Chowdury, W. Al-Holou, L. Wang, et al. "Foundation models for fast, label-free detection of glioma infiltration." In: *Nature* (2024), pp. 1–7 (cit. on p. 24).
- [88] R. Krishnan, P. Rajpurkar, and E. J. Topol. "Self-supervised learning in medicine and healthcare." In: *Nature Biomedical Engineering* 6.12 (2022), pp. 1346–1352 (cit. on p. 18).
- [89] E. M. de Kruijf, J. G. van Nes, C. J. van de Velde, H. Putter, V. T. Smit, G. J. Liefers, P. J. Kuppen, R. A. Tollenaar, and W. E. Mesker. "Tumor–stroma ratio in the primary tumor is a prognostic factor in early breast cancer patients, especially in triple-negative carcinoma patients." In: *Breast cancer research and treatment* 125 (2011), pp. 687–696 (cit. on p. 3).
- [90] V. L'Imperio, E. Wulczyn, M. Plass, H. Müller, N. Tamini, L. Gianotti, N. Zucchini, R. Reihs, G. S. Corrado, D. R. Webster, et al. "Pathologist validation of a machine learning–derived feature for colon cancer risk stratification." In: *JAMA Network Open* 6.3 (2023), e2254891–e2254891 (cit. on p. 4).
- [91] G. Larsson, M. Maire, and G. Shakhnarovich. "Learning representations for automatic colorization." In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV* 14. Springer. 2016, pp. 577–593 (cit. on p. 14).
- [92] T. Lazard, M. Lerousseau, E. Decencière, and T. Walter. "Giga-ssl: Self-supervised learning for gigapixel images." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 4305–4314 (cit. on pp. 24, 25).
- [93] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang. "Diverse Image-to-Image Translation via Disentangled Representations." In: *European Conference on Computer Vision*. 2018 (cit. on pp. 10, 12, 27).
- [94] N. I. Lindeman, P. T. Cagle, M. B. Beasley, D. A. Chitale, S. Dacic, G. Giaccone, R. B. Jenkins, D. J. Kwiatkowski, J.-S. Saldivar, J. Squire, et al. "Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology." In: *Journal of Thoracic Oncology* 8.7 (2013), pp. 823–859 (cit. on p. 1).

- [95] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. Van de Loo, R. Vogels, et al. "1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset." In: *GigaScience* 7.6 (2018), giy065 (cit. on p. 9).
- [96] M.-Y. Liu, T. Breuel, and J. Kautz. "Unsupervised image-to-image translation networks." In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 11).
- [97] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, et al. "A visual-language foundation model for computational pathology." In: *Nature Medicine* 30.3 (2024), pp. 863–874 (cit. on pp. 4, 5, 19, 20, 37).
- [98] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, M. Zhao, A. K. Chow, K. Ikemura, A. Kim, D. Pouli, A. Patel, et al. "A multimodal generative AI copilot for human pathology." In: *Nature* 634.8033 (2024), pp. 466–473 (cit. on pp. 4, 38).
- [99] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood. "Data-efficient and weakly supervised computational pathology on whole-slide images." In: *Nature biomedical engineering* 5.6 (2021), pp. 555–570 (cit. on pp. 4, 6, 22, 32).
- [100] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas. "A method for normalizing histology slides for quantitative analysis." In: *2009 IEEE international symposium on biomedical imaging: from nano to macro*. IEEE. 2009, pp. 1107–1110 (cit. on pp. 5, 10).
- [101] N. Marini, S. Otalora, M. Wodzinski, S. Tomassini, A. F. Dragoni, S. Marchand-Maillet, J. P. D. Morales, L. Duran-Lopez, S. Vatrano, H. Müller, et al. "Data-driven color augmentation for H&E stained images in computational pathology." In: *Journal of Pathology Informatics* 14 (2023), p. 100183 (cit. on p. 5).
- [102] B. Märkl, L. Füzesi, R. Huss, S. Bauer, and T. Schaller. "Number of pathologists in Germany: comparison with European countries, USA, and Canada." In: *Virchows Archiv* 478 (2021), pp. 335–341 (cit. on p. 2).
- [103] C. Matek, S. Krappe, C. Münzenmayer, T. Haferlach, and C. Marr. "Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set." In: *Blood, The Journal of the American Society of Hematology* 138.20 (2021), pp. 1917–1927 (cit. on p. 31).
- [104] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr. "Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks." In: *Nature Machine Intelligence* 1.11 (2019), pp. 538–544 (cit. on pp. 2, 29, 31).

- [105] T. Mathew, J. R. Kini, and J. Rajan. "Computational methods for automated mitosis detection in histopathology images: A review." In: *Biocybernetics and Biomedical Engineering* 41.1 (2021), pp. 64–82 (cit. on p. 3).
- [106] W. E. Mesker, J. M. Junggeburdt, K. Szuhai, P. de Heer, H. Morreau, H. J. Tanke, and R. A. Tollenaar. "The carcinoma–stromal ratio of colon carcinoma is an independent factor for survival compared to lymph node status and tumor stage." In: *Analytical Cellular Pathology* 29.5 (2007), pp. 387–398 (cit. on p. 3).
- [107] D. M. Metter, T. J. Colgan, S. T. Leung, C. F. Timmons, and J. Y. Park. "Trends in the US and Canadian pathologist workforces from 2007 to 2017." In: *JAMA network open* 2.5 (2019), e194337–e194337 (cit. on p. 2).
- [108] N. Mu, A. Kirillov, D. Wagner, and S. Xie. "Slip: Self-supervision meets language-image pre-training." In: *European conference on computer vision*. Springer. 2022, pp. 529–544 (cit. on p. 18).
- [109] D. Nechaev, A. Pchelnikov, and E. Ivanova. "Hibou: A Family of Foundational Vision Transformers for Pathology." In: *arXiv preprint arXiv:2406.05074* (2024) (cit. on p. 19).
- [110] P. Neidlinger, O. S. El Nahhas, H. S. Muti, T. Lenz, M. Hoffmeister, H. Brenner, M. van Treeck, R. Langer, B. Dislich, H. M. Behrens, et al. "Benchmarking foundation models as feature extractors for weakly-supervised computational pathology." In: *arXiv preprint arXiv:2408.15823* (2024) (cit. on p. 6).
- [111] P. C. Neto, D. Montezuma, S. P. Oliveira, D. Oliveira, J. Fraga, A. Monteiro, L. Ribeiro, S. Gonçalves, S. Reinhard, I. Zlobec, I. M. Pinto, and J. S. Cardoso. *IMP Whole-Slide Images of Colorectal Samples 2024 [Data set]*. 2024. DOI: 10.25747/FB1Q-J507 (cit. on p. 9).
- [112] J. M. Niehues, G. Müller-Franzes, Y. Schirris, S. J. Wagner, M. Jendrusch, M. Kloor, A. T. Pearson, H. S. Muti, K. J. Hewitt, G. P. Veldhuizen, et al. "Using histopathology latent diffusion models as privacy-preserving dataset augmenters improves downstream classification performance." In: *Computers in Biology and Medicine* 175 (2024), p. 108410 (cit. on p. 12).
- [113] J. M. Niehues, P. Quirke, N. P. West, H. I. Grabsch, M. van Treeck, Y. Schirris, G. P. Veldhuizen, G. G. Hutchins, S. D. Richman, S. Foersch, et al. "Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: A retrospective multi-centric study." In: *Cell reports Medicine* 4.4 (2023) (cit. on p. 33).

- [114] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash. “Boosting self-supervised learning via knowledge transfer.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9359–9367 (cit. on p. 14).
- [115] S. W. Omholt. “Bridging the Genotype-Phenotype Gap.” In: *Science* 295.5563 (2002), pp. 2220–2220. DOI: 10.1126/science.1070267. eprint: <https://www.science.org/doi/pdf/10.1126/science.1070267> (cit. on p. 1).
- [116] A. v. d. Oord, Y. Li, and O. Vinyals. “Representation learning with contrastive predictive coding.” In: *arXiv preprint arXiv:1807.03748* (2018) (cit. on p. 15).
- [117] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. “Dinov2: Learning robust visual features without supervision.” In: *arXiv preprint arXiv:2304.07193* (2023) (cit. on pp. 5, 17, 19, 25, 30, 31).
- [118] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. “Context encoders: Feature learning by inpainting.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544 (cit. on p. 14).
- [119] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. “Learning transferable visual models from natural language supervision.” In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763 (cit. on pp. 17, 20).
- [120] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. “Color transfer between images.” In: *IEEE Computer graphics and applications* 21.5 (2001), pp. 34–41 (cit. on pp. 5, 10).
- [121] D. Reisenbüchler, S. J. Wagner, M. Boxberg, and T. Peng. “Local attention graph-based transformer for multi-target genetic alteration prediction.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 377–386 (cit. on pp. 24, 33).
- [122] A. Riasatian, M. Babaie, D. Maleki, S. Kalra, M. Valipour, S. Hemati, M. Zaveri, A. Safarpour, S. Shafiei, M. Afshari, et al. “Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides.” In: *Medical image analysis* 70 (2021), p. 102032 (cit. on pp. 18, 19).
- [123] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation.” In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III* 18. Springer. 2015, pp. 234–241 (cit. on p. 28).

- [124] J. Rosai. “Why microscopy will remain a cornerstone of surgical pathology.” In: *Laboratory investigation* 87.5 (2007), pp. 403–408 (cit. on pp. 1, 38).
- [125] C. Saillard, R. Jenatton, F. Llinares-López, Z. Mariet, D. Cahané, E. Durand, and J.-P. Vert. *H-optimus-0*. 2024 (cit. on pp. 5, 19).
- [126] O. L. Saldanha, C. M. Loeffler, J. M. Niehues, M. van Treeck, T. P. Seraphin, K. J. Hewitt, D. Cifci, G. P. Veldhuizen, S. Ramesh, A. T. Pearson, et al. “Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology.” In: *NPJ Precision Oncology* 7.1 (2023), p. 35 (cit. on p. 4).
- [127] R. Schaer, S. Otálora, O. Jimenez-del-Toro, M. Atzori, and H. Müller. “Deep learning-based retrieval system for gigapixel histopathology cases and the open access literature.” In: *Journal of pathology informatics* 10.1 (2019), p. 19 (cit. on p. 4).
- [128] Y. Schirris, E. Gavves, I. Nederlof, H. M. Horlings, and J. Teuwen. “DeepSMILE: Contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer.” In: *Medical Image Analysis* 79 (2022), p. 102464 (cit. on pp. 6, 32).
- [129] P. J. Schöffler, T. J. Fuchs, C. S. Ong, P. J. Wild, N. J. Rupp, and J. M. Buhmann. “TMARKER: A free software toolkit for histopathological cell counting and staining estimation.” In: *Journal of pathology informatics* 4.2 (2013), p. 2 (cit. on p. 3).
- [130] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni. “Staingan: Stain style transfer for digital histological images.” In: *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*. IEEE. 2019, pp. 953–956 (cit. on p. 27).
- [131] G. Shaikovski, A. Casson, K. Severson, E. Zimmermann, Y. K. Wang, J. D. Kunz, J. A. Retamero, G. Oakley, D. Klimstra, C. Kanan, et al. “PRISM: A Multi-Modal Generative Foundation Model for Slide-Level Histopathology.” In: *arXiv preprint arXiv:2405.10254* (2024) (cit. on pp. 4, 25).
- [132] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al. “Transmil: Transformer based correlated multiple instance learning for whole slide image classification.” In: *Advances in neural information processing systems* 34 (2021), pp. 2136–2147 (cit. on pp. 6, 24, 32).
- [133] Y. Shen and J. Ke. “StainDiff: Transfer Stain Styles of Histology Images with Denoising Diffusion Probabilistic Models and Self-ensemble.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 549–559 (cit. on p. 12).

- [134] A. H. Song, G. Jaume, D. F. Williamson, M. Y. Lu, A. Vaidya, T. R. Miller, and F. Mahmood. "Artificial intelligence for digital and computational pathology." In: *Nature Reviews Bioengineering* 1.12 (2023), pp. 930–949 (cit. on pp. 4, 36).
- [135] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. Van Der Laak. "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology." In: *Medical image analysis* 58 (2019), p. 101544 (cit. on pp. 3, 5, 9, 27).
- [136] K. Thandiackal, B. Chen, P. Pati, G. Jaume, D. F. Williamson, M. Gabrani, and O. Goksel. "Differentiable zooming for multiple instance learning on whole-slide images." In: *European Conference on Computer Vision*. Springer. 2022, pp. 699–715 (cit. on p. 23).
- [137] M. Tittford. "A short history of histopathology technique." In: *Journal of Histotechnology* 29.2 (2006), pp. 99–110 (cit. on p. 1).
- [138] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. "Training data-efficient image transformers & distillation through attention." In: *International conference on machine learning*. PMLR. 2021, pp. 10347–10357 (cit. on p. 30).
- [139] M. Tran, P. Schmidle, S. J. Wagner, V. Koch, B. Novotny, V. Lupperger, A. Feuchtinger, A. Böhner, R. Kaczmarczyk, T. Biedermann, et al. "Generating clinical-grade pathology reports from gigapixel whole slide images with HistogPT." In: *medRxiv* (2024), pp. 2024–03 (cit. on pp. 4, 25).
- [140] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab. "Structure-preserving color normalization and sparse stain separation for histological images." In: *IEEE transactions on medical imaging* 35.8 (2016), pp. 1962–1971 (cit. on pp. 5, 10).
- [141] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, et al. "Applications of machine learning in drug discovery and development." In: *Nature reviews Drug discovery* 18.6 (2019), pp. 463–477 (cit. on p. 2).
- [142] P. J. Van Diest, E. van der Wall, and J. P. Baak. "Prognostic value of proliferation in invasive breast cancer: a review." In: *Journal of clinical pathology* 57.7 (2004), pp. 675–681 (cit. on p. 3).
- [143] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy. "Tracking emerges by colorizing videos." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 391–408 (cit. on p. 14).

- [144] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, K. Severson, E. Zimmermann, J. Hall, N. Tenenholtz, N. Fusi, et al. "A foundation model for clinical-grade computational pathology and rare cancers detection." In: *Nature medicine* (2024), pp. 1–12 (cit. on pp. 5, 19).
- [145] S. J. Wagner, N. Khalili, R. Sharma, M. Boxberg, C. Marr, W. De Back, and T. Peng. "Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations." In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII* 24. Springer. 2021, pp. 257–266 (cit. on p. 5).
- [146] S. J. Wagner, D. Reisenbüchler, N. P. West, J. M. Niehues, J. Zhu, S. Foersch, G. P. Veldhuizen, P. Quirke, H. I. Grabsch, P. A. van den Brandt, et al. "Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study." In: *Cancer Cell* 41.9 (2023), pp. 1650–1661 (cit. on pp. 4, 6, 23).
- [147] W. Wang, S. Ma, H. Xu, N. Usuyama, J. Ding, H. Poon, and F. Wei. "When an image is worth 1,024 x 1,024 words: A case study in computational pathology." In: *arXiv preprint arXiv:2312.03558* (2023) (cit. on p. 24).
- [148] X. Wang, Y. Du, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han. "RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval." In: *Medical image analysis* 83 (2023), p. 102645 (cit. on pp. 4, 19, 30).
- [149] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han. "Transformer-based unsupervised contrastive learning for histopathological image classification." In: *Medical image analysis* 81 (2022), p. 102559 (cit. on pp. 19, 30).
- [150] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li, et al. "A pathology foundation model for cancer diagnosis and prognosis prediction." In: *Nature* 634.8035 (2024), pp. 970–978 (cit. on pp. 24, 25).
- [151] K. Q. Weinberger and L. K. Saul. "Distance metric learning for large margin nearest neighbor classification." In: *Journal of machine learning research* 10.2 (2009) (cit. on p. 14).
- [152] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. "The cancer genome atlas pan-cancer analysis project." In: *Nature genetics* 45.10 (2013), pp. 1113–1120 (cit. on pp. 8, 9, 18, 32).

- [153] Z. Weng, A. Seper, A. Pryalukhin, et al. "GrandQC: A comprehensive solution to quality control problem in digital pathology." In: *Nature Communications* 15 (2024), p. 10685. doi: 10.1038/s41467-024-54769-y (cit. on pp. 3, 27).
- [154] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. "Unsupervised feature learning via non-parametric instance discrimination." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3733–3742 (cit. on p. 15).
- [155] K.-X. Xi, Y.-S. Wen, C.-M. Zhu, X.-Y. Yu, R.-Q. Qin, X.-W. Zhang, Y.-B. Lin, T.-H. Rong, W.-D. Wang, Y.-Q. Chen, et al. "Tumor-stroma ratio (TSR) in non-small cell lung cancer (NSCLC) patients after lung resection is a prognostic factor for survival." In: *Journal of thoracic disease* 9.10 (2017), p. 4017 (cit. on p. 3).
- [156] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh. "Nyströmformer: A nyström-based algorithm for approximating self-attention." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 16. 2021, pp. 14138–14148 (cit. on p. 32).
- [157] F. Xu, C. Zhu, W. Tang, Y. Wang, Y. Zhang, J. Li, H. Jiang, Z. Shi, J. Liu, and M. Jin. "Predicting Axillary Lymph Node Metastasis in Early Breast Cancer Using Deep Learning on Primary Tumor Biopsy Slides." In: *Frontiers in Oncology* (2021), p. 4133 (cit. on p. 9).
- [158] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, et al. "A whole-slide foundation model for digital pathology from real-world data." In: *Nature* (2024), pp. 1–8 (cit. on pp. 4, 5, 19, 24, 25).
- [159] S. Yang, Y. Wang, and H. Chen. "Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 296–306 (cit. on p. 24).
- [160] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. "Coca: Contrastive captioners are image-text foundation models." In: *arXiv preprint arXiv:2205.01917* (2022) (cit. on p. 18).
- [161] R. Zhang, P. Isola, and A. A. Efros. "Colorful image colorization." In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III* 14. Springer. 2016, pp. 649–666 (cit. on p. 13).
- [162] Y. Zheng, R. H. Gindra, E. J. Green, E. J. Burks, M. Betke, J. E. Beane, and V. B. Kolachalama. "A graph-transformer for whole slide image classification." In: *IEEE transactions on medical imaging* 41.11 (2022), pp. 3003–3015 (cit. on p. 22).

- [163] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. “ibot: Image bert pre-training with online tokenizer.” In: *arXiv preprint arXiv:2111.07832* (2021) (cit. on pp. 5, 17, 19, 25).
- [164] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.” In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. 2017 (cit. on p. 11).
- [165] E. Zimmermann, E. Vorontsov, J. Viret, A. Casson, M. Zelechowski, G. Shaikovski, N. Tenenholtz, J. Hall, D. Klimstra, R. Yousfi, et al. “Virchow2: Scaling self-supervised mixed magnification models in pathology.” In: *arXiv:2408.00738* (2024) (cit. on p. 19).

A. Original Publications

A.1. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations

This publication was published in the proceedings of the peer-reviewed international conference Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2021. The one-page summary can be found in Section 3.1.

Wagner, S. J., Khalili, N., Sharma, R., Boxberg, M., Marr, C., De Back, W., & Peng, T. (2021). Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII* 24 (pp. 257-266). Springer International Publishing.

Copyright: Copyright © 2021, Springer Nature. Reproduced with permission from Springer Nature. The license is attached after the original publication.

Structure-Preserving Multi-Domain Stain Color Augmentation using Style-Transfer with Disentangled Representations

Sophia J. Wagner^{1,2,4}, Nadieh Khalili⁵, Raghav Sharma³, Melanie Boxberg^{1,4},
Carsten Marr³, Walter de Back⁵, Tingying Peng^{1,2,4}

¹ Technical University Munich, Munich, Germany

² Helmholtz AI, Oberschleißheim, Germany

³ Insitute for Computational Biology, HelmholtzZentrum Munich, Germany

⁴ Munich School of Data Science (MuDS), Munich, Germany

⁵ ContextVision AB, Stockholm, Sweden

Abstract. In digital pathology, different staining procedures and scanners cause substantial color variations in whole-slide images (WSIs), especially across different laboratories. These color shifts result in a poor generalization of deep learning-based methods from the training domain to external pathology data. To increase test performance, stain normalization techniques are used to reduce the variance between training and test domain. Alternatively, color augmentation can be applied during training leading to a more robust model without the extra step of color normalization at test time. We propose a novel color augmentation technique, HistAuGAN, that can simulate a wide variety of realistic histology stain colors, thus making neural networks stain-invariant when applied during training. Based on a generative adversarial network (GAN) for image-to-image translation, our model disentangles the content of the image, i.e., the morphological tissue structure, from the stain color attributes. It can be trained on multiple domains and, therefore, learns to cover different stain colors as well as other domain-specific variations introduced in the slide preparation and imaging process. We demonstrate that HistAuGAN outperforms conventional color augmentation techniques on a classification task on the publicly available dataset Camelyon17 and show that it is able to mitigate present batch effects.¹

Keywords: color augmentation · style-transfer · disentangled representations.

1 Introduction

Modern cancer diagnosis relies on the expert analysis of tumor specimen and biopsies. To highlight its structure and morphological properties, conventionally, the tissue is stained with hematoxylin and eosin (H&E) [5]. The path from the raw tissue to the final digitized image slide however consists of many different

¹ Code and model weights are available at <https://github.com/sophiajw/HistAuGAN>.

processing steps that can introduce variances, such as tissue fixation duration, the age and the composition of the H&E-staining, or scanner settings. Therefore, histological images show a large variety of colors, not only differing between laboratories, but also within one laboratory [3].

This variability can lead to poor generalization of algorithms that are trained on WSIs from a single source. One strategy to account for this is stain color normalization. Traditionally, this is either done by aligning the color distribution of the test images to a reference tile in the training domain [12] or by decomposing the color space of a reference tile into hematoxylin and eosin components [10,17]. Then, H&E components of the test tiles can be aligned while keeping the structure intact.

Recently, the focus shifted toward the application of style-transfer methods such as cycle-consistent generative adversarial networks, cycleGAN [19], for stain normalization [16]. However, these models aim to match the target distribution possibly leading to undesired changes in the morphological structure [6]. To circumvent this, other approaches propose color space transformations [14], structural similarity loss functions [9], or residual learning [4].

We propose a novel histological color transfer model, HistAuGAN, based on a GAN architecture for image-to-image translation. In contrast to previous approaches, HistAuGAN disentangles the content of a histological image, i.e., the morphological tissue structure, from the stain color attributes, hence preserving the structure while altering the color. Therefore, HistAuGAN can be used as a stain augmentation technique during training of a task-specific convolutional neural network (CNN). We demonstrate that this helps to render the trained network color-invariant and makes it transferable to external datasets without an extra normalization step at test time. Applied as an augmentation technique, HistAuGAN significantly outperforms other color augmentation techniques on a binary tumor-classification task. Furthermore, clustering results suggest that HistAuGAN can capture sources of domain shifts beyond color variations, such as noise and artefacts introduced in the staining or digitization process, e.g., image compression or blurring.

To the best of our knowledge, HistAuGAN is the first GAN-based color augmentation technique that generates realistic histological color variations.

2 Method

2.1 Model architecture

We build our model based on a multi-domain GAN using disentangled representations, inspired by DRIT++ [8]. Originally designed for image-to-image translation on natural images using a predefined style, we propose its application on histological images to disentangle the morphological tissue structure from the visual appearance. In contrast to previous cycleGAN-based color normalization methods that use only a single encoder, HistAuGAN is able to separate two essential image properties from each other as visualized in Figure 1b: the

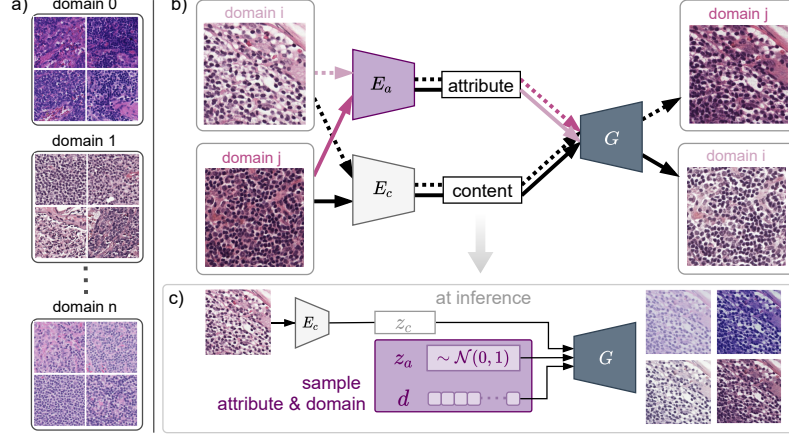


Fig. 1: We propose HistAuGAN for structure-preserving multi-domain stain color augmentation. (a) Histological slides from different laboratories (domains) exhibit color variations. (b) Model architecture. Here, the domain information flow is visualized by colored arrows. (c) At inference, HistAuGAN can be used as an augmentation technique by sampling attribute z_a and domain d .

domain-invariant content encoder E_c encodes the histopathological structure of the tissue, e.g., size and position of the nuclei, whereas the domain-specific attribute encoder E_a learns the domain-specific color appearance. The model can be trained on data from multiple domains and thereby captures both inter-laboratory variability between multiple domains and intra-laboratory variability within each domain at the same time. Finally, the generator G takes as input a content vector z_c , an attribute vector z_a , and the one-hot-encoded domain vector d and outputs a simulated histological image. The objective function is given by

$$L_{total} = w_{cc}L^{cc} + w_cL^c + w_dL^d + w_{recon}L^{recon} + w_{latent}L^{latent} + w_{KL}L^{KL}, \quad (1)$$

where L^{cc} is the cycle-consistency loss, L^c and L^d are adversarial losses for the content and the attribute encoder, L^{recon} is an L_1 -loss for image reconstruction, L^{latent} is an L_1 -loss for latent space reconstruction, and L^{KL} enforces the latent attribute space to be distributed according to the standard normal distribution. Please refer to [8] for a detailed explanation of each loss and the precise hyperparameter setting.

At inference, using the fixed content encoding of the input image z_c , we can sample the attribute vector z_a and the one-hot encoded domain vector d as visualized in Figure 1c. Hence, we can map one image to many different structure-preserving augmentations. More specifically, we sample a random color attribute z_a from a normal distribution that parametrizes the stain color variabilities in one domain. Figure 2b shows randomly sampled outcomes of intra-domain aug-

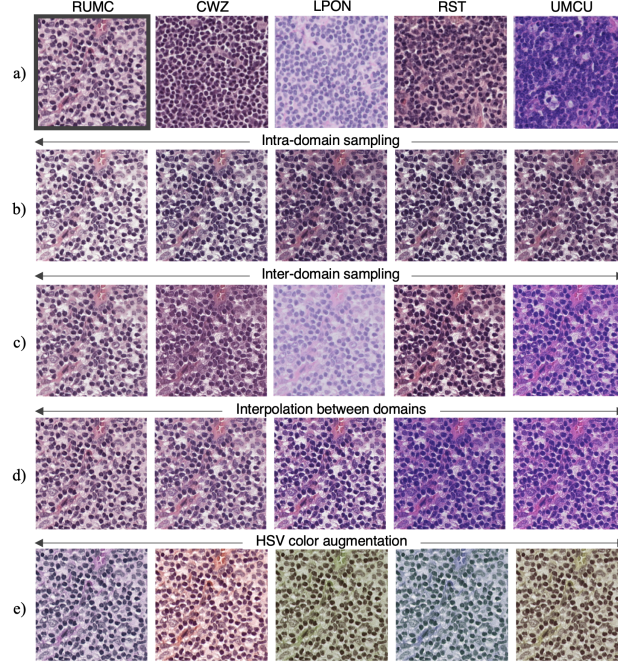


Fig. 2: Overview of the color variation in the dataset and the augmentation techniques used in this paper using the framed image as example tile.

mentations. Additionally, we can change the one-hot-encoded domain vector d to project the input image into multiple target domains as visualized in Figure 2c. In addition to sampling from the training domains, we can also interpolate between these domains to obtain an even broader variety of realistic color appearances for histopathological images. Figure 2d demonstrates this by linearly interpolating the domain from domain RUMC to domain UMCU according to

$$d = (1 - t) \cdot d_{\text{RUMC}} + t \cdot d_{\text{UMCU}}, \quad \text{for } t \in [0, 1]. \quad (2)$$

2.2 Competitive methods for stain color augmentation

Most existing stain color transfer methods are used for stain normalization, i.e., to transfer the stain color of the test domain to that of the training domain. Recently, it has been shown that simple stain color augmentations, such as perturbing the HSV color space of the histological images, perform better and lead to more robust models than traditional and network-based normalization techniques [15]. Therefore, we compare our HistAuGAN to the HSV augmentations used in [15]. Besides HSV augmentation, there is a more complicated augmentation technique based on the Wasserstein distance of different domains [11]. But

the method is much slower than HSV and HistAuGAN, thus difficult to be used as an on-the-fly augmentation technique.

For a quantitative evaluation of our augmentation technique, we consider the following augmentation methods:

- *Geometric augmentations*: vertical and horizontal flipping, as well as 90° , 180° , and 270° rotations.
- *HSV color augmentations*: geometric augmentations with Gaussian blur and contrast and brightness perturbations applied with probability 0.25 and 0.5, respectively. We tried both light and strong color augmentations, as suggested in [15]. Strong color augmentations can generate unrealistic color appearances. However, applying hue and saturation jittering with factor 0.5 and probability 0.5, which results in relatively strong color perturbances as shown in Figure 2e, performed best for us.
- *HistAuGAN*: geometric augmentations combined with our augmentation technique applied to half of the images during training. For each image, we randomly pick a target domain from the training domains and sample a color attribute vector $z_a \in \mathbb{R}^8$ from the standard normal distribution.

2.3 Evaluation

We evaluate HistAuGAN on three different aspects, in particular, i) whether it can remove batch effects present in histological images collected from multiple medical laboratories, ii) how it affects the out-of-domain generalization of a deep learning model trained for a specific down-stream task, and

iii) how HistAuGAN preserves morphological structure during augmentation.

For ii), we choose a binary classification task of classifying WSI tiles into the classes *tumor* versus *non-tumor* as described in more detail in Section 3.3. Question iii) is evaluated by asking a pathology expert to check image similarity before and after augmentation. To explore how generalizable our model is, we extend the HistAuGAN training data (lypmh nodes) by tiles from unseen tissue and tumor types, in particular, breast tissue [13].

3 Results and Discussion

3.1 Dataset

For the quantitative evaluation of HistAuGAN, we choose the publicly available Camelyon17 dataset [1] that provides WSIs from five different medical centers (denoted by RUMC, CWZ, UMCU, RST, and LPON) with different scanning properties and stain colors as shown in Figure 2a. Pixel-wise annotations are given for 50 WSIs in total, 10 from each medical center. To create the training patches, we first threshold the images with naive RGB thresholding combined with Otsu thresholding and then patch the tissue regions of each WSI at the highest resolution based on a grid into tiles of size 512×512 pixels. Each tile is labeled as *tumor* if the ratio of pixels annotated as tumor pixels is larger than

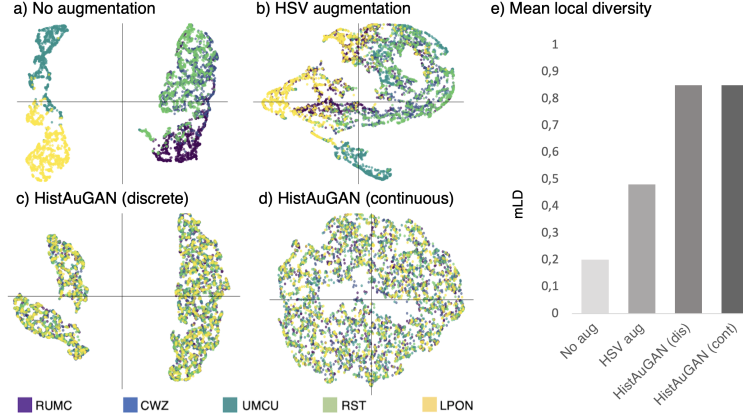


Fig. 3: Effect of color augmentation on batch effects in color statistics. (a-d) UMAP embeddings of color statistics of training data, color-coded by source domains. (e) The quantification of mixing based on mean local diversity (mLD, higher is better) suggests HistAuGAN effectively mitigates batch effects.

1%, otherwise, it is labelled as *non-tumor*. The tiled dataset has an *imbalanced* class distribution, i.e., overall, 7% of the tiles are labeled as *tumor* and the ratio of tumor tiles is in the same order of magnitude across all medical centers.

3.2 Evaluation of batch-effect removal

To evaluate how color augmentation mitigates batch effects, we quantify the mixing of images from different medical centers with respect to their color statistics. A random set of 1000 image tiles were extracted from the WSIs from each center and analyzed in terms of the average values of each component after transformation to various color spaces (RGB, HSV, LAB, HED, grayscale). To visually observe batch effects, we reduced dimensionality to 2D using UMAP [2] and labeled points according to their domain as shown in Figure 3a-d. To quantify mixing of different domains, we measured the mean over the local diversity (mLD) for all k -nearest neighborhoods ($k = 10$) in the 2D projection using Shannon’s equitability which varies between 0 for non-mixed and 1 for perfectly mixed populations (cf. Figure 3e).

Without color augmentation, we observe a clear batch effect: tiles from different domains form distinct clusters (mLD=0.20, Figure 3a). HSV augmentations improve data mixing but domain-correlated clusters are still visible (mLD=0.48, Figure 3b) and single domains, e.g. LPON, are not mixed with other domains. In contrast, HistAuGAN mixes data from multiple domains (Figure 3c,d) with a high local diversity (mLD=0.85). If HistAuGAN is used to transfer colors to discrete domains, the distinct domain clusters are retained, but each cluster contains well-mixed image samples transferred from all domains (Figure 3c). When HistAuGAN is used to randomly interpolate between domains, a continuous

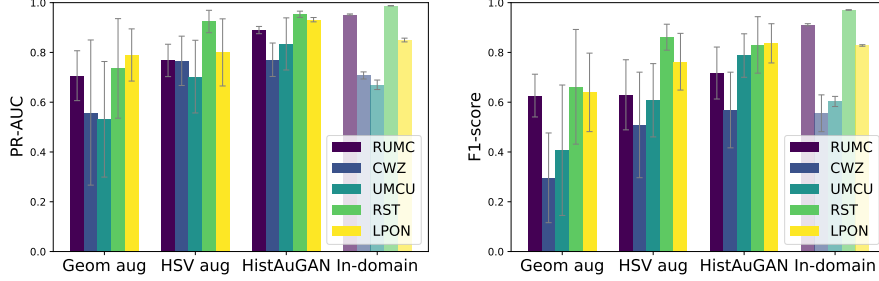


Fig. 4: Precision-recall AUC (left) and F1-score (right) of our binary classification task. The bold bars depict the results on the out-of-domain centers averaged across all runs. The most-right, pale bars denote the in-domain test performance of the classifiers trained with geometric augmentations.

well-mixed color subspace is obtained without any clustering structure (Figure 3d).

These results show that HistAuGAN is highly effective in removing batch effects present in color statistics of images sampled from different medical centers.

3.3 Evaluation on down-stream classification task

To evaluate the effect of our proposed augmentation method, we trained a CNN on a binary tumor classification task and compare the performance on different out-of-domain test sets based on the Camelyon17 dataset. Due to the relatively small size of our dataset, in particular the small number of tumor tiles, we choose a small CNN, namely, a pre-trained ResNet18 [7], and fine-tune the last two ResNet-blocks together with the fully-connected layer on our dataset. For training, we use weighted cross-entropy-loss to rebalance the contribution of each class, with a learning rate of $1e-5$ and an L_2 -regularization of $1e-5$ across all runs and for all augmentation techniques. Furthermore, we used random erasing on all augmentation techniques [18]. Since our dataset is highly imbalanced, we report the F1-score of the tumor class in addition to the area under the precision-recall curve (PR-AUC).

Figure 4 shows the results of the quantitative evaluation of different augmentation techniques on the binary tumor-classification task. For each medical center, we trained three classifiers, one for each augmentation type, and aggregated the results evaluated on the test domains. All experiments were repeated three times. On both metrics, HistAuGAN shows better performance on all of the out-of-domain test sets. As visualized in Figure 2, the appearance of images from medical center UMCU and LPON deviates strongly from the other centers, explaining their lower scores. In comparison to HSV color augmentation, HistAuGAN performs better in handling the stain color discrepancy between training and test domain and is therefore able to generate a more robust classification model that generalizes better to out-of-domain test sets. This can also be

Tissue type	High	Moderate	Low	Total
Lymph nodes	10	7	3	20
Breast	14	4	2	20

Table 1: Expert evaluation.

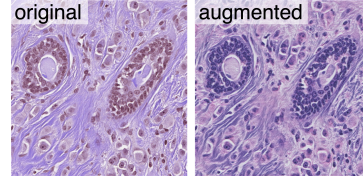


Fig. 5: HistAuGAN on unseen tissue.

measured in the standard deviation of the results across the out-of-domain test sets centers. For our model, the standard deviation of the PR-AUC for the tumor class amounts to 0.08, whereas it is higher for geometric (0.22) and color (0.14) augmentations, respectively, which demonstrates that our model is more robust to underlying stain color variations. The right-most group shows the in-domain test results for geometric augmentations. It can be seen as an upper bound for any stain normalization technique, and thus shows that HistAuGAN can even outperform stain normalization techniques on some of the five domains.

3.4 Qualitative evaluation by an expert pathologist

We further check the quality of HistAuGAN by an expert pathologist on the structural similarity of original and augmented WSI tiles from the training set, i.e., the Camelyon17 dataset, and an unseen dataset of breast tissue [13]. We define three levels of similarity: a) High similarity: a pathologist would find it difficult to distinguish the original tile from the augmented tile. b) Moderate similarity: some structural variations are observed, but do not affect pathological diagnosis. c) Low similarity: the augmented tiles can not be used for diagnostic purposes. As shown in Table 1, most of the augmented images do not have a structural modification that affects diagnosis and over half of them can even *fool* an expert pathologist. It is worth to mention that HistAuGAN is not trained on any of the breast cancer images but is still able to transfer its color in a structure-preserving manner as shown in Figure 5 on a sample tile.

4 Conclusion

In summary, we propose a novel GAN-based technique, HistAuGAN, for color augmentation of histopathological images. Based on the disentangled representations of content and style, HistAuGAN is able to change the color appearance of an histological image while preserving its morphological structure. Moreover, HistAuGAN captures both intra-domain and inter-domain color variations. It is able to interpolate between domains and can therefore span a continuous color space covering a large variety of realistic stain colors. When applied as an augmentation technique, HistAuGAN yields a robust down-stream classifier that generalizes better to out-of-domain test sets than other color augmentations

techniques and, therefore, renders additional stain normalization steps unnecessary. Finally, HistAuGAN can mitigate batch effects present in histopathological data which suggests that it is also able to cover domain shifts beyond color variations, such as noise and artefacts introduced in image compression. The code is publicly available at <https://github.com/sophiajw/HistAuGAN> together with a model of HistAuGAN trained on the five medical centers of the Camelyon17 dataset.

References

1. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermesen, M., Ehteshami Bejnordi, B., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F.G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A.B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Cetin, M., Halici, E., Jackson, H., Chen, R., Both, F., Franke, J., Kusters-Vandeveld, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J., Litjens, G.: From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Trans. Med. Imaging* **38**(2), 550–560 (Feb 2019)
2. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., Newell, E.W.: Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (Dec 2018)
3. Bejnordi, B.E., Litjens, G., Timofeeva, N., Otte-Holler, I., Homeyer, A., Karssemeijer, N., van der Laak, J.A.: Stain specific standardization of Whole-Slide histopathological images (2016)
4. de Bel, T., Bokhorst, J.M., van der Laak, J., Litjens, G.: Residual cyclegan for robust domain transformation of histopathological tissue slides. *Med. Image Anal.* **70**, 102004 (May 2021)
5. Chan, J.K.C.: The wonderful colors of the Hematoxylin–Eosin stain in diagnostic surgical pathology. *Int. J. Surg. Pathol.* **22**(1), 12–32 (Feb 2014)
6. Cohen, J.P., Luck, M., Honari, S.: Distribution matching losses can hallucinate features in medical image translation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part I. Lecture Notes in Computer Science*, vol. 11070, pp. 529–536. Springer (2018). https://doi.org/10.1007/978-3-030-00928-1_60, https://doi.org/10.1007/978-3-030-00928-1_60
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016)
8. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H.: DRIT : Diverse Image-to-Image translation via disentangled representations (2020)
9. Liang, H., Plataniotis, K.N., Li, X.: Stain style transfer of histopathology images via Structure-Preserved generative learning. In: *Machine Learning for Medical Image Reconstruction*. pp. 153–162. Springer International Publishing (2020)
10. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. pp. 1107–1110. IEEE (2009)

11. Nadeem, S., Hollmann, T., Tannenbaum, A.: Multimarginal wasserstein barycenter for stain normalization and augmentation. *Med. Image Comput. Comput. Assist. Interv.* **12265**, 362–371 (Oct 2020)
12. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. *IEEE Comput. Graph. Appl.* **21**(5), 34–41 (Jul 2001)
13. Roux, L.: Mitos-atypia-14 grand challenge. <https://mitos-atypia-14.grand-challenge.org/>, accessed: 2021-03-03
14. Shaban, M.T., Tarek Shaban, M., Baur, C., Navab, N., Albarqouni, S.: Staingan: Stain style transfer for digital histological images (2019)
15. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* **58**, 101544 (Dec 2019)
16. Tschuchnig, M.E., Oostingh, G.J., Gadermayr, M.: Generative adversarial networks in digital pathology: A survey on trends and future potential. *Patterns (N Y)* **1**(6), 100089 (Sep 2020)
17. Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N.: Structure-Preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* **35**(8), 1962–1971 (Aug 2016)
18. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(07), 13001–13008 (Apr 2020). <https://doi.org/10.1609/aaai.v34i07.7000>, <https://ojs.aaai.org/index.php/AAAI/article/view/7000>
19. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2223–2232 (2017)

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Dec 16, 2024

This Agreement between Sophia J. Wagner ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5930430618998
License date	Dec 15, 2024
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Structure-Preserving Multi-domain Stain Color Augmentation Using Style-Transfer with Disentangled Representations
Licensed Content Author	Sophia J. Wagner, Nadieh Khalili, Raghav Sharma et al
Licensed Content Date	Jan 1, 2021
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter

Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	Generalizable and Data- efficient Deep Learning for Computer-aided Cancer Diagnosis
Institution name	Technical University Munich
Expected presentation date	Feb 2025
The Requesting Person / Organization to Appear on the License	Sophia J. Wagner
Requestor Location	Sophia Wagner Ingolstädter Landstraße 1 Neuherberg, 85764 Germany
Billing Type	Invoice
Billing Address	Sophia Wagner Ingolstädter Landstraße 1 Neuherberg, Germany 85764
Total	0.00 EUR
Terms and Conditions	

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed

Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at <https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/>.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a	

journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: *'Reproduced with permission from Springer Nature'*.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany's choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit <https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution>

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

A.2. Multi-modality microscopy image style augmentation for nuclei segmentation

This publication was published in the peer-reviewed international journal *Journal of Imaging* in 2022. The one-page summary can be found in Section 3.2.

Liu, Y., Wagner, S. J., & Peng, T. (2022). Multi-modality microscopy image style augmentation for nuclei segmentation. *Journal of Imaging*, 8(3), 71.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>). The license is attached after the original publication.

Article

Multi-Modality Microscopy Image Style Augmentation for Nuclei Segmentation

Ye Liu ¹, Sophia J. Wagner ^{1,2}  and Tingying Peng ^{2,*} 

¹ Department of Mathematics, Technical University Munich, 85748 Garching, Germany; yeliu0724@hotmail.com (Y.L.); sophia.wagner@helmholtz-muenchen.de (S.J.W.)

² Helmholtz AI, Helmholtz Munich, 85764 Neuherberg, Germany

* Correspondence: tingying.peng@helmholtz-muenchen.de

Abstract: Annotating microscopy images for nuclei segmentation by medical experts is laborious and time-consuming. To leverage the few existing annotations, also across multiple modalities, we propose a novel microscopy-style augmentation technique based on a generative adversarial network (GAN). Unlike other style transfer methods, it can not only deal with different cell assay types and lighting conditions, but also with different imaging modalities, such as bright-field and fluorescence microscopy. Using disentangled representations for content and style, we can preserve the structure of the original image while altering its style during augmentation. We evaluate our data augmentation on the 2018 Data Science Bowl dataset consisting of various cell assays, lighting conditions, and imaging modalities. With our style augmentation, the segmentation accuracy of the two top-ranked Mask R-CNN-based nuclei segmentation algorithms in the competition increases significantly. Thus, our augmentation technique renders the downstream task more robust to the test data heterogeneity and helps counteract class imbalance without resampling of minority classes.

Keywords: style transfer; data augmentation; nuclei segmentation



Citation: Liu, Y.; Wagner, S.J.; Peng, T. Multi-Modality Microscopy Image Style Transfer for Nuclei Segmentation. *J. Imaging* **2022**, *8*, 71. <https://doi.org/10.3390/jimaging8030071>

Academic Editor: Lotfi Chaari

Received: 1 February 2022

Accepted: 9 March 2022

Published: 11 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The evaluation of cell-level features is a key task in the histopathological workflow. Features, such as nuclei shape and distribution, are used to determine cell, tissue, and cancer types and are therefore relevant for the clinical diagnosis of cancer, e.g., for cancer identification, grading, and prognosis [1]. Deep learning serves as a promising tool to quantify these features in an automated manner [2]. However, the performance of these models depends heavily on the quality and quantity of training data since they require accurate and time-consuming segmentation masks. To tackle this, the Kaggle competition 2018 Data Science Bowl (DSB) (<https://www.kaggle.com/c/data-science-bowl-2018> (accessed on 15 January 2022)) was initiated to find segmentation algorithms that are robust and effective in various microscopy set-ups, cell lines, and, most significantly, types of light microscopy. The nuclei presented in the images are from a variety of creatures, including humans, mice, and flies. In this dataset, most of the images are stained primarily with DAPI or Hoechst. Hematoxylin and eosin-stained tissue samples are also included in the dataset. Among 3891 submissions during the challenge, the top-performing methods were mostly based on the common segmentation network architectures, Mask R-CNN [3] and U-Net [4]. The most important factors influencing the competition ranking were the amount of training data (some teams also used private training data aside from the data provided by the challenge organizer), complex pre-processing and post-processing, and ensemble learning. A common problem that encumbers the performance is the dataset's bias: while segmenting fluorescent images, which make up most of the training set, is relatively easy and accurate, segmenting less-represented image types, such as bright-field images of stained tissue sections, is difficult and inaccurate [5].

Style transfer has been widely used for image synthesis from one modality to another, e.g., from computed tomography (CT) to magnetic resonance imaging (MRI) [6], from MRI to CT [7], from MRI to Positron Emission Tomography (PET) [8]. It has also been shown to be effective for histological images, e.g., for stain color augmentation [9]. In terms of style transfer architectures, conditional GAN (cGAN) [10] can be used for aligned image pairs and its generator can generate images of a certain class. For instance, Pix2pix [11] is a type of cGAN for paired image translation, in which the input image is conditional on the output pair. It receives a whole image as input rather than noise and class vector. The discriminator of Pix2Pix [11], called PatchGAN, generates a matrix of classifications instead of a single output for optimization. Additionally, Pix2pix [11] changed the generator to a U-Net [4] architecture. CycleGAN [12], on the other hand, introduces the cycle-consistency loss to overcome the dependency of paired images. It can be trained on two independent datasets. Unlike CycleGAN which uses a single encoder, DRIT++ [13] introduces a structure with two encoders and is capable of decomposing the image into domain-invariant content and domain-specific feature encoding. Another well-known architecture is StyleGAN [14], which increases the diversity of outputs and has more control over image features by better disentangling the latent factors of variations.

Among the top-25 submissions in the competition, the team BIOMAGic introduced the only approach using style transfer. They refined their approach after the competition and named it NucleAlzer [15]. In addition to the DSB dataset, they included annotated data from twelve data sources increasing the number of training images from 735 to 1102 and the number of annotated nuclei from 33,814 to 80,692. For style transfer, they trained a Pix2pix [11] style transfer network (from the original image to segmentation mask) for each cluster of similar images in the unannotated data, resulting in a total of 134 networks. With these, 20 synthetic image/mask pairs of each style with nuclei at certain places are generated and added to the training set to train Mask R-CNN [3] for instance segmentation. The winner [5] of the competition used an ensemble technique of a total of 32 models based on eight different model architectures. Pretrained Resnets, Dual Path networks, and Inception-Resnets are used as encoders and U-Net or feature pyramid networks (FPN) [16] as decoder architectures. The second-best approach used a single pretrained FPN-based neural network model and two customized output layers. A multitask framework is used to train the customized output layers. Additionally, they introduced a new loss function that penalizes the instance errors based on the object's size.

In this paper, we propose to facilitate network training by a GAN-based style transfer data augmentation technique. By synthesizing less-represented image types from well-represented ones, our style augmentation increases the number of images of minority types in the training set. More specifically, we use disentangled representations in the style transfer, i.e., we decompose the input into its content and style representations and only alter the style without changing the content. Since the structure is thereby preserved, the style augmented images have the same nuclei locations and shapes as the original image. Therefore, our augmentation technique presents the segmentation network a multitude of styles already during training, rendering it more robust to the heterogeneity in the test data. Our experiments on the DSB challenge dataset demonstrate that by increasing the variety in the training dataset, our augmentation technique improves the overall segmentation accuracy.

2. Materials and Methods

Our nuclei segmentation workflow consists of three steps: data augmentation (split into clustering and training of a style transfer network), training of an instance segmentation network, and evaluating the segmentation network with test time augmentation (see Figure 1). In the following, we detail each of the three steps.

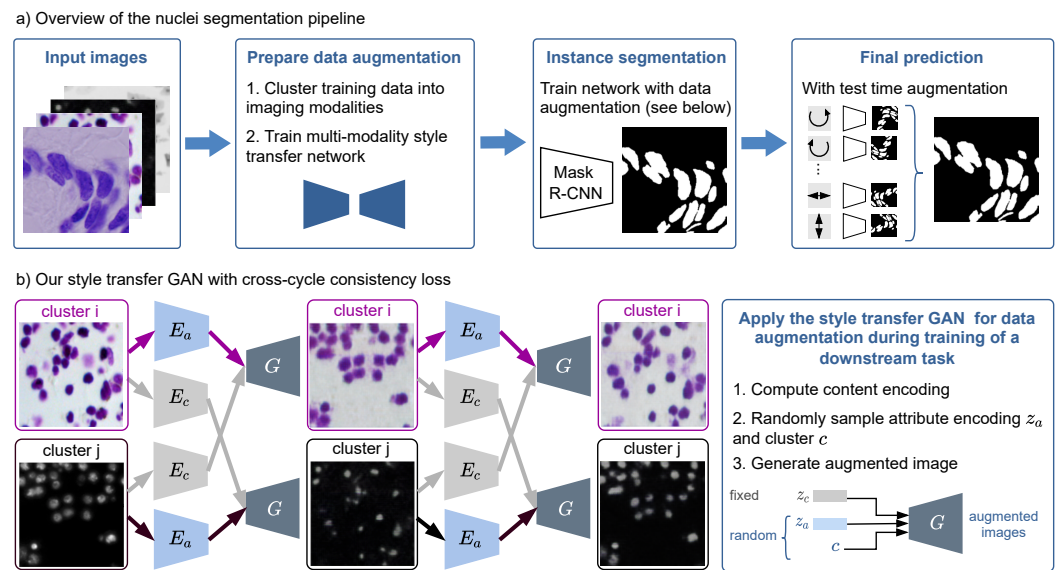


Figure 1. (a) Overview of our nuclei segmentation pipeline with multi-modality style transfer as data augmentation, Mask R-CNN as instance segmentation network, and test time augmentation for the final prediction. (b) Our style transfer GAN is trained with a cross-cycle consistency loss that does not need paired data. The right box shows how it is applied for data augmentation during training of the nuclei segmentation network.

2.1. Dataset

The dataset was provided by the competition organizers and is publicly available in the Broad Bioimage Benchmark Collection with access number BBBC038 (<https://bbbc.broadinstitute.org/BBBC038> (accessed on 15 January 2022)). It is divided into three sets: a training set of 670 images with nuclei segmentation masks provided (29,464 nuclei), a first-stage test set of 65 images (4152 nuclei), and a second-stage test set, which includes 3019 images (37,333 nuclei). The evaluation of nuclei segmentation for the annotated first-stage test set was disclosed later by the organizers. However, the official evaluation and ranking of a segmentation algorithm are based on the performance of the second-stage test set, which hides the annotated nuclei masks from the participants. Both training and test images are highly varied in terms of large experimental settings such as lighting conditions, cell lines, nuclei densities, and microscopy imaging modalities (with a few examples shown in Figures 1a and 2). It is worth noting that the test set is not only substantially larger than the training set, but also includes novel image modalities rendering the segmentation task extremely challenging.

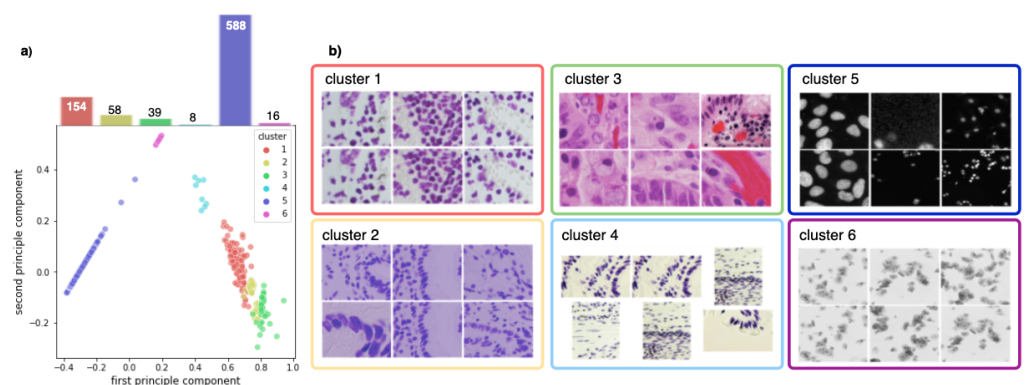


Figure 2. (a) PCA decomposition of hue saturation values of all training images and their class distribution. Each point represents one image, colored according to the corresponding cluster. (b) Exemplary images for each cluster.

2.2. Data Augmentation Step I: Clustering Training Images into Different Modalities

Since the dataset does not contain any metadata that can be used to determine the image modalities, we first divide our training images into multiple domains to perform image style transfer. Each cluster should ideally correspond to one imaging modality. We used the classic K-means algorithm to divide the training images into six distinct clusters based on their hue saturation value (HSV). Furthermore, we observed that the fluorescent images are typically dark and have low contrast. So we use the Contrast Limited Adaptive Histogram Equalization (CLAHE) (https://scikit-image.org/docs/dev/auto_examples/color_exposure/plot_equalize.html (accessed on 15 January 2022)) for contrast enhancement. For each section of an image, CLAHE computes a contrast histogram and adjusts the local contrast accordingly if that section is darker or brighter than the rest of the image.

2.3. Data Augmentation Step II: Multi-Modality Style Transfer

As shown in Figure 1b, for two exemplary images from different domains, the modality transfer model based on [13] first decomposes each image into a domain-invariant content encoding using the content encoder E_c and a domain-specific attribute encoding using the attribute encoder E_a , respectively. The content encoding and the attribute encoding, together with the domain information, can then be used to generate a synthetic image, where the structure is preserved. To create various synthetic images as augmentation during training of the segmentation network, we sample attribute encoding and domain vector randomly, while leaving the content-encoding fixed (see Figure 1b). We apply this augmentation technique randomly to half of the training images additionally to standard augmentations.

The underlying assumption of the style-transfer model is that the content space is shared across the domains. Therefore, during training, the content vector z_c can be exchanged between the modalities. Since the dataset is unpaired, i.e., we do not have ground truth for the style transfer, we use a cross-cycle consistency loss L_{cc} for training (see Figure 1b): The content encoding of two images from different modalities is swapped and a first image translation is generated. Subsequently, the translation is repeated, again by computing the encodings and then swapping the content encoding. The second image translation should resemble the original input, which is enforced by applying a loss based on the L_1 -norm. A classical adversarial loss L_d to enforce realistic synthetic images and an adversarial loss on the content encoder L_c to eliminate domain-specific information in the content encodings are also applied. Additionally, an L_1 -loss for image reconstruction, L_{recon} , is applied: the original image is decomposed into its content and attribute encodings and then reconstructed using those representations. To enable sampling from the latent space, L_{latent} is an L_1 -loss for latent space reconstruction, and the loss L_{KL} enforces the latent attribute space to be distributed according to a prior Gaussian distribution. Please refer to [13] for a detailed explanation of each loss. The final objective function is given by

$$L_{total} = w_{cc}L_{cc} + w_cL_c + w_dL_d + w_{recon}L_{recon} + w_{latent}L_{latent} + w_{KL}L_{KL}, \quad (1)$$

where the parameters w are used to weight the loss terms.

Since the content encoding is fixed during augmentation, the augmented image has the same nuclei locations and shapes as the input image and thus inherits the nuclei segmentation mask from the original image. This is a key difference between our approach and a common CycleGAN-based image style transfer [12], as there is no guarantee of content invariance after the CycleGAN style transfer.

2.4. Mask R-CNN for Instance Segmentation

For nuclei segmentation, we used two implementations published in the competition leaderboard (https://github.com/Lopezurrutia/DSB_2018 (accessed on 15 January 2022)) (https://github.com/mirzaevinom/data_science_bowl_2018 (accessed on 15 Jan-

uary 2022)) as baselines. Both teams use Mask R-CNN [3] for segmentation. The implementations differ in the image pre-processing and preparation of the network training. We choose these two methods because of their publicly accessible code, sufficient documentation, excellent performance (both methods were ranked among the top-5 in the leaderboard), as well as relatively simple network training (only one model needs to be trained, by contrast, the top-1 approach trained 32 models, which would consume considerably more training time and resources). Notably, the second approach developed by the team mirzaevinom. used additional data beyond the data provided by the organizer. However, we only used the annotated training data provided by the organizer to train our model, so our model performance differs slightly from their results reported in the leaderboard. For each method, we train two models, one with and one without multi-modality style transfer as augmentation technique.

2.5. Test Time Augmentation

In the inference stage, we use test time augmentation (TTA). TTA applies simple augmentations on the test images, e.g., rotating, flipping, color jittering, and image rescaling, and aggregates the model predictions on the augmented images. Thereby, TTA improves the robustness of Mask R-CNN without any additional training cost, leading to an increase in segmentation accuracy when compared to predicting only on the original image [17].

2.6. Implementation Details

All models were trained on Intel Xeon 6230 system with 128 GB of RAM and a Tesla V100 graphics card. The style transfer model is trained with batch size two and a learning rate of 0.0001 for 600 epochs. For the optimizer, we used the Adam optimizer with β_1 0.5 and β_2 0.999 and a weight decay of 0.0001. We trained the model on five domains and applied the style transfer augmentation with probability 0.5 in the segmentation downstream task.

3. Results

3.1. Clustering into Modalities

Figure 2b shows that the training set can be clustered into six sets that correspond indeed to the different imaging modalities present in the dataset. These clusters can also be seen in the low-dimensional principal component analysis (PCA) embedding of the training data visualized in Figure 2a. We observe that the training set is highly imbalanced: over 75% of all training images are fluorescent images, whilst the remaining bright-field images are grouped into five distinct, partly very small clusters.

3.2. Improved Segmentation Performance by Style Transfer Augmentation

Figure 3 shows a few examples of images created by our multi-modality style transfer GAN from one domain to the others. Visual assessment reveals that our generated images resemble real ones, implying that they are well suited for the augmentation task.

We quantify the add-on value of our proposed augmentation method by training models with and without augmentation of the multi-modality style transfer GAN. The final evaluation score is based on the object-level errors. If the predicted mask and segmented object have an Intersection-over-Union (IoU) score greater than a predefined threshold, they are regarded as correctly identified. The official competition metric S computes a precision score for semantic segmentation by evaluating true positives (TP), false positives (FP), and false negatives (FN) at several IoU thresholds t and then averaging over these, i.e.,

$$S = \frac{1}{|T|} \sum_{t \in T} \frac{TP(t)}{TP(t) + FP(t) + FN(t)}, \text{ where } T = \{0.1, 0.15, \dots, 0.95\}. \quad (2)$$

After that, the average of all the score is calculated over all images. By submitting our segmentation results of the second-stage test dataset to the Kaggle competition (<https://www.kaggle.com/competitions/2022-ai-challenge>).

[//www.kaggle.com/c/data-science-bowl-2018/submit](https://www.kaggle.com/c/data-science-bowl-2018/submit) (accessed on 15 January 2022)), we obtain the final evaluation score. As shown in Table 1, including style transfer augmentation in the method increased the nuclei segmentation accuracy from 53.2% to 60.9% for the method Deep Retina and 59.9% to 61.3% for method Inom Mirzaev, respectively. Notably, a score around 61% was ranked among the top-5 and almost only achieved by using additional datasets. In Figure 4, we can observe that our augmentation technique using the multi-modality style transfer GAN improves the segmentation results, especially in less-represented image modalities such as bright-field images of stained tissue sections and fluorescent images of large nuclei (most fluorescent images in the training set contain only small nuclei).

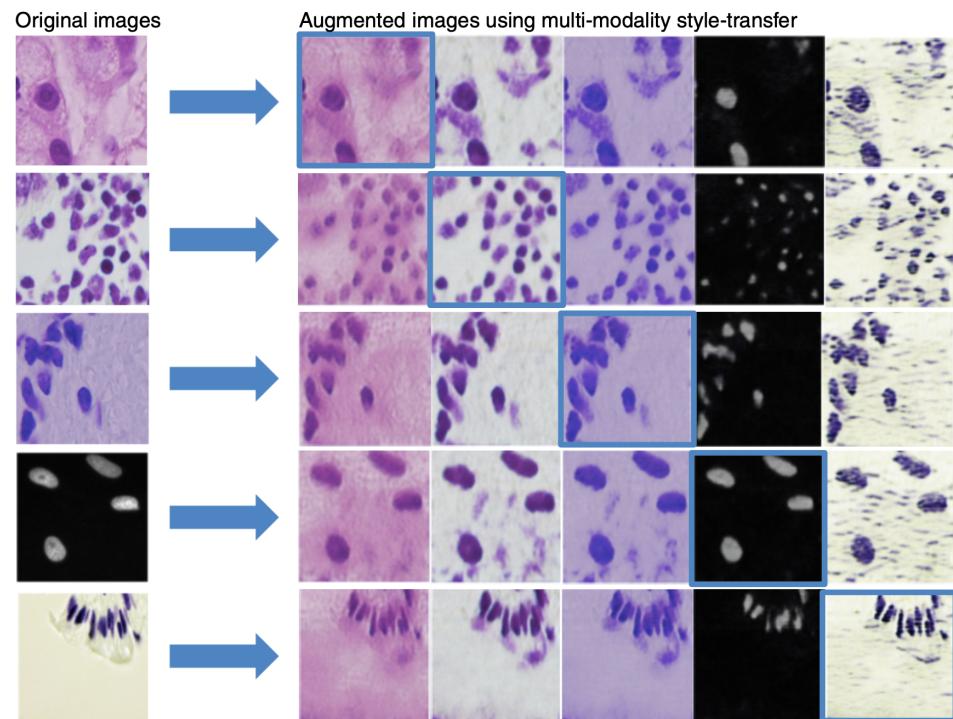


Figure 3. Examples of the outcome of multi-modality style transfer. **First column:** original images. **Remaining columns:** result of each domain's multi-modality style transfer with self-reconstruction on the diagonal, highlighted in blue boxes.

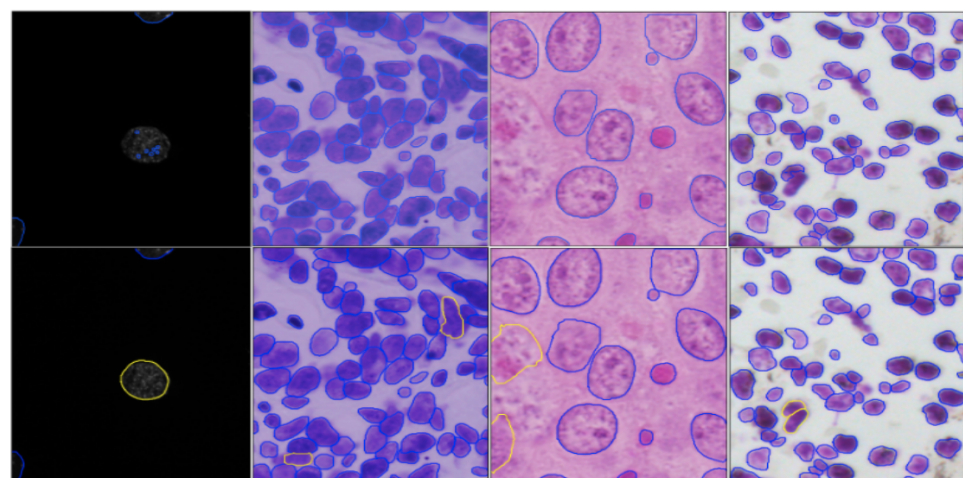


Figure 4. Nuclei segmentation results without (first row) and with (second row) our augmentation. Agreeing mask contours are shown in blue. Yellow contours in the second row show the improvements compared to the first row.

Table 1. Evaluation scores of different segmentation methods.

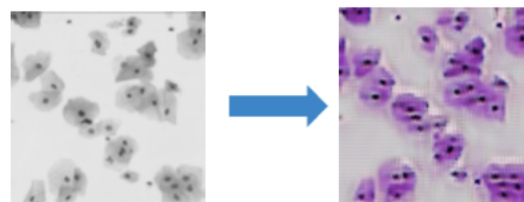
Method	Score [↑]
BIOMAGic	0.570
Deep Retina w/o our augmentation	0.532
Deep Retina w/ our augmentation	0.609
Inom Mirzaev w/o our augmentation	0.599
Inom Mirzaev w/our augmentation	0.613

In addition to the above two baseline methods, we also quote the result from the team BIOMAGic (57.0% IoU), as it was the only one to use style transfer among the top-25 submissions during the competition. Unlike our style transfer between multiple microscopy image modalities, BIOMAGic (and its successor nucleAIzer [15]) performed style transfer between images and their segmentation masks, which can be trained only by annotated images. In contrast, our method can be trained in an unsupervised fashion using images without annotated segmentation masks. In particular, their style transfer approach required training 134 networks for fine-grained image clusters for the same competition, whereas our approach only required the training of one network.

4. Discussion

Although our approach is effective in transferring style across the majority of modalities, it also has limitations. When using our style transfer GAN from image modalities that only show nuclei to another modality where both nuclei and cytoplasm are visible, unrealistic images are generated as demonstrated in Figure 5. This is because the underlying assumption of our model is violated. In particular, these two modalities do not only differ in style, but also in content; hence, the content and style cannot be efficiently disentangled.

Our style transfer model is built based on the DSB dataset, which mostly consists of images from fluorescence microscopy, primarily stained with DAPI or Hoechst, but also includes tissue samples stained with hematoxylin and eosin. Currently, our model is able to transfer the style between the five modalities contained in this dataset. Generally, it can be trained on all kinds of imaging modalities that share the same morphological structure and thereby fulfill our training assumption. Apart from the standard microscopy approaches above, our work can also be applied to images acquired by alternative microscopy technologies such as lens-free microscopy, a more cost-effective option that does not require expensive optical lenses and has a wider field of view. As a result, a much more affordable microscopy technique capable of monitoring cells continuously within an incubator and ideal for low-cost point-of-care devices aiming at resource-limited settings [18]. Our future work will involve transferring the style of lens-free microscopy images to the existing modalities.

**Figure 5.** The style transfer GAN fails when images at two modalities contain different content.

5. Conclusions

In summary, we developed an augmentation technique using a multi-modality style transfer GAN to transfer microscopy nuclei images from one modality to another. During training a Mask R-CNN for nuclei segmentation, this augmentation strategy facilitates the

training by increasing the diversity of the training images, hence making it more robust to the test data heterogeneity and resulting in better segmentation accuracy.

Author Contributions: Conceptualization and methodology, S.J.W. and T.P.; software, validation, formal analysis, investigation, Y.L.; resources, S.J.W.; data curation, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, S.J.W., T.P.; visualization, S.J.W., Y.L.; supervision, S.J.W., T.P.; project administration, T.P. All authors have read and agreed to the published version of the manuscript.

Funding: The second author was supported by the Munich School for Data Science (MUDS).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available in the public domain: Broad Bioimage Benchmark Collection at <https://bbbc.broadinstitute.org/BBBC038/> (accessed on 15 January 2022) with accession number BBBC038.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gurcan, M.N.; Boucheron, L.E.; Can, A.; Madabhushi, A.; Rajpoot, N.M.; Yener, B. Histopathological image analysis: A review. *IEEE Rev. Biomed. Eng.* **2009**, *2*, 147–171. [[CrossRef](#)] [[PubMed](#)]
2. Moen, E.; Bannon, D.; Kudo, T.; Graf, W.; Covert, M.; Valen, D.V. *Deep Learning for Cellular Image Analysis*; Springer Science and Business Media LLC: New York, NY, USA, 2019; Volume 16, pp. 1233–1246. [[CrossRef](#)]
3. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: Venice, Italy, 2017. [[CrossRef](#)]
4. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Lecture Notes in Computer Science*; Springer International Publishing: New York, NY, USA, 2015; pp. 234–241. [[CrossRef](#)]
5. Caicedo, J.C.; Goodman, A.; Karhohs, K.W.; Cimini, B.A.; Ackerman, J.; Haghighi, M.; Heng, C.; Becker, T.; Doan, M.; McQuin, C.; et al. Nucleus segmentation across imaging experiments: The 2018 Data Science Bowl. *Nat. Methods* **2019**, *16*, 1247–1253. [[CrossRef](#)]
6. Jiang, J.; Hu, Y.C.; Tyagi, N.; Zhang, P.; Rimner, A.; Mageras, G.S.; Deasy, J.O.; Veeraraghavan, H. Tumor-Aware, Adversarial Domain Adaptation from CT to MRI for Lung Cancer Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*; Springer International Publishing: New York, NY, USA, 2018; pp. 777–785. [[CrossRef](#)]
7. Emami, H.; Dong, M.; Nejad-Davarani, S.P.; Glide-Hurst, C.K. SA-GAN: Structure-Aware GAN for Organ-Preserving Synthetic CT Generation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*; Springer International Publishing: New York, NY, USA, 2021; pp. 471–481. [[CrossRef](#)]
8. Sikka, A.; Virk, J.S.; Bathula, D.R. MRI to PET Cross-Modality Translation Using Globally and Locally Aware GAN (GLA-GAN) for Multi-Modal Diagnosis of Alzheimer’s Disease. *arXiv* **2021**, arXiv:2108.02160.
9. Wagner, S.J.; Khalili, N.; Sharma, R.; Boxberg, M.; Marr, C.; Back, W.d.; Peng, T. Structure-Preserving Multi-domain Stain Color Augmentation Using Style-Transfer with Disentangled Representations. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: New York, NY, USA, 2021; pp. 257–266.
10. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:cs.LG/1411.1784.
11. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv* **2017**, arXiv:1611.07004.
12. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; IEEE: Venice, Italy, 2017. [[CrossRef](#)]
13. Lee, H.Y.; Tseng, H.Y.; Huang, J.B.; Singh, M.K.; Yang, M.H. Diverse Image-to-Image Translation via Disentangled Representations. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
14. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv* **2018**, arXiv:1812.04948v3.
15. Hollandi, R.; Szkalitsky, A.; Toth, T.; Tasnadi, E.; Molnar, C.; Mathe, B.; Grexa, I.; Molnar, J.; Balind, A.; Gorbe, M.; et al. nucleAIzer: A Parameter-free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer. *Cell Syst.* **2020**, *10*, 453–458.e6. [[CrossRef](#)] [[PubMed](#)]
16. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. *arXiv* **2016**, arXiv:1612.03144.

17. Moshkov, N.; Mathe, B.; Kertesz-Farkas, A.; Hollandi, R.; Horvath, P. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Sci. Rep.* **2020**, *10*, 5068. [[CrossRef](#)] [[PubMed](#)]
18. Kim, S.B.; Bae, H.; Koo, K.I.; Dokmeci, M.R.; Ozcan, A.; Khademhosseini, A. Lens-Free Imaging for Biological Applications. *J. Lab. Autom.* **2012**, *17*, 43–49. [[CrossRef](#)] [[PubMed](#)]

License:

Attribution 4.0 International

=====

Creative Commons Corporation ("Creative Commons") is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an "as-is" basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

Using Creative Commons Public Licenses

Creative Commons public licenses provide a standard set of terms and conditions that creators and other rights holders may use to share original works of authorship and other material subject to copyright and certain other rights specified in the public license below. The following considerations are for informational purposes only, are not exhaustive, and do not form part of our licenses.

Considerations for licensors: Our public licenses are intended for use by those authorized to give the public permission to use material in ways otherwise restricted by copyright and certain other rights. Our licenses are irrevocable. Licensors should read and understand the terms and conditions of the license they choose before applying it. Licensors should also secure all rights necessary before applying our licenses so that the public can reuse the material as expected. Licensors should clearly mark any material not subject to the license. This includes other CC-licensed material, or material used under an exception or limitation to copyright. More considerations for licensors: wiki.creativecommons.org/Considerations_for_licensors

Considerations for the public: By using one of our public licenses, a licensor grants the public permission to use the licensed material under specified terms and conditions. If the licensor's permission is not necessary for any reason--for example, because of any applicable exception or limitation to copyright--then that use is not regulated by the license. Our licenses grant only permissions under copyright and certain other rights that a licensor has authority to grant. Use of the licensed material may still be restricted for other

A. Original Publications

reasons, including because others have copyright or other rights in the material. A licensor may make special requests, such as asking that all changes be marked or described.

Although not required by our licenses, you are encouraged to respect those requests where reasonable. More considerations for the public:

wiki.creativecommons.org/Considerations_for_licensees

Creative Commons Attribution 4.0 International Public License

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 -- Definitions.

- a. Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.
- b. Adapter's License means the license You apply to Your Copyright and Similar Rights in Your contributions to Adapted Material in accordance with the terms and conditions of this Public License.
- c. Copyright and Similar Rights means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section 2(b)(1)-(2) are not Copyright and Similar Rights.
- d. Effective Technological Measures means those measures that, in the absence of proper authority, may not be circumvented under laws

A. Original Publications

fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements.

- e. Exceptions and Limitations means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material.
- f. Licensed Material means the artistic or literary work, database, or other material to which the Licensor applied this Public License.
- g. Licensed Rights means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of the Licensed Material and that the Licensor has authority to license.
- h. Licensor means the individual(s) or entity(ies) granting rights under this Public License.
- i. Share means to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them.
- j. Sui Generis Database Rights means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world.
- k. You means the individual or entity exercising the Licensed Rights under this Public License. Your has a corresponding meaning.

Section 2 -- Scope.

a. License grant.

- 1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:

- a. reproduce and Share the Licensed Material, in whole or

A. Original Publications

in part; and

b. produce, reproduce, and Share Adapted Material.

2. Exceptions and Limitations. For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.
3. Term. The term of this Public License is specified in Section 6(a).
4. Media and formats; technical modifications allowed. The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)(4) never produces Adapted Material.
5. Downstream recipients.
 - a. Offer from the Licensor -- Licensed Material. Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License.
 - b. No downstream restrictions. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material.
6. No endorsement. Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others designated to receive attribution as provided in Section 3(a)(1)(A)(i).

b. Other rights.

A. Original Publications

1. Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise.
2. Patent and trademark rights are not licensed under this Public License.
3. To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases the Licensor expressly reserves any right to collect such royalties.

Section 3 -- License Conditions.

Your exercise of the Licensed Rights is expressly made subject to the following conditions.

a. Attribution.

1. If You Share the Licensed Material (including in modified form), You must:
 - a. retain the following if it is supplied by the Licensor with the Licensed Material:
 - i. identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated);
 - ii. a copyright notice;
 - iii. a notice that refers to this Public License;
 - iv. a notice that refers to the disclaimer of warranties;
 - v. a URI or hyperlink to the Licensed Material to the extent reasonably practicable;

A. Original Publications

- b. indicate if You modified the Licensed Material and retain an indication of any previous modifications; and
 - c. indicate the Licensed Material is licensed under this Public License, and include the text of, or the URI or hyperlink to, this Public License.
- 2. You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information.
- 3. If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable.
- 4. If You Share Adapted Material You produce, the Adapter's License You apply must not prevent recipients of the Adapted Material from complying with this Public License.

Section 4 -- Sui Generis Database Rights.

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

- a. for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database;
- b. if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database Rights (but not its individual contents) is Adapted Material; and
- c. You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database.

For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.

Section 5 -- Disclaimer of Warranties and Limitation of Liability.

- a. UNLESS OTHERWISE SEPARATELY UNDERTAKEN BY THE LICENSOR, TO THE

A. Original Publications

EXTENT POSSIBLE, THE LICENSOR OFFERS THE LICENSED MATERIAL AS-IS AND AS-AVAILABLE, AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE LICENSED MATERIAL, WHETHER EXPRESS, IMPLIED, STATUTORY, OR OTHER. THIS INCLUDES, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT, ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OR ABSENCE OF ERRORS, WHETHER OR NOT KNOWN OR DISCOVERABLE. WHERE DISCLAIMERS OF WARRANTIES ARE NOT ALLOWED IN FULL OR IN PART, THIS DISCLAIMER MAY NOT APPLY TO YOU.

- b. TO THE EXTENT POSSIBLE, IN NO EVENT WILL THE LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY (INCLUDING, WITHOUT LIMITATION, NEGLIGENCE) OR OTHERWISE FOR ANY DIRECT, SPECIAL, INDIRECT, INCIDENTAL, CONSEQUENTIAL, PUNITIVE, EXEMPLARY, OR OTHER LOSSES, COSTS, EXPENSES, OR DAMAGES ARISING OUT OF THIS PUBLIC LICENSE OR USE OF THE LICENSED MATERIAL, EVEN IF THE LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH LOSSES, COSTS, EXPENSES, OR DAMAGES. WHERE A LIMITATION OF LIABILITY IS NOT ALLOWED IN FULL OR IN PART, THIS LIMITATION MAY NOT APPLY TO YOU.
- c. The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.

Section 6 -- Term and Termination.

- a. This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically.
- b. Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:
 - 1. automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or
 - 2. upon express reinstatement by the Licensor.

For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.

- c. For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop

A. Original Publications

distributing the Licensed Material at any time; however, doing so will not terminate this Public License.

- d. Sections 1, 5, 6, 7, and 8 survive termination of this Public License.

Section 7 -- Other Terms and Conditions.

- a. The Licensor shall not be bound by any additional or different terms or conditions communicated by You unless expressly agreed.
- b. Any arrangements, understandings, or agreements regarding the Licensed Material not stated herein are separate from and independent of the terms and conditions of this Public License.

Section 8 -- Interpretation.

- a. For the avoidance of doubt, this Public License does not, and shall not be interpreted to, reduce, limit, restrict, or impose conditions on any use of the Licensed Material that could lawfully be made without permission under this Public License.
- b. To the extent possible, if any provision of this Public License is deemed unenforceable, it shall be automatically reformed to the minimum extent necessary to make it enforceable. If the provision cannot be reformed, it shall be severed from this Public License without affecting the enforceability of the remaining terms and conditions.
- c. No term or condition of this Public License will be waived and no failure to comply consented to unless expressly agreed to by the Licensor.
- d. Nothing in this Public License constitutes or may be interpreted as a limitation upon, or waiver of, any privileges and immunities that apply to the Licensor or You, including from the legal processes of any jurisdiction or authority.

=====
Creative Commons is not a party to its public licenses. Notwithstanding, Creative Commons may elect to apply one of its public licenses to material it publishes and in those instances will be considered the "Licensor." The text of the Creative Commons public licenses is dedicated to the public domain under the CC0 Public

A. Original Publications

Domain Dedication. Except for the limited purpose of indicating that material is shared under a Creative Commons public license or as otherwise permitted by the Creative Commons policies published at creativecommons.org/policies, Creative Commons does not authorize the use of the trademark "Creative Commons" or any other trademark or logo of Creative Commons without its prior written consent including, without limitation, in connection with any unauthorized modifications to any of its public licenses or any other arrangements, understandings, or agreements concerning use of licensed material. For the avoidance of doubt, this paragraph does not form part of the public licenses.

Creative Commons may be contacted at creativecommons.org.

A.3. S5CL: unifying fully-supervised, self-supervised, and semi-supervised learning through hierarchical contrastive learning

This publication was published in the proceedings of the peer-reviewed international conference Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2022. The one-page summary can be found in Section 3.3.

Tran, M., Wagner, S. J., Boxberg, M., & Peng, T. (2022, September). S5CL: unifying fully-supervised, self-supervised, and semi-supervised learning through hierarchical contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 99-108). Cham: Springer Nature Switzerland.

Copyright: Copyright © 2022, Springer Nature. Reproduced with permission from Springer Nature. The license is attached after the original publication.



S5CL: Unifying Fully-Supervised, Self-supervised, and Semi-supervised Learning Through Hierarchical Contrastive Learning

Manuel Tran^{1,3(✉)}, Sophia J. Wagner^{1,3}, Melanie Boxberg^{1,2},
and Tingying Peng³

¹ Technical University of Munich, Munich, Germany

`manuel.tran@helmholtz-muenchen.de`

² Pathology Munich-North, Munich, Germany

³ Helmholtz AI, Helmholtz Munich – German Research Center for Environmental Health, Neuherberg, Germany

Abstract. In computational pathology, we often face a scarcity of annotations and a large amount of unlabeled data. One method for dealing with this is semi-supervised learning which is commonly split into a self-supervised pretext task and a subsequent model fine-tuning. Here, we compress this two-stage training into one by introducing S5CL, a unified framework for fully-supervised, self-supervised, and semi-supervised learning. With three contrastive losses defined for labeled, unlabeled, and pseudo-labeled images, S5CL can learn feature representations that reflect the hierarchy of distance relationships: similar images and augmentations are embedded the closest, followed by different looking images of the same class, while images from separate classes have the largest distance. Moreover, S5CL allows us to flexibly combine these losses to adapt to different scenarios. Evaluations of our framework on two public histopathological datasets show strong improvements in the case of sparse labels: for a H&E-stained colorectal cancer dataset, the accuracy increases by up to 9% compared to supervised cross-entropy loss; for a highly imbalanced dataset of single white blood cells from leukemia patient blood smears, the F1-score increases by up to 6% (Code: <https://github.com/manuel-tran/s5cl>).

Keywords: Contrastive learning · Self-supervision · Semi-supervision

1 Introduction

Pixel-wise annotation of histopathological data is highly time-consuming due to the large scale of whole-slide images (WSIs). Therefore, sparse annotations are commonly used [2], but this results in small amounts of labeled data and an abundance of unlabeled data – an ideal use case for semi-supervised learning.

Current semi-supervised methods often depend on self-supervised pre-training which applies domain-specific pretext tasks or contrastive learning techniques to extract useful features from the input data without relying on label

information [12, 18]. Examples of methods that use domain-specific pretext tasks are COCO [23] and PC-CHiP [5] which utilize cross-stain prediction and tissue prediction on histopathological datasets, respectively. The quality of the learned representations, however, highly depends on the designed pretext task [11]. SimCLR [3], BYOL [7], and Barlow Twins [24], on the other hand, use contrastive techniques to learn feature representations that are invariant to various distortions of the input sample. These methods usually define positive pairs (two differently augmented versions of the same image) and negative pairs (two different images). A contrastive loss then pushes the representations of positive pairs toward each other while pushing the representations of negative pairs away. A common problem with this approach is that the extracted features can differ significantly from the ones learned through full-supervision [17]. Therefore, recent approaches in semi-supervised learning propose to train with both labeled and unlabeled images at the same time. S4L-Exemplar [25], for instance, combines a supervised cross-entropy loss on labeled images and a self-supervised triplet loss on the whole dataset. MOAM [25] and FixMatch [19] extend this approach by further applying pseudo-labels on unlabeled images. To overcome a confirmation bias toward large classes [1], Noisy Student [22] and Meta Pseudo Labels [16]

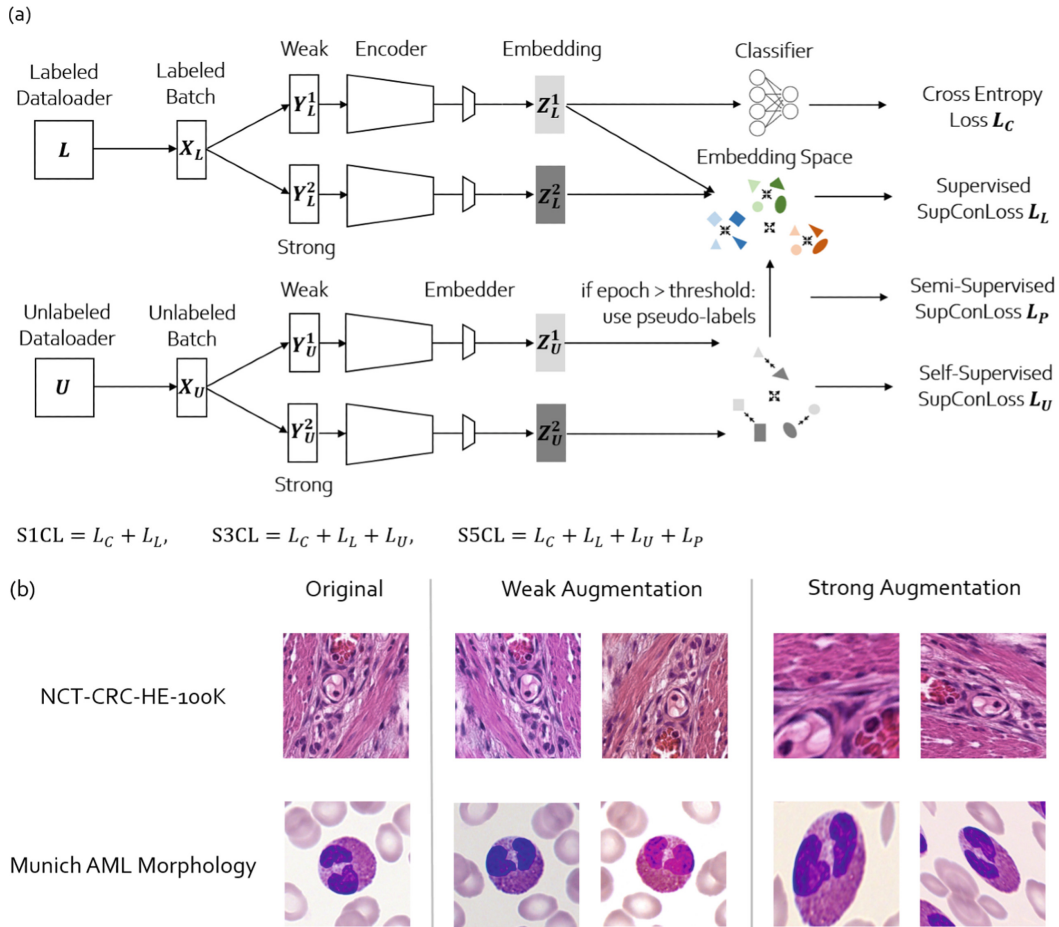


Fig. 1. (a) Overview of S5CL: Colors represent classes; shapes represent instances. Different shades or stretches indicate strongly and weakly augmented data points. (b) Examples of weak and strong augmentations on the two datasets we use.

use a teacher-student framework and additionally apply noise or a feedback system, respectively. Yet, these models still have difficulties with highly imbalanced datasets as they are usually biased toward the majority classes [15].

Contributions. To solve these issues, we propose a novel framework, called S5CL, that unifies fully-supervised, self-supervised, and semi-supervised learning through hierarchical contrastive learning. We train our models with three contrastive losses simultaneously to construct an embedding space that takes into account the hierarchy of distance relationships between images with respect to their class labels and consistency with different degrees of augmentations. The resulting framework is easy to use and highly flexible: We can omit unlabeled images and train fully-supervised; we can also set the weights of the supervised and semi-supervised loss to zero and train self-supervised; or we train with both labeled and unlabeled images in a semi-supervised way.

Unlike CoCo or PC-CHiP, S5CL does not require an artificially created pre-text task. And different from SimCLR or BYOL, S5CL does not require, large batch sizes, long training times, and high-capacity networks [4]. Further, our approach does not need to be fine-tuned on labeled data in a second step like all self-supervised methods. In fact, as mentioned above, self-supervised training is even a special case of our framework. Similar to semi-supervised techniques like S4L or FixMatch, S5CL also learns from labeled and unlabeled images simultaneously, but neither applies a contrastive loss on labeled images directly: S4L treats labeled images as unlabeled instances in their triplet loss and FixMatch applies a consistency loss. It can be seen in Sect. 3 that a qualitative embedding space does indeed correspond with better classification performance.

We validate our method on two public datasets: a colorectal cancer dataset with H&E-stained image patches (NCT-CRC-HE-100K) [9] and a highly imbalanced dataset of single white blood cells from leukemia patients (Munich AML Morphology) [13].

2 Method

S5CL combines three contrastive losses as visualized in Fig. 1a. First, a supervised contrastive loss embeds images with the same class label into the same cluster and pushes images from other classes away. We improve this by adding weakly and strongly augmented images (see Fig. 1b for some examples) to the loss function (S1CL). Now, weak augmentations and similar images are embedded the closest, followed by different-looking images of the same class or strongly augmented views. Images from different classes are the furthest away. Next, a self-supervised contrastive loss additionally pulls the feature representation of unlabeled images and their augmentations toward each other (S3CL). Since both losses act on the same embedding space, the unlabeled image is also forced to have a representation close to its actual but not available class cluster. We further improve the performance by adding pseudo-labels. Instead of applying them to the supervised branch of our method, we add a semi-supervised contrastive loss to circumvent the confirmation bias mentioned above (S5CL). As contrastive

loss, we choose the state-of-the-art SupConLoss [10]. It outperforms other losses like SimCLR [3], only has one hyperparameter, the temperature, and does not require hard negative mining like the triplet loss. However, our framework is more general and also works with other contrastive losses.

2.1 Description of S5CL

Given a labeled dataset L , we sample a batch X_L of labeled images. From X_L two batches of distorted images Y_L^1 and Y_L^2 are created via data augmentation. Y_L^1 is weakly augmented with simple color and geometric transformations. It is used for the supervised contrastive loss and the cross-entropy loss. The second batch Y_L^2 , on the contrary, is strongly augmented and used for supervised contrastive learning only. Both batches are passed on to the encoder network. The last layer of the encoder is followed by a bottleneck layer called the embedder, producing embeddings Z_L^1 and Z_L^2 . The labeled embeddings are then assessed by the supervised SupConLoss L_L [10]:

$$L_L(\tau) = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}. \quad (1)$$

Here, the anchor $\mathbf{z}_i \in Z_L^j$ is an embedding vector with class label y_i . As a measure of similarity, SupConLoss calculates the inner product of the anchor with all positive samples in the batch I that belong to the same class, $P(i) = \{p \in A(i) : y_i = y_p\}$. It then applies an exponential function that amplifies large values. The outputs are summed up and normalized over all samples $A = I \setminus \{i\}$. An important hyperparameter is the supervised temperature $\tau = T_L$ that helps distinguishing positive and negative samples. T_L controls the cluster density since intuitively, after training, we have $\mathbf{z}_i \cdot \mathbf{z}_p > T_L > \mathbf{z}_i \cdot \mathbf{z}_n$ with $n \in P(i)^C$.

Similar to the labeled dataset, given an unlabeled dataset U , a batch X_U of unlabeled images is sampled and two distorted views Y_U^1 and Y_U^2 of weakly and strongly augmented images are created. The unlabeled embeddings Z_U^1 and Z_U^2 are then processed by the self-supervised SupConLoss L_U . It has the same form as Eq. 1 but with a different temperature T_U . In L_U , the positive samples $\mathbf{z}'_i$ are the different augmented versions of the same image while the negative samples are different images. This could give rise to two contradicting loss objectives: two different samples from the same class are pulled together in the supervised loss L_L (as they belong to the same class) but get pushed away in the self-supervised loss L_U (as they are different samples). In S5CL, we solve this by choosing $T_U > T_L$ and achieve a hierarchy of distance relationships:

$$\mathbf{z}_i \cdot \mathbf{z}'_i > T_U > \mathbf{z}_i \cdot \mathbf{z}_p > T_L > \mathbf{z}_i \cdot \mathbf{z}_n. \quad (2)$$

After a fixed number t of epochs e , the classifier can be applied to Z_U^1 , yielding pseudo-labels. This is passed on to L_U . The loss is now semi-supervised and called L_P with temperature T_P . The total loss function is thus

$$L_T = \mathbb{1}_{e < t} \cdot w_U L_U + \mathbb{1}_{e > t} \cdot w_P L_P + w_L L_L + w_C L_C, \quad (3)$$

where $\mathbb{1}_{e < t}$ and $\mathbb{1}_{e > t}$ are indicator functions. The positive parameters w_U , w_P , w_L , and w_C are the weights of each loss which are always one in our case.

2.2 Augmentations

Each input image is augmented weakly and strongly. Weak augmentations consist of rotations by 0° , 90° , 180° , and 270° ; vertical and horizontal flipping; as well as color jitter (with brightness, contrast, saturation, and hue values all set to 0.1). We always apply rotations and color jitter while flipping is performed with a probability of 0.5.

Strong augmentations include all augmentations above except that color jitter is replaced by light HED jitter. Additionally, we apply Inception-crop for tissue images [20]. For single-cell images, we crop between 50% and 100% of the image around the center and re-scale the image to its original size. We lastly employ random affine linear transformations with interpolation (see Fig. 1b).

2.3 Competitive Methods and Implementation Details

We compare S5CL to the following baseline models: (i) a fully-supervised model that is trained with a cross-entropy loss only (CrossEntropy); (ii) another fully-supervised model that is trained with both a supervised contrastive loss and a cross-entropy loss (SupConLoss); (iii) a state-of-the-art semi-supervised learning method based on a teacher-student network, Meta Pseudo Labels (MPL) [16], which outperforms other frameworks such as BYOL [7] or Noisy Student [22].

All models use the same encoder – a ResNet18 [8] pre-trained on ImageNet without the final classification layer. The output of dimension 512 is fed into the embedder which consists of a linear layer without activation and batch normalization layer. The embedder is followed by a linear classifier. MPL uses the same architecture for both student and teacher networks. We employ the following training script for both S5CL and MPL: For each unlabeled batch sampled from the unlabeled dataloader, a labeled batch is sampled from the labeled dataloader.

We tune the hyperparameters, including all temperatures and batch sizes, for each model separately using an internal validation set that differs from the test set. Afterward, we fix the parameters and report the mean $\pm$ std over 5 runs. In particular, for MPL training, we use Adam with a learning rate of $1e-5$ for the student and a learning rate of $3e-5$ for the teacher. The weight decay is always $1e-4$. On NCT-CRC-HE-100K, the remaining models work well with Adamax while Adam is better on Munich AML Morphology. The learning rate and weight decay are set to $1e-4$.

3 Results and Discussions

3.1 Evaluation on the Colon Cancer Histology Dataset

The NCT-CRC-HE-100K dataset¹ contains 100,000 non-overlapping image patches from H&E-stained WSIs of colorectal cancer and normal tissue. The

¹ <https://zenodo.org/record/1214456#.YhZI1ZY0-Uk>.

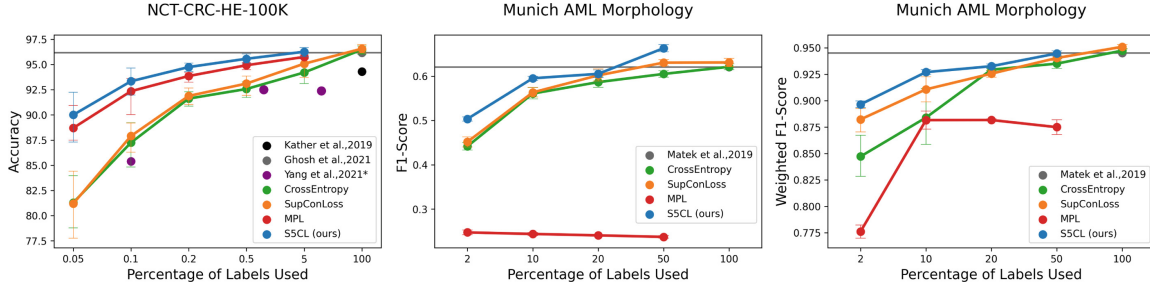


Fig. 2. Results for (a) NCT-CRC-HE-100K and (b, c) Munich AML Morphology. Percentages refer to the full dataset for NCT-CRC-HE-100K and to the majority classes in the case of Munich AML Morphology. Models depicted with a gray line are SOTA supervised models in the literature (Ghosh et al. [6] and Matek et al. [13]). *Yang et al. [23] is a SOTA semi-supervised baseline on the colon cancer dataset. They train with 8 labels without the background, here we assume 100% accuracy on that class.

images are 224×224 pixels large at 0.5 mpp and are color-normalized using Macenko’s method. They are annotated into nine well-balanced classes.

We split the 100,000 images into a training set of 91,000 images and a validation set of 9,000 images (1,000 per class). Next, the original dataset is split into a labeled set containing 5, 10, 20, 50, and 500 images per class which are approximately 0.05%, 0.1%, 0.2%, 0.5%, and 5% of all labeled images. The remaining images are assigned to the unlabeled set. For the test set, we use CRC-VAL-HE-7K, an independent dataset with 7,180 patches from an external cohort [9].

The CrossEntropy and SupConLoss models are trained for 130, 30, and 5 epochs on the sets with 0.05%–0.2%, 0.05%–5%, and 100% labeled images. Both S5CL and MPL are trained for five epochs. The unlabeled set has a batch size of 128 while the labeled batch size is 32 (for 100% labeled images, the batch size is also 128). All the embedding spaces have a dimension of 64. The supervised temperature is set to 0.2, and both the self-supervised as well as the semi-supervised temperature are set to 0.7 (see Ablations 3.3 for details).

As shown in Fig. 2, on all reduced labeled sets, S5CL achieves higher accuracy than the three competitive models including MPL. With only 50 labeled images per class (0.5%), S5CL has an accuracy of 95%, which is already comparable to the two reference models in the literature [9] and [6] that are trained with the full dataset. With 500 labeled images per class (5%), S5CL almost reaches the accuracy of the full SupConLoss model with 96.6% accuracy.

In addition to improving the classification accuracy on small labeled datasets, S5CL also makes the feature embedding space more compact and explicable. To see this, we plot the feature representations as a UMAP in Fig. 3 and measure the embedding quality with the MAP@R score [14], where 100% implies perfect clustering and separation. S5CL achieves a MAP@R of 74% whereas CrossEntropy has a score of 67%. In Fig. 3, we can see that CrossEntropy, might split the representation of the same class (e.g. DEB = debris) into two different clusters – unlike S5CL. Within a cluster, S5CL also achieves more robust embeddings, i.e., images are ordered hierarchically: weakly augmented images and similar images

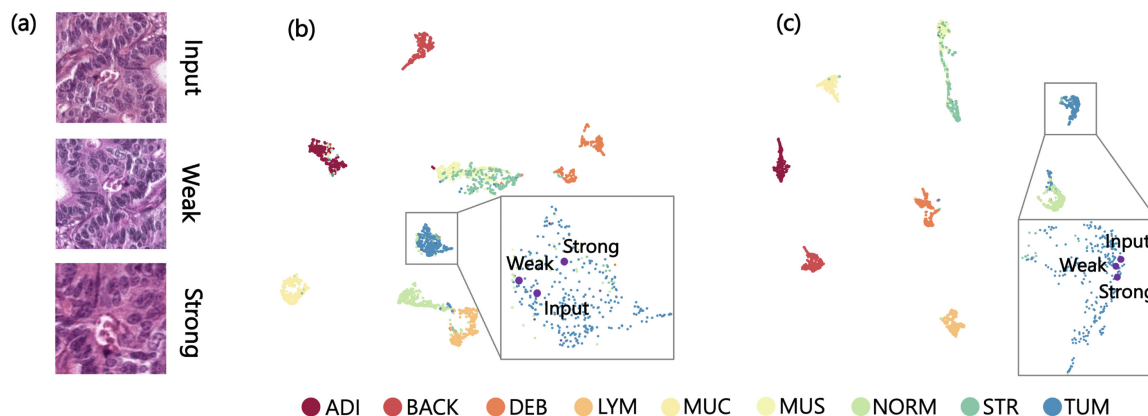


Fig. 3. UMAP of feature embeddings for the test set with 300 sampled images per class including weak and strong augmentations using (b) CrossEntropy and (c) S5CL.

are embedded the closest to their origins, then come strong augmentations as well as different looking images from the same class.

3.2 Evaluation on the Leukemia Single-Cell Dataset

The Munich AML Morphology dataset² contains 18,365 labeled single-cell images taken from peripheral blood smears of 100 patients diagnosed with acute myeloid leukemia (AML). All smears are scanned at $100\times$ magnification with oil immersion at a resolution of 14.14 mpp and have a patch size of 400×400 pixels. The dataset’s classes are heavily imbalanced, ranging from 8,484 images in the largest class to 11 images in the smallest class. In addition, some of the 15 cell types are morphologically similar as they are biologically related.

The dataset is split into a training set of size 11,025; a validation set of size 3,666; and a test set of size 3,674. The training set is further split into a labeled and unlabeled set. To preserve the class distribution, we sample 2%, 10%, 20%, 50%, and 100% images from the large classes for the labeled set and keep all labeled images from small classes that have less than 100 instances per class.

We train the CrossEntropy and SupConLoss models with a labeled batch size of 128; S5CL in addition has an unlabeled batch size of 64. Early stopping is used to avoid overfitting as in [13]. The number of epochs is thus 40, 25, 20, 15, and 15, respectively for each data subset. MPL models use smaller labeled and unlabeled batches of size 32 since training is very memory intensive. However, the batch size does not have a large effect on the results. Due to overfitting, we train the models for two epochs. In addition, we have to include the validation set into the training set, as otherwise, the student model would never see images from the minority class [16]. But we note the class proportions are still the same. For all models, the embedding dimension is 256. The supervised temperature is always set to 0.1 and the self-supervised, as well as the semi-supervised one, are both set to 0.6.

² <https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=61080958>.

As shown in Fig. 2, S5CL also outperforms CrossEntropy and SupConLoss for all reduced labeled datasets. Remarkable is that S5CL also exceeds the full models in terms of F1-score, indicating that S5CL is less biased toward the large classes and can also achieve good performance within the minority classes. It is also worth noting that MPL does not perform well on this dataset, which could be due to a confirmation bias toward the majority classes, as described in [21].

3.3 Ablation Study

To investigate the contribution of each loss, we first start from a model trained with supervised SupConLoss and progressively build two intermediate models, S1CL and S3CL, to arrive at S5CL (please refer to Fig. 1a). We can see that as the model complexity grows, so does the test accuracy (Fig. 4a). Next, we analyze the effect of different augmentation techniques. We start with no augmentations, add weak augmentations and then vary different types of strong augmentations. Our results suggest that both weak and strong augmentations are important and different types of strong augmentations can change the accuracy by $\sim 3\%$ (Fig. 4b). We also investigate the choice of the self-supervised temperature T_U with respect to the supervised temperature T_L . We set $T_L = 0.2$ and vary the temperature T_U . As expected, a higher distance to T_L resolves the problem of conflicting loss objectives (Fig. 4c). The last ablation analyzes the effect of pseudo-labels. If no pseudo-labels are used after the first epoch, the accuracy drops (Fig. 4d). Also, if we use pseudo-labels for the classification loss, the accuracy decreases, possibly due to flawed pseudo-labels.

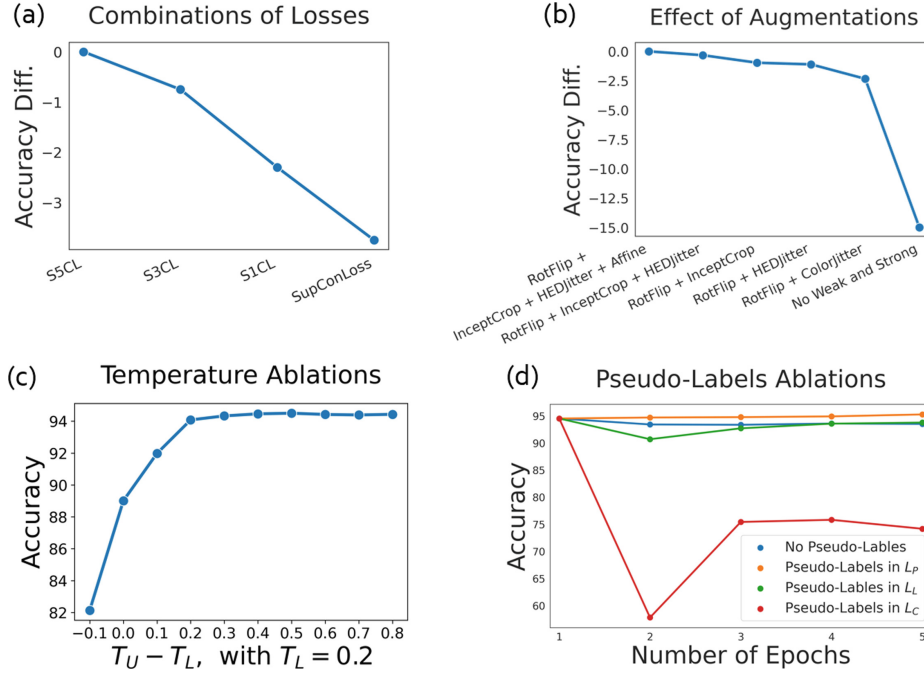


Fig. 4. Ablation study for different choices of hyperparameters and architectures.

4 Conclusion

We propose S5CL as a unified framework for supervised, self-supervised, and semi-supervised contrastive learning. It combines the cross-entropy loss with three contrastive losses for labeled, unlabeled and pseudo-labeled data. In particular, by altering the temperature of the SupConLoss, S5CL can simultaneously train supervised and self-supervised on both labeled and unlabeled data. We evaluate S5CL on two public datasets where it outperforms several baseline methods including Meta Pseudo Labels, a SOTA semi-supervised method. Due to its easy implementation and flexibility, we believe that S5CL can be employed in many problems in computational pathology with sparsely labeled datasets.

Acknowledgements. S.J.W. is supported by the Helmholtz Association under the joint research school “Munich School for Data Science - MUDS”.

References

1. Arazo, E., Ortego, D., Albert, P., O’Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. Eprint [arXiv:1908.02983](https://arxiv.org/abs/1908.02983) (2020)
2. Bokhorst, J.M., Pinckaers, H., van Zwam, P., Nagtegaal, I., van der Laak, J., Ciompi, F.: Learning from sparsely annotated data for semantic segmentation in histopathology images. In: Cardoso, M.J., et al. (eds.) Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning, pp. 84–91. Proceedings of Machine Learning Research (2019)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. Eprint [arXiv:2002.05709](https://arxiv.org/abs/2002.05709) (2020)
4. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. Eprint [arXiv:2006.10029](https://arxiv.org/abs/2006.10029) (2020)
5. Fu, Y., et al.: Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat. Cancer **1**(8) (2020). <https://www.nature.com/articles/s43018-020-0085-8>
6. Ghosh, S., Bandyopadhyay, A., Sahay, S., Gosh, R., Kundu, I., Santosh, K.: Colorectal histology tumor detection using ensemble deep neural network. Eng. Appl. Artif. Intell. **100**, 104202 (2021). https://www.sciencedirect.com/science/article/abs/pii/S095219762100049X?dgcid=rss_sd_all
7. Grill, J.B., et al.: Bootstrap your own latent: a new approach to self-supervised learning. Eprint [arXiv:2006.07733](https://arxiv.org/abs/2006.07733) (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
9. Kather, J.N., et al.: Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med. **16**(1), 1–22 (2019). <https://doi.org/10.1371/journal.pmed.1002730>
10. Khosla, P., et al.: Supervised contrastive learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 18661–18673. Curran Associates, Inc. (2020). <https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf>

11. Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N.: Self-path: self-supervision for classification of pathology images with limited annotations. *IEEE Trans. Med. Imaging* **40**(10), 2845–2856 (2021). <https://doi.org/10.1109/TMI.2021.3056023>
12. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: a framework and review. Eprint [arXiv:2010.05113](https://arxiv.org/abs/2010.05113) (2020)
13. Matek, C., Schwarz, S., Spiekermann, K., Marr, C.: Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks. *Nat. Mach. Intell.* **1**, 538–544 (2019). <https://doi.org/10.1038/s42256-019-0101-9>
14. Musgrave, K., Belongie, S.J., Lim, S.: A metric learning reality check. Eprint [arXiv:2003.08505](https://arxiv.org/abs/2003.08505) (2020)
15. Oh, Y., Kim, D., Kweon, I.S.: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. Eprint [arXiv:2106.05682](https://arxiv.org/abs/2106.05682) (2020)
16. Pham, H., Dai, Z., Xie, Q., Le, Q.V.: Meta pseudo labels. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11552–11563 (2021). <https://doi.org/10.1109/CVPR46437.2021.01139>
17. Purushwalkam, S., Gupta, A.: Demystifying contrastive self-supervised learning: invariances, augmentations and dataset biases. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) *Advances in Neural Information Processing Systems*, vol. 33, pp. 3407–3418. Curran Associates, Inc. (2020). <https://proceedings.neurips.cc/paper/2020/file/22f791da07b0d8a2504c2537c560001c-Paper.pdf>
18. Sahito, A., Frank, E., Pfahringer, B.: Semi-supervised learning using Siamese networks. In: Liu, J., Bailey, J. (eds.) *AI 2019. LNCS (LNAI)*, vol. 11919, pp. 586–597. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35288-2_47
19. Sohn, K., et al.: FixMatch: simplifying semi-supervised learning with consistency and confidence. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608. Curran Associates, Inc. (2020). <https://proceedings.neurips.cc/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf>
20. Szegedy, C., et al.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015). <https://doi.org/10.1109/CVPR.2015.7298594>
21. Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F.: CReST: a class-rebalancing self-training framework for imbalanced semi-supervised learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10852–10861 (2021). <https://doi.org/10.1109/CVPR46437.2021.01071>
22. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves ImageNet classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695 (2020). <https://doi.org/10.1109/CVPR42600.2020.01070>
23. Yang, P., Hong, Z., Yin, X., Zhu, C., Jiang, R.: Self-supervised visual representation learning for histopathological images. In: de Bruijne, M., et al. (eds.) *MICCAI 2021. LNCS*, vol. 12902, pp. 47–57. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_5
24. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: self-supervised learning via redundancy reduction. Eprint [arXiv:2103.03230](https://arxiv.org/abs/2103.03230) (2020)
25. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4L: self-supervised semi-supervised learning. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1476–1485 (2019). <https://doi.org/10.1109/ICCV.2019.00156>

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Dec 16, 2024

This Agreement between Helmholtz Munich ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number 5930960720485

License date Dec 16, 2024

Licensed Content Publisher Springer Nature

Licensed Content Publication Springer eBook

Licensed Content Title S5CL: Unifying Fully-Supervised, Self-supervised, and Semi-supervised Learning Through Hierarchical Contrastive Learning

Licensed Content Author Manuel Tran, Sophia J. Wagner, Melanie Boxberg et al

Licensed Content Date Jan 1, 2022

Type of Use Thesis/Dissertation

Requestor type academic/university or research institute

Format print and electronic

Portion full article/chapter

Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	Generalizable and Data- efficient Deep Learning for Computer-aided Cancer Diagnosis
Institution name	Technical University Munich
Expected presentation date	Feb 2025
The Requesting Person / Organization to Appear on the Helmholtz Munich License	
	Sophia Wagner 
Requestor Location	 Germany
Billing Type	Invoice
Billing Address	Sophia Wagner Ingolstädter Landstraße 1 Neuherberg, Germany 85764
Total	0.00 EUR
Terms and Conditions	

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed

Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at <https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/>.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a	

journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: *'Reproduced with permission from Springer Nature'*.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licenser warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licenser and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licenser reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany's choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit <https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution>

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

A.4. Low-resource finetuning of foundation models beats state-of-the-art in histopathology

This publication was published in the proceedings of the peer-reviewed international conference 2024 IEEE International Symposium on Biomedical Imaging (ISBI). The one-page summary can be found in Section 3.4.

Roth, B.*, Koch, V.*, Wagner, S. J.*, Schnabel, J. A., Marr, C., & Peng, T. (2024). Low-resource finetuning of foundation models beats state-of-the-art in histo-pathology. *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. *Contributed equally

Copyright: Copyright © 2024, IEEE. Reprinted, with permission, from Roth, B., Koch, V., Wagner, S. J., Schnabel, J. A., Marr, C., Peng, T., Low-resource finetuning of foundation models beats state-of-the-art in histo-pathology, 2024 IEEE International Symposium on Biomedical Imaging (ISBI), May 2024. The copyright clearance is attached after the original publication.

LOW-RESOURCE FINETUNING OF FOUNDATION MODELS BEATS STATE-OF-THE-ART IN HISTOPATHOLOGY

Benedikt Roth ^{1,2,*}, Valentin Koch ^{1,2,*}, Sophia J. Wagner ^{1,2,*}, Julia A. Schnabel ^{1,2,3}
Carsten Marr ^{2,†}, Tingying Peng ^{2,†}

¹ School of Computation and Information Technology, Technical University of Munich, Germany

² Helmholtz Zentrum München - German Research Center for Environmental Health, Germany

³ School of Biomedical Engineering and Imaging Sciences, King's College London, UK

ABSTRACT

To handle the large scale of whole slide images in computational pathology, most approaches first tessellate the images into smaller patches, extract features from these patches, and finally aggregate the feature vectors with weakly-supervised learning. The performance of this workflow strongly depends on the quality of the extracted features. Recently, foundation models in computer vision showed that leveraging huge amounts of data through supervised or self-supervised learning improves feature quality and generalizability for a variety of tasks. In this study, we benchmark the most popular vision foundation models as feature extractors for histopathology data. We evaluate the models in two settings: slide-level classification and patch-level classification. We show that foundation models are a strong baseline. Our experiments demonstrate that by finetuning a foundation model on a single GPU for only two hours or three days depending on the dataset, we can match or outperform state-of-the-art feature extractors for computational pathology. These findings imply that even with little resources one can finetune a feature extractor tailored towards a specific downstream task and dataset. This is a considerable shift from the current state, where only few institutions with large amounts of resources and datasets are able to train a feature extractor. We publish all code used for training and evaluation as well as the finetuned models ¹.

Index Terms— Self-supervised learning, foundation models, medical imaging, histopathology

1. INTRODUCTION

Recent progress in natural language processing, in particular the success of large language models [1], has shown that huge data sources can be leveraged to build models that generalize well to a wide variety of tasks. These models,

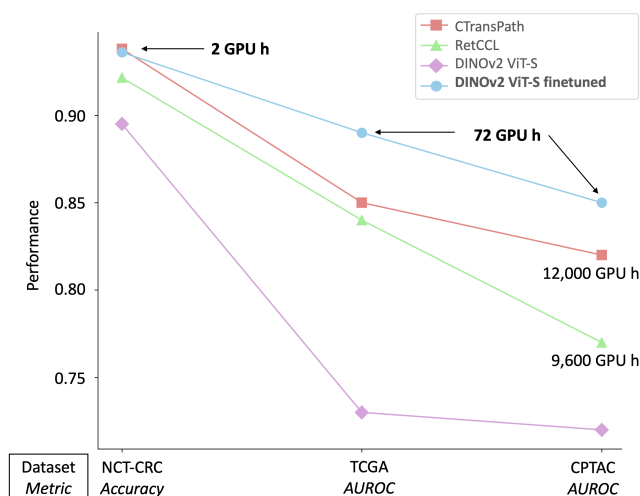


Fig. 1. We propose finetuning a DINOv2 ViT-S, which yields at least equal performance compared to CTransPath and RetCCL but in a fraction of domain specific training time. Performance is measured on three datasets: TCGA & CPTAC (WSI-level classification) and NCT-CRC (patch-level classification).

also called foundation models [2], are trained on large-scale datasets without task-specific supervision and can be adapted for specialized downstream applications.

Similar approaches have been developed for computer vision, where the most influential aspects for a good performance lie in large-scale datasets and self-supervised learning techniques. The core concept is to efficiently represent complex visual data in a lower-dimensional space. One common approach is contrastive learning, which involves pushing similar instances closer in the learned representation space while separating dissimilar ones [3]. Training networks with contrastive loss often needs a large batch size and has the risk to collapse to one representation. To account for this, non-contrastive methods, such as MoCov3 [4], were proposed. MoCov3 creates a dynamic dictionary of negative examples

* contributed equally to this work

† corresponding author, carsten.marr@helmholtz-munich.de

† corresponding author, tingying.peng@helmholtz-munich.de

¹<https://github.com/beneroth13/dinov2>

using a momentum update mechanism. A similar approach is student-teacher network training, where the student only gets a masked version of the input and tries to learn similar representations as the teacher network. The student is updated during training whereas the teacher is updated with the exponential moving average of the student weights [5]. This strategy can be used for knowledge distillation, where the knowledge from a larger model is transferred to a smaller one.

Translating these efforts to medical tasks would be highly beneficial but is challenging for several reasons: Large-scale and publicly available data sources are scarce, imaging modalities and downstream tasks are very diverse, high performance on very specific tasks is preferred over a good performance on a broad set of tasks, and, finally, the patients' privacy needs to be preserved. Nevertheless, a few models were proposed as potential foundation models that are trained on medical imaging data, but not extensively on histopathology data [6, 7].

In computational pathology, self-supervised feature extractors are commonly used to compute a low-dimensional representation of a whole slide image that was tessellated into smaller patches [8, 9, 10]. These low-dimensional features are then usually aggregated for a downstream task using weakly-supervised learning. The largest dataset used for pre-training these feature extractors comprises 15 million patches [10]. However, this is still magnitudes smaller than the large-scale pre-training datasets used in classical computer vision, e.g., 142 million images curated from a dataset of 1.2 billion images for DINOv2 [11]. There has been a recent effort to scale-up the pre-training for histopathology-specific feature extractors using private datasets [12, 13, 14] but none of the models is currently publicly available and could be included in this study.

Chen et al. [12] report using 32 NVIDIA A100 80GB GPUs and a batch size of 3,072 for 125,000 iterations without specifying the exact training time. CTransPath was trained for 250 hours on 48 NVIDIA V100 GPUs [10]. RetCCL [9] used 32 NVIDIA V100 32G GPU for 300 hours. Virchow [13] and REMEDIS [6] do not specify their setup but the scale of the data and the trained model suggest a similar compute heavy setup. As the compute resources needed to train such a model is out of reach for many institutions, only publicly available models such as CTransPath [10] or RetCCL [9] are used.

In our work, we propose a shift towards low-resource fine-tuning on specific datasets and show its applicability. We present a benchmark of four of the largest foundation models in classical computer vision on histopathology downstream tasks and compare the models to the state-of-the-art domain-specific feature extractors CTransPath and RetCCL. We fine-tune ViT models of the best performing self-supervised learning method DINOv2 on a single GPU and show that fine-tuning foundation models on task-specific data can outperform large-scale trained domain-specific models while needing only a fraction of resources and training time (Figure 1).

2. EXPERIMENTS & METHODS

Our comparative analysis includes ResNet50 [15], pre-trained on ImageNet and its truncated variant as baseline models. We benchmark four vision foundation models that have been trained on natural images: ImageBind, a multi-modal model trained on paired data consisting of images and a corresponding second modality such as text or depth-maps [16]; Segment Anything (SAM), a segmentation model, trained with supervision on a large amount of images with segmentation masks [17]; BEiT, a vision transformer (ViT) trained with self-supervision based on masked image modeling [18]; DINOv2, a self-supervised teacher-student model trained on a large curated dataset [11]. As histopathology-specific feature extractors, we evaluate the state-of-the-art models CTransPath [10] and RetCCL [9], which are specifically designed for patches from whole-slide images and are trained on 15 million histopathology patches.

2.1. Experimental setup

We evaluate the methods on three datasets of colorectal cancer tissue, two with annotations for slide-level classification and one with patch-level annotations.

Slide-level classification. The public repository The Cancer Genome Atlas (TCGA) [19] contains whole-slide images of 632 patients with colorectal cancer (CRC) from the cohorts COAD and READ with annotations of microsatellite instability (MSI). We perform 5-fold cross validation using three folds for training, one for validation, and one for testing. As an external test set, we use the CRC cohort of the public database The Clinical Proteomic Tumor Analysis Consortium, (CPTAC) [20] with 110 patients.

We tessellate the whole slide images at $10\times$ magnification in patches of size 256×256 pixels (px) and extract the features². Subsequently, we train a slide-level classifier using transformer-based aggregation following Wagner et al. [21].

Patch-level classification. The NCT-CRC-100K (NCT) dataset [22] contains 100,000 non-overlapping image patches from H&E-stained WSIs from nine classes of colorectal cancer and normal colon tissue. The images are of size 224×224 pixels (px) at $0.5 \mu\text{m}/\text{px}$ magnification and are color-normalized using Macenko's method [23]. We use the full dataset for training. For testing, we use CRC-VAL-HE-7K [22], an independent dataset of 7,180 patches from an external cohort with the same tissue classes.

We extract the features using the introduced models as feature extractors and evaluate the classification using the sklearn K -nearest neighbors implementation with $K = 20$ as well as the sklearn implementation of logistic regression classifier with 1,000 iterations and ℓ_2 -regularization coefficient λ of $\frac{100}{MC}$, where M is the embedding dimension and C

²Code for the pre-processing pipeline available at https://github.com/peng-lab/HistoBistro/tree/feature_extraction.

Model	Histo	AUROC	
		TCGA	CPTAC
ResNet50		0.67±0.10	0.65±0.05
ResNet50 truncated		0.68±0.02	0.68±0.02
SAM (ViT-B)		0.56±0.07	0.62±0.03
SAM (ViT-H)		0.55±0.05	0.64±0.02
BEiT (ViT-B)		0.52±0.07	0.51±0.02
ImageBind (ViT-h)		0.64±0.06	0.71±0.02
DINOv2 (ViT-S)		0.73±0.07	0.72±0.06
DINOv2 (ViT-g)		0.66±0.12	0.60±0.03
RetCCL (ResNet50)	✓	0.84±0.08	0.77±0.05
CTransPath (Swin-T)	✓	0.85±0.05	0.82±0.04
DINOv2 (ViT-S)	✓	0.89 ±0.05	0.85 ±0.02
DINOv2 (ViT-g)	✓	0.84±0.05	0.79±0.03

Table 1. Finetuned DINOv2 outperforms dedicated feature extractors CTransPath and RetCCL on AUROC scores on WSI classification (MSI detection) in CRC tissue from TCGA and CPTAC (external). Also, DINOv2 ViT-S notably outperforms the ViT-g variant.

is the number of classes. The choice of hyperparameters and evaluation methods follows Chen et al. [24].

2.2. Finetuning and implementation details

To evaluate performance improvements of the foundation models by finetuning, we focus on DINOv2 in the smallest and largest variants ViT-S/14 (21M parameters, Table 3) and ViT-g/14 (1.1B parameters), respectively. We finetune both models using the original DINOv2 implementation³ on the NCT and TCGA dataset. As a starting point, the publicly available pretrained weights for the backbones were loaded into both the student and the teacher network. Since the weights for the DINO-heads are not publicly available, a random initialization is used. As an additional augmentation we add vertical flipping.

Similar as proposed in [11], we use input images of size 224×224 px with a local crop size of 98×98 px. In order to adapt the positional encoding from its original size of 518×518 px to the desired size of 224×224 px, we employ bicubic interpolation.

In our finetuning experiments, we utilize the provided configuration file from the training pipeline with a base learning rate of 2×10^{-4} . To account for the comparatively small datasets, we adjust the number of iterations per pseudo-epoch and the number of pseudo-epochs. As batch-size we use 256 for ViT-S and 32 for ViT-g. Using a smaller batch size of 128 for ViT-S yielded similar results as with batch size 256. In the KoLeoLoss the hyperparameter ϵ was changed from 1×10^{-8} to 1×10^{-4} to avoid infinite values in the loss. The remaining

³<https://github.com/facebookresearch/dinov2>

Model	Histo	20-NN		linear probe	
		ACC	F1	ACC	F1
ResNet50		0.78	0.81	0.86	0.88
ResNet50 truncated		0.87	0.91	0.88	0.89
SAM (ViT-B)		0.75	0.76	0.83	0.81
SAM (ViT-H)		0.69	0.71	0.79	0.84
BEiT (ViT-B)		0.63	0.66	0.59	0.65
ImageBind (ViT-h)		0.87	0.89	0.90	0.91
DINOv2 (ViT-S)		0.88	0.90	0.90	0.92
DINOv2 (ViT-g)		0.91	0.93	0.92	0.94
RetCCL (ResNet50)	✓	0.91	0.93	0.92	0.94
CTransPath (Swin-T)	✓	0.95	0.96	0.94	0.95
DINOv2 (ViT-S)	✓	0.94	0.95	0.93	0.94
DINOv2 (ViT-g)	✓	0.93	0.95	0.94	0.96

Table 2. Finetuned DINOv2 matches the performance of the best-scoring feature extractor, CTransPath, on balanced accuracy (ACC) and weighted F1-scores (F1) for CRC tissue classification on NCT-patches.

hyperparameters are kept consistent with the values specified in the configuration file provided in the official implementation. Training was done on a single NVIDIA A100 GPU (80GB).

Slide-level classification. For finetuning on TCGA, we tessellate the whole slide image into patches of size 512×512 px at $20\times$ magnification. We thereby obtain a training dataset of 2.5M patches, which are downsampled using RandomResizeCrop to 224×224 px as input to the vision transformer. We set the iterations per epoch to 500 and the number of pseudo-epochs to 200.

Patch-level classification. We use the full training dataset NCT of 100,000 images for finetuning and set the iterations per epoch to 100 and the number of pseudo-epochs to 100, respectively.

2.3. Results

Slide-level classification. In our experiments, DINOv2 outperforms all other computer vision foundation models, i.e., SAM, BEiT, and ImageBind, in both the small (ViT-S) and giant (ViT-g) variant (Table 1). Notably, the giant variant performs substantially worse than the small variant (e.g., 0.66 vs. 0.73 on TCGA, 0.6 vs. 0.72 on CPTAC). The performance of the larger network is surpassed by the truncated version of ResNet50 trained on ImageNet, which consistently outperforms its full counterpart. Interestingly, the features from ImageBind lead to a strong generalization on CPTAC, where the external validation almost matches the performance of DINOv2 (0.71 vs. 0.72) despite a low performance on the in-domain test set TCGA (0.64). These models are all outperformed by the histopathology feature extractors, RetCCL and CTransPath, showing that feature quality benefits from

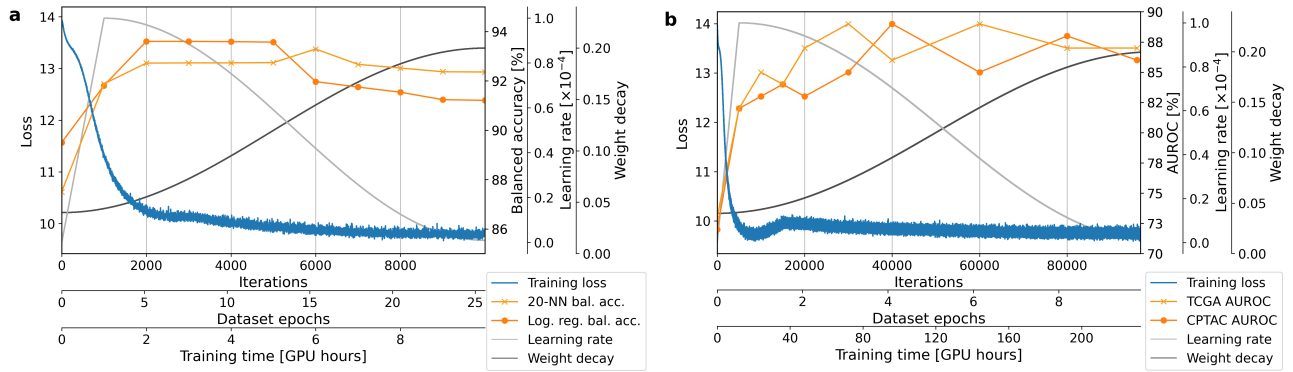


Fig. 2. Performance over time of finetuning a ViT-s with DINOv2: a) on NCT-CRC and evaluating on the external NCT-CRC testset on patch-level classification and b) on TCGA and testing on TCGA (5-fold cross-validation) and CPTAC (external testset) on WSI-level classification.

training on histopathology-specific data. However, finetuning DINOv2 on our 2.5M patches from TCGA further improves the performance and outperforms the state-of-the-art feature extractors, where the smaller variant based on ViT-S performs better than the large version. Compared to CTransPath we only utilize 0.6% (72 h vs. 12,000 h) of GPU hours to reach peak performance for training the ViT-S models (Figure 1). Peak performance is reached at approximately 30,000 to 40,000 iterations for ViT-S compared to around 60,000 for ViT-g (Figure 2).

Patch-level classification. As for slide-level classification, DINOv2 outperforms all other computer vision foundation models, where the margin to ImageBind and the truncated ResNet50 version is small. Histopathology-specific pre-training improves also in this case, and CTransPath achieves best results accross all metrics (Table 2). However, RetCCL obtains an almost identical performance. Fine-tuning both DINOv2 variants also almost match the performance of CTransPath with only two GPU hours (0.08% of CTransPath’s) of training (Figure 1). The peak performance is approximately reached at 2,000 iterations for ViT-S compared to around 10,000 for ViT-g (Figure 2).

In both tasks, all foundation models show a very similar behavior: ImageBind and the truncated ResNet50 follow DINOv2 regarding the performance, where ImageBind has roughly 20 times more parameters (Table 3). Conversely, both variants of SAM as well as BEiT underperform compared to other models. The training duration for NCT is significantly shorter than for TCGA due to the smaller number of slides in the training set. To ensure no overfitting is happening, we evaluate each task on an external test set, CPTAC and CRC-VAL-HE-7K, respectively, as well as the in-domain TCGA test set.

Model	#params	feature size
ResNet50	25.6 M	1024
ResNet50 truncated	8.5 M	1024
SAM (ViT-B)	91 M	256
SAM (ViT-H)	636 M	256
BEiT (ViT-B)	85.7 M	768
ImageBind (ViT-h)	630 M	1024
RetCCL (ResNet50)	23.5 M	2048
CTransPath (Swin-T)	27.5 M	768
DINOv2 (ViT-S)	21 M	384
DINOv2 (ViT-g)	1.1 B	1536

Table 3. Parameter count and feature dimension of benchmarked models.

3. CONCLUSION

In our study, we benchmarked four of the most popular publicly available state-of-the-art foundations models in the field of computer vision as feature extractors. By fine-tuning the most promising model with the DINOv2 framework, we achieved notable improvements that are comparable or better than the current state-of-the-art feature extractor in histopathology CTransPath, while only using 0.08% or 0.6% of their compute budget, which amounts to two hours or three days of training on one single A100 GPU depending on the dataset. Our experiments were conducted with a limited scope, utilizing two datasets for training and three for testing purposes. To fully conclude a superior performance compared to established feature extractors, more experiments are needed on more diverse benchmarks. Using the provided code, further experiments can be conducted in the future. These may include exploring different benchmark datasets, and tasks like segmentation or detection.

4. ETHICS STATEMENT

This research study was conducted retrospectively using human subject data made available in open access. Ethical approval was not required as confirmed by the license attached with the open access data.

5. ACKNOWLEDGMENTS

V.K. and S.J.W. were supported by the Helmholtz Association under the joint research school “Munich School for Data Science - MUDS”. S.J.W. was supported by the Add-on Fellowship of the Joachim Herz Foundation. This work was also supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). C.M. has received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation program (grant agreement number 866411) and is supported by the Hightech Agenda Bayern. The results shown here are in part based upon data generated by the TCGA Research Network ⁴.

6. REFERENCES

- [1] Wayne Xin Zhao et al., “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023.
- [2] Rishi Bommasani et al., “On the opportunities and risks of foundation models,” 2022.
- [3] Ting Chen et al., “A simple framework for contrastive learning of visual representations,” 2020.
- [4] Xinlei Chen et al., “An empirical study of training self-supervised vision transformers,” 2021.
- [5] Chengming Hu et al., “Teacher-student architecture for knowledge learning: A survey,” 2022.
- [6] Shekoofeh Azizi et al., “Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging,” *Nature Biomedical Engineering*, pp. 1–24, 2023.
- [7] Michael Moor et al., “Med-flamingo: a multimodal medical few-shot learner,” 2023.
- [8] Abtin Riasatian et al., “Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides,” *Medical Image Analysis*, vol. 70, pp. 102032, 2021.
- [9] Xiyue Wang et al., “Retccl: clustering-guided contrastive learning for whole-slide image retrieval,” *Medical image analysis*, vol. 83, pp. 102645, 2023.
- [10] Xiyue Wang et al., “Transformer-based unsupervised contrastive learning for histopathological image classification,” *Medical Image Analysis*, vol. 81, pp. 102559, 2022.
- [11] Maxime Oquab et al., “Dinov2: Learning robust visual features without supervision,” 2023.
- [12] Richard J. Chen et al., “A general-purpose self-supervised model for computational pathology,” 2023.
- [13] Eugene Vorontsov et al., “Virchow: A million-slide digital pathology foundation model,” *arXiv preprint arXiv:2309.07778*, 2023.
- [14] Gabriele Campanella et al., “Computational pathology at health system scale—self-supervised foundation models from three billion images,” *arXiv preprint arXiv:2310.07033*, 2023.
- [15] Kaiming He et al., “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015.
- [16] Rohit Girdhar et al., “Imagebind: One embedding space to bind them all,” 2023.
- [17] Alexander Kirillov et al., “Segment anything,” 2023.
- [18] Hangbo Bao et al., “Beit: Bert pre-training of image transformers,” 2022.
- [19] Cancer Genome Atlas Network et al., “Comprehensive molecular characterization of human colon and rectal cancer,” *Nature*, vol. 487, no. 7407, pp. 330, 2012.
- [20] Nathan J. Edwards et al., “The cptac data portal: A resource for cancer proteomics research,” *Journal of Proteome Research*, vol. 14, no. 6, pp. 2707–2713, 2015, PMID: 25873244.
- [21] Sophia J. Wagner et al., “Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study,” *Cancer Cell*, vol. 41, no. 9, pp. 1650–1661.e4, 2023.
- [22] Jakob Nikolas Kather et al., “100,000 histological images of human colorectal cancer and healthy tissue,” May 2018.
- [23] Marc Macenko et al., “A method for normalizing histology slides for quantitative analysis,” in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009, pp. 1107–1110.
- [24] Richard J. Chen et al., “A general-purpose self-supervised model for computational pathology,” 2023.

⁴<https://www.cancer.gov/tcga>



RightsLink



Low-Resource Finetuning of Foundation Models Beats State-of-the-Art in Histopathology

Conference Proceedings: 2024 IEEE International Symposium on Biomedical Imaging (ISBI)

Author: Benedikt Roth

Publisher: IEEE

Date: 27 May 2024

Copyright © 2024, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis online.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK

CLOSE WINDOW

A.5. DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology

This publication was published in the proceedings of the peer-reviewed international conference Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2024. The one-page summary can be found in Section 3.5.

Koch, V.*, Wagner, S. J.*, Kazeminia, S., Sancar, E., Hehr, M., Schnabel, J. A., ... & Marr, C. (2024, October). DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 520-530). Cham: Springer Nature Switzerland.

Copyright: Copyright © 2024, Springer Nature. Reproduced with permission from Springer Nature. The license is attached after the original publication.



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology

Valentin Koch^{1,2,*}, Sophia J. Wagner^{1,2,*}, Salome Kazemina^{1,2}, Ece Sancar^{1,2}, Matthias Hehr^{2,3}, Julia A. Schnabel^{1,2,4}, Tingying Peng^{1,2}, Carsten Marr^{2,✉}

¹ School of Computation, Information and Technology, Technical University of Munich, Munich, Germany

² Computational Health Center, Helmholtz Munich, Munich, Germany

³ Dr. von Haunersches Kinderspital, Ludwig-Maximilians-University Munich, Munich, Germany

⁴ School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

Abstract. In hematology, computational models offer significant potential to improve diagnostic accuracy, streamline workflows, and reduce the tedious work of analyzing single cells in peripheral blood or bone marrow smears. However, clinical adoption of computational models has been hampered by the lack of generalization due to large batch effects, small dataset sizes, and poor performance in transfer learning from natural images. To address these challenges, we introduce DinoBloom, the first foundation model for single cell images in hematology, utilizing a tailored DINOv2 pipeline. Our model is built upon an extensive collection of 13 diverse, publicly available datasets of peripheral blood and bone marrow smears, the most substantial open-source cohort in hematology so far, comprising over 380,000 white blood cell images. To assess its generalization capability, we evaluate it on an external dataset with a challenging domain shift. We show that our model outperforms existing medical and non-medical vision models in (i) linear probing and k -nearest neighbor evaluations for cell-type classification on blood and bone marrow smears and (ii) weakly supervised multiple instance learning for acute myeloid leukemia subtyping by a large margin. A family of four DinoBloom models (small, base, large, and giant) can be adapted for a wide range of downstream applications, be a strong baseline for classification problems, and facilitate the assessment of batch effects in new datasets. All models are available at github.com/marrlab/DinoBloom.

Keywords: Self-supervised learning · Foundation model · Hematology

1 Introduction

Hematology, the study of blood and blood-related diseases, relies heavily on the microscopic examination of peripheral blood and bone marrow smears. This

* Contributed equally

✉ carsten.marr@helmholtz-munich.de

practice is integral to diagnosing hematological diseases, such as acute myeloid leukemia (AML) [15,3]. Currently, differential blood counts still rely on manual cytomorphological analysis of at least 200 individual white blood cells (WBC) per patient, where exact evaluation is crucial for early and precise diagnosis [13]. This labor-intensive process has resisted automation, remaining a domain for trained experts. However, it suffers from significant intra- and inter-expert variability, complicating diagnosis in environments that lack trained personnel [9].

Recent advances in deep learning propose solutions to the challenges in hematology, such as classifying leukemia subtypes from microscopic images [20,18]. However, the transition from manual to automated analysis requires robust models that can deal with limited data, strong batch effects, and largely varying cell phenotypes. In particular, for weakly-supervised settings, such as multiple instance learning (MIL) for patient-level disease prediction, a strong feature extractor for single blood cells is necessary as supervised learning is not possible.

Still, most approaches so far rely on supervised training sets with corresponding datasets. Rastogi et al. [24] train a convolutional neural network (CNN) as a feature extractor on the AML Matek dataset [20] on 18,365 single-cell images from peripheral blood smears. Hehr et al. [12] train a feature extractor fully supervised on a cell classification task on additional data from the same domain and subsequently train the MIL aggregation model.

Large-scale self-supervised training on diverse datasets has transformed the domain of computer vision on natural images [10,5]. In the medical imaging domain, especially, in histopathology, domain specific self-supervised representation learning on large sets of unlabeled images [30,31] has shown to improve downstream tasks in MIL settings [29]. To this end, DINO [5] and its successor DINOv2 [23] have emerged as a pipeline well-suited to train these feature extractors [6,28,8]. However, to date, there is no comparable effort tackling the challenges in the domain of hematology.

We propose DinoBloom (Dino Blood Model), a model family based on vision transformers trained with a customized DINOv2 [23] pipeline to provide rich visual features for single-cell image analysis in hematology. These models are able to extract predictive features even on unseen datasets that can be used for few-shot classification, multiple instance learning, or cell embeddings potentially characterizing disease profiles while offering explainable features for enhanced interpretability. The main contributions of our work are:

- We introduce DinoBloom, the first large-scale self-supervised trained models designed explicitly for single-cell hematology image analysis.
- We assemble the largest cohort in hematology comprising 13 datasets of peripheral blood and bone marrow smears.
- We show that DinoBloom models are effective in capturing diverse visual features of single cells across tasks on both in-domain and out-of-domain datasets for cell-type classification and leukemia subtyping.
- We provide open access to all DinoBloom models as well as the source code and parameters used for training, encouraging the research community to collaboratively build upon our work.

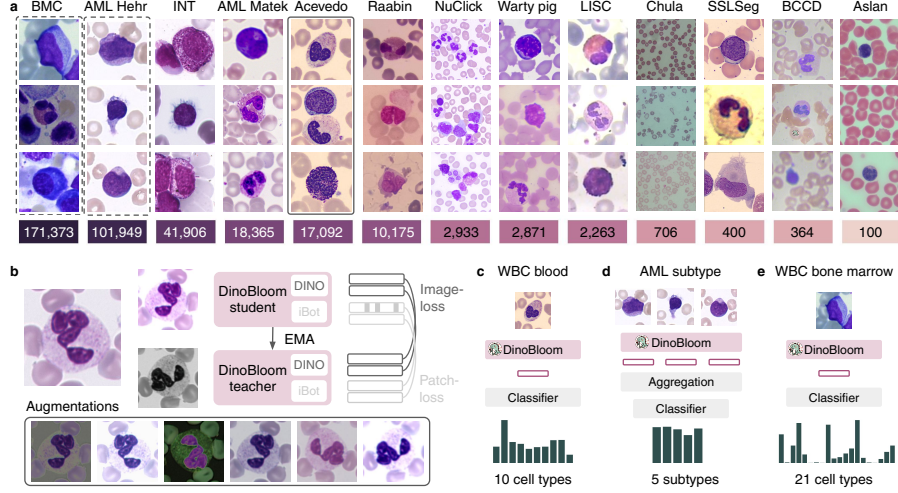


Fig. 1. Data and model overview of our pipeline. (a) All 13 datasets used in this study: dashed lines indicate datasets split into training data for DinoBloom and test data for downstream evaluations, continuous line indicates the dataset was completely held out for testing purposes. (b) Modified DINOv2 pipeline without local crops for model training. We evaluate the performance on three downstream tasks: (c) WBC type classification on the external dataset Acevedo, (d) AML subtyping via multiple instance learning, and (e) bone marrow WBC type classification.

2 Datasets

To the best of our knowledge, we collected the largest hematology cohort with 13 publicly available datasets of in total over 380,000 blood cell images, fusing the domains of peripheral blood and bone marrow smears (Fig. 1a). It consists of the following datasets: The *Bone Marrow Cytomorphology* (BMC) dataset contains 171,373 de-identified, expert-annotated cells from bone marrow smears of 945 patients [18]. In contrast to all other datasets included in this study, the bone marrow smears were stained using the May-Grünwald-Giemsa/Pappenheim stain. The *AML Hehr* dataset includes 81,214 single-cell images from 189 patients, covering four genetic AML subtypes and a healthy control group, sourced from annotated patient-level blood smears. The *INT* dataset is an internal dataset that consists of 41,906 images of peripheral blood smears. The dataset will be made publicly available through a different publication. The *AML Matek* dataset consists of 18,365 expert-labeled single-cell images with 15 heavily imbalanced classes taken from peripheral blood smears of 100 patients diagnosed with AML, as well as 100 healthy patients [19,20]. *Acevedo* encompasses 17,092 images of WBCs labeled into 11 classes [1]. The *Raabin* dataset features 10,175 cropped WBCs, which are labeled by experts into five classes [17]. The *NuClick* dataset was created from 11,000 WBC images of four classes to generate 2,689 images of

Table 1. Training configuration of the DinoBloom model family.

Model	Batch size	Train time	Feature dim	#params
DinoBloom-S	1216	1:30 h	384	22 M
DinoBloom-B	960	0:45 h	768	86 M
DinoBloom-L	448	1:00 h	1024	304 M
DinoBloom-G	208	4:00 h	1536	1136 M

artificially overlapping nuclei [16]. The *Warty pig* dataset contains 1,408 cropped and classified, plus 1,463 augmented WBC images of juvenile Visayan warty pigs [2]. The public part of the *LISC* dataset consists of 157 WBC images as well as several augmented versions of them, totaling 2263 images [25]. *Chula* [22] is a red blood cell segmentation dataset and holds 706 single images. The *SSLSeg* datasets contain 300 images of WBCs, stemming from two different sources (200/100 images) [32]. The *Blood Cell Count and Detection* (BCCD) dataset was created to detect blood cells and includes 364 images with bounding box labels [21]. The *Aslan* blood cell detection dataset offers 100 images of white and red blood cells taken from a light microscope [4].

3 Methods

DINOv2 finetuning. We train our models that are based on vision transformers (ViT) [7] in different sizes using the DINOv2 framework. Following [26], we use the pretrained checkpoints to efficiently finetune the vision transformer on our multi-cohort dataset. The self-supervised learning framework DINOv2 employs a teacher-student architecture with an image-based loss on the class token of the DINO head and a patch-based loss on the class token of the masked patches from the iBot head (Fig. 1b). We remove the global-local crop loss as we found it hampers performance when learning representations on the single-cell images in blood and bone marrow datasets. Images are resized to 224×224 pixels. The models are trained on 8 NVIDIA A100-SXM4-40GB GPUs with an AMD EPYC 7742 64-Core CPU. All models show similar convergence patterns and reached their peak performance between 4,000 and 8,000 iterations, after which downstream task performance drops slightly. Depending on the model and corresponding batch size, 4,000 iterations cover the training set between 1.7 (batch size 208, DinoBloom-G) up to 10 times (batch size 1,216, DinoBloom-S). Used batch size, feature dimension, and training time, as well as the number of parameters, can be inferred from Table 1.

Train and test data. We train the DinoBloom models on all datasets except the Acevedo dataset, which is kept as external test set. The other two datasets used for evaluation were split into train/test (80/20). Only training data was used to train our DinoBloom models. We evaluate the downstream task performance using the same train/test split. The following datasets and settings are used:

Table 2. Evaluation on peripheral blood: Image-level WBC classification on Acevedo dataset and patient-level AML subtyping on AML Hehr dataset. Best results are marked in bold, second best results are underlined. Standard deviation in the ABMIL setting was obtained by 5-fold cross-validation. Performance is measured in weighted F1-score (wF1) and balanced Accuracy (bAcc).

	Acevedo						AML Hehr	
	1-NN		20-NN		Linear probe		ABMIL	
	wF1	bAcc	wF1	bAcc	wF1	bAcc	wF1	bAcc
ResNet 50	58.8	52.6	65.6	58.7	81.3	75.4	81.9 $\pm$ 9.7	81.5 $\pm$ 9.6
ResNet 50 trunc	68.5	62.4	74.0	67.8	87.5	81.6	41.5 $\pm$ 11.7	45.9 $\pm$ 10.0
DINOv2 ViT-S	72.9	65.6	78.5	70.8	87.7	82.0	52.5 $\pm$ 11.8	54.5 $\pm$ 10.0
DINOv2 ViT-B	71.9	64.8	77.3	69.9	87.8	81.8	49.6 $\pm$ 17.3	52.2 $\pm$ 14.3
DINOv2 ViT-L	72.2	64.9	77.8	72.4	89.1	83.5	51.5 $\pm$ 16.6	53.6 $\pm$ 13.5
DINOv2 ViT-G	77.8	70.4	81.9	74.2	90.1	84.5	21.1 $\pm$ 13.5	28.6 $\pm$ 10.2
CTransPath	80.8	73.9	83.1	76.8	88.0	82.5	60.2 $\pm$ 12.6	60.9 $\pm$ 11.8
Phikon ViT-B	83.3	76.5	85.1	78.7	88.2	82.7	81.8 $\pm$ 8.3	81.5 $\pm$ 8.5
DinoBloom-S (ours)	86.4	80.5	90.0	84.5	90.1	84.5	<u>93.0</u> $\pm$ 3.0	<u>92.3</u> $\pm$ 3.4
DinoBloom-B (ours)	87.4	81.9	90.5	85.4	90.7	85.5	92.7 $\pm$ 2.9	91.9 $\pm$ 3.1
DinoBloom-L (ours)	<u>88.9</u>	<u>83.2</u>	<u>91.3</u>	<u>86.1</u>	<u>91.2</u>	<u>86.0</u>	91.7 $\pm$ 2.4	91.0 $\pm$ 2.7
DinoBloom-G (ours)	89.1	83.5	91.4	86.4	91.8	86.6	93.1 $\pm$ 2.5	92.4 $\pm$ 2.8

- Acevedo is the smallest of the peripheral blood datasets with complete image-level classification exhibiting a strong batch effect [27], (Fig. 1a) hence serving as a good measure of the generalization capabilities of our model.
- The AML Hehr dataset is divided into train/test (80/20) on patient level. It includes 101,949 WBCs from 242 patients across four AML subtypes: CBFB::MYH11, NPM1, PML::RARA, and RUNX1::RUNX1T1 and a healthy control class.
- The BMC dataset is split into train/test (80/20) as it is the only bone marrow dataset. It contains 21 heavily imbalanced classes (Fig. 1e).

Downstream evaluations. In all downstream experiments, we compare the following feature extractors: (i) the non-medical-domain models ImageNet-pretrained ResNet50 [11] (full and truncated) and (ii) the pretrained DINOv2 checkpoints, trained on LVD-142M [23]; (iii) the medical-image domain feature extractors CTransPath [31], trained on 14M patches from TCGA and PAIP, (iv) the Phikon ViT-B model [8], trained on PanCancer40M from TCGA; and (v) our models DinoBloom-S, -B, -L, and -G.

We evaluate the performance of all supervised classifier models by linear probe and k -nearest neighbors (k -NN) for cell-type classification and in a weakly-supervised multiple instance learning (MIL) setting for AML subtyping. For linear probe, the sklearn LogisticRegression class is used with l2-regularization coefficient of $\frac{c \times n}{100}$ where n is the number of training samples, and c is the number of classes. For the MIL evaluation, similar to the Hehr et. al [12] framework, we deploy a dedicated classifier head, structured as a two linear layer architecture with an intermediary ReLU activation, tailored to map the aggregated latent

Table 3. Evaluation on bone marrow: WBC classification on the dataset BMC with 21 highly imbalanced classes.

	1-NN			20-NN			Linear probe		
	wF1	Acc	bAcc	wF1	Acc	bAcc	wF1	Acc	bAcc
ResNet 50	37.6	37.3	21.1	47.4	50.0	23.0	64.1	65.2	39.6
ResNet 50 trunc	46.7	46.4	31.2	57.5	59.8	33.0	74.5	75.0	49.8
DINOv2 ViT-S	43.2	43.2	25.0	52.4	55.6	26.5	68.1	69.0	44.9
DINOv2 ViT-B	39.8	39.6	23.9	49.3	55.6	24.1	70.8	71.5	48.5
DINOv2 ViT-L	39.4	39.2	24.5	48.9	52.2	24.0	71.0	71.6	47.8
DINOv2 ViT-G	41.0	41.0	22.6	50.4	53.7	24.4	73.5	74.0	52.1
CTransPath	49.1	48.7	42.0	58.5	60.3	36.1	74.1	74.9	52.2
Phikon ViT-B	47.5	47.2	40.8	57.1	59.0	35.5	73.2	73.8	54.4
DinoBloom-S (ours)	78.4	78.3	<u>62.0</u>	84.2 84.8	55.6		85.7 85.9	71.4	
DinoBloom-B (ours)	<u>79.6</u>	<u>79.5</u>	65.8	83.7	84.1	57.1	<u>85.5</u>	<u>85.6</u>	<u>70.7</u>
DinoBloom-L (ours)	78.8	78.8	57.7	83.6	84.0	<u>56.3</u>	84.9	85.0	64.4
DinoBloom-G (ours)	80.0 79.9	59.4		<u>83.8</u>	<u>84.2</u>	56.2	84.9	85.0	69.3

vectors of a patient to a class prediction.

4 Results

Peripheral blood. DinoBloom models outperform existing models of same size on single WBC classification on the external dataset Acevedo. There is a strong performance gain over the original DINOv2 models, e.g., ViT-S with 71.9 weighted F1-score on 1-NN vs. DinoBloom-S 86.4, ViT-G 77.8 vs. DinoBloom-G 89.1 (Table 2). The histopathology domain-specific feature extractors CTransPath and the recently released Phikon (Vit-B) model perform better than models with non-medical pretraining, both models perform roughly equally in linear probing, while Phikon has a slightly better performance in k-NN evaluations. However, DinoBloom models do not only perform better than their corresponding baseline from DINOv2, but even our smallest model performs better than all other tested models, irrespective of their size. One can also observe that the larger variants of DinoBloom models perform better compared to smaller versions, e.g., DinoBloom-G vs. DinoBloom-B vs. DinoBloom-S with the weighted F1-score on 1-NN (89.1 vs. 87.4 vs. 86.4).

DinoBloom effectively serves as a feature extractor for training a weakly-supervised AML subtype classifier with ABMIL [14] aggregation. In our experiments, DinoBloom models achieve a weighted F1-score between 91.7 (DinoBloom-L) and 93.1 (DinoBloom-G), while the second best models are ResNet50 and Phikon ViT-B with 81.9 and 81.8, respectively.

Bonemarrow cytology. In line with the results on the Acevedo dataset, DinoBloom models outperform both non-medical and medical models on bone marrow WBC classification by even larger margins. The classification task on BMC is heavily imbalanced with 21 classes and over 170,000 samples in total,

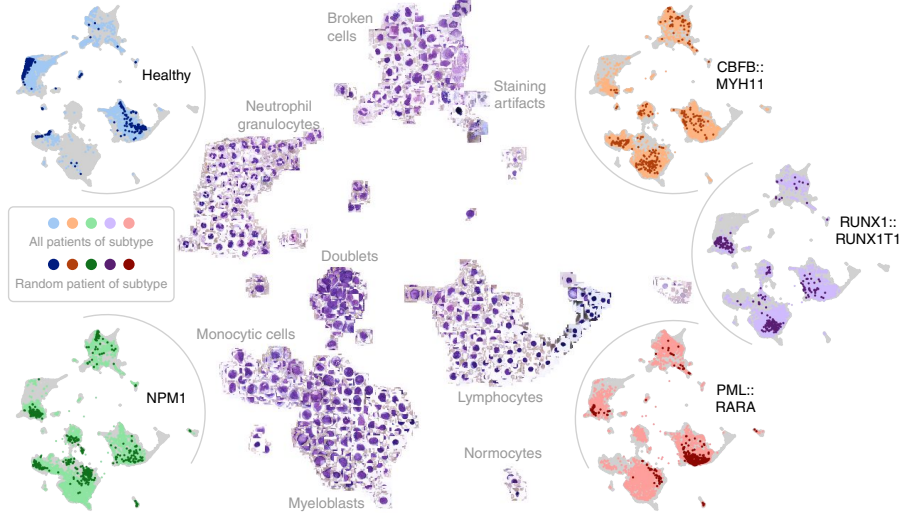


Fig. 2. Low dimensional representation (UMAP) of DinoBloom-B features fitted on over 80,000 single cells from the AML Hehr training dataset. Center: UMAP with original images. Five arcs: UMAP of patients from the training set (grey) and from the test set for healthy patients (blue), patients with CBFB::MYH11 (orange), NPM1 (green), PML::RARA (red), and RUNX1::RUNX1T1 (purple). In each arc, all test set patients are shown in bright dots, with one random patient highlighted in dark dots.

where some classes have a very low sample count, such as abnormal eosinophils (8) or smudge cells (42, Fig. 1e). Despite the challenging task, DinoBloom-B reaches a balanced accuracy of 65.8 and an accuracy of 79.5 in 1-NN evaluation compared to 42.0 and 48.7 of the next best model (Table 3). Similar large gaps in performance are also observed in linear probing, where our best model, DinoBloom-S, achieves a balanced accuracy of 71.4 compared to 52.2 of CTransPath. Notably, the performance in 1-NN weighted-F1 score is doubled compared to the DINOv2 baseline that can be observed for the ViT-B variant. Compared to the peripheral blood tasks, the performance of the larger variants of DinoBloom is not consistently better than that of the smaller variants.

Patient embeddings. We show a potential clinical application of our model: The low-dimensional embedding of AML and healthy patients from the training set of the AML Hehr dataset shows that related cell types cluster well (Fig. 2). Based on this fit, new patients (from the test set) are embedded into the same lower dimensional feature representation. The distribution of cell types is clearly distinct between healthy patients and all AML subtypes: as expected there are almost no cells in the Myeloblast related cluster in the healthy group. More subtle differences can be seen between clustering profiles of distinct AML subtypes, e.g., CBFB::MYH11 and RUNX1::RUNX1T1. The cell embeddings of our model could give clinicians an easy-to-grasp visualization of the cell distribution of a

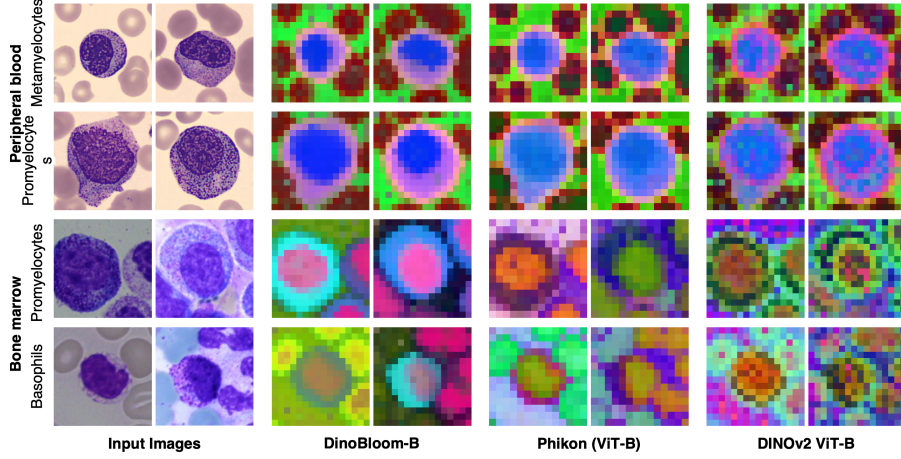


Fig. 3. PCA visualization of the patch tokens on the test data of Acevedo (external) and BMC. Comparison between DinoBloom-B, the second best model Phikon (ViT-B), and the pretrained DINOv2 ViT-B. Colors represent the values of the first three PCA components. DinoBloom-B can differentiate between nuclei, cytoplasm, surrounding red blood cells, and background.

patient and help to verify the manual quantification of different cell types. For instance, experts could identify the presence of myeloblasts within an entire smear, and gates could be applied to the embedding to facilitate morphology-based cell counting, similar to FACS analysis.

Interpretability. In Fig. 3, we show that our model learns robust features that detect semantic concepts across domains and compare it to its baseline, the pretrained DINOv2, and the second best performing model Phikon. We compute the principal components for the encoded patches of four images per dataset and visualize the first three components of each patch of size 14×14 (DinoBloom-B, DINOv2 ViT-B) and 16×16 pixels (Phikon), respectively, in RGB colors. One can clearly observe that DinoBloom captures shapes of nuclei, cell body, and the surrounding cells. While the other two models also capture shape and roughly outline cells, they do not catch fine grained details, as can be especially seen in the case of the nuclei, that only DinoBloom is able to differentiate from the whole cell.

5 Conclusion

With DinoBloom, we introduce a publicly available family of foundation models for single cell images in hematology, trained on the largest multi-cohort dataset with over 380,000 WBC images of 13 datasets. We show its strong generalization capabilities to external datasets despite strong batch effects. Our experiments demonstrate that DinoBloom extracts rich features from hematology images,

with its effectiveness demonstrated for cell-type classification and AML subtyping compared to non-medical and medical-imaging specific models. We also support this claim through visualizations showing that our model detects important hematological concepts, such as nuclei, cytoplasm, and red blood cells, which could be further leveraged for zero-shot segmentation. We believe that the generalizable cell embedding capabilities of our DinoBloom models offer great potential in assisting clinicians in their tedious manual work of cell detection and classification.

6 Disclosure of Interests

VK, SJW and SK are supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUdS". SJW is supported by the Add-on Fellowship of the Joachim Herz Foundation. This work was also supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). CM has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement number 866411).

References

1. Acevedo, A., Merino, A., Alf  rez, S., Molina,   ., Bold  , L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in brief* **30** (2020)
2. Alipo-on, J.R., Escobar, F.I., Novia, J.L., Atienza, M.M., Mana-ay, S., Tan, M.J., AlDahoul, N., Yu, E.: Dataset for machine learning-based classification of white blood cells of the juvenile visayan warty pig (2022). <https://doi.org/10.21227/3qsb-d447>
3. Arber, D.A., Orazi, A., Hasserjian, R.P., Borowitz, M.J., Calvo, K.R., Kvasnicka, H.M., Wang, S.A., Bagg, A., Barbui, T., Branford, S., et al.: International consensus classification of myeloid neoplasms and acute leukemias: integrating morphology, clinical, and genomic data. *Blood, The Journal of the American Society of Hematology* **140**(11), 1200–1228 (2022)
4. Aslan, A.: Blood cell detection dataset (2020), accessed: 2024-03-05
5. Caron, M., Touvron, H., Misra, I., J  gou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers (2021)
6. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., Williams, M., Vaidya, A., Sahai, S., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Williams, W., Le, L.P., Gerber, G., Mahmood, F.: A general-purpose self-supervised model for computational pathology (2023)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)

8. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Kain, A.M., Saillard, C., Schiratti, J.B.: Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv* (2023)
9. Font, P., Loscertales, J., Soto, C., Ricard, P., Novas, C.M., Martín-Clavero, E., López-Rubio, M., Garcia-Alonso, L., Callejas, M., Bermejo, A., et al.: Interobserver variance in myelodysplastic syndromes with less than 5% bone marrow blasts: unilineage vs. multilineage dysplasia and reproducibility of the threshold of 2% blasts. *Annals of hematology* **94**, 565–573 (2015)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proc. of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
12. Hehr, M., Sadafi, A., Matek, C., Lienemann, P., Pohlkamp, C., Haferlach, T., Spiekermann, K., Marr, C.: Explainable ai identifies diagnostic cells of genetic aml subtypes. *PLOS Digital Health* **2**(3), e0000187 (2023)
13. Houwen, B.: The differential cell count. *Laboratory Hematology* **7**, 89–100 (2001)
14. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: *International conference on machine learning*. pp. 2127–2136. PMLR (2018)
15. Khoury, J.D., Solary, E., Abla, O., Akkari, Y., Alaggio, R., Apperley, J.F., Bejar, R., Berti, E., Busque, L., Chan, J.K., et al.: The 5th edition of the world health organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms. *Leukemia* **36**(7), 1703–1719 (2022)
16. Koohbanani, N.A., Jahanifar, M., Tajadin, N.Z., Rajpoot, N.: Nuclick: a deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis* **65**, 101771 (2020)
17. Kouzehkanan, Z., Saghari, S., Tavakoli, E., Rostami, P., Abaszadeh, M., Mirzadeh, F., Satlsar, E., Gheidishahran, M., Gorgi, F., Mohammadi, S., et al.: Raabin-wbc: a large free access dataset of white blood cells from normal peripheral blood. (2021)
18. Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T., Marr, C.: Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood, The Journal of the American Society of Hematology* **138**(20), 1917–1927 (2021)
19. Matek, C., Schwarz, S., Marr, C., Spiekermann, K.: A single-cell morphological dataset of leukocytes from aml patients and non-malignant controls (aml-cytomorphology_lmu). *The Cancer Imaging Archive (TCIA)* (2019)
20. Matek, C., Schwarz, S., Spiekermann, K., Marr, C.: Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nature Machine Intelligence* **1**(11), 538–544 (2019)
21. Mohamed, M., Far, B., Guaily, A.: An efficient technique for white blood cells nuclei automatic segmentation. In: *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. pp. 220–225. IEEE (2012)
22. Naruenatthanaset, K., Chalidabhongse, T.H., Palasuwan, D., Anantrasirichai, N., Palasuwan, A.: Red blood cell segmentation with overlapping cell separation and classification on imbalanced dataset (2023)
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
24. Rastogi, P., Khanna, K., Singh, V.: Leufeatx: Deep learning-based feature extractor for the diagnosis of acute leukemia from microscopic images of peripheral blood smear. *Computers in Biology and Medicine* **142**, 105236 (2022)

25. Rezatofighi, S.H., Soltanian-Zadeh, H.: Automatic recognition of five types of white blood cells in peripheral blood. *Computerized Medical Imaging and Graphics* **35**(4), 333–343 (2011)
26. Roth, B., Koch, V., Wagner, S.J., Schnabel, J.A., Marr, C., Peng, T.: Low-resource finetuning of foundation models beats state-of-the-art in histopathology. *arXiv preprint arXiv:2401.04720* (2024)
27. Sadafi, A., Salehi, R., Gruber, A., Boushehri, S.S., Giehr, P., Navab, N., Marr, C.: A continual learning approach for cross-domain white blood cell classification. In: Koch, L., Cardoso, M.J., Ferrante, E., Kamnitsas, K., Islam, M., Jiang, M., Rieke, N., Tsafaris, S.A., Yang, D. (eds.) *Domain Adaptation and Representation Transfer*. pp. 136–146. Springer Nature Switzerland, Cham (2024)
28. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., Mathieu, P., van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y.K., Kunz, J.D., Lee, M.C.H., Bernhard, J., Godrich, R.A., Oakley, G., Millar, E., Hanna, M., Retamero, J., Moye, W.A., Yousfi, R., Kanan, C., Klimstra, D., Rothrock, B., Fuchs, T.J.: Virchow: A million-slide digital pathology foundation model (2024)
29. Wagner, S.J., Reisenbüchler, D., West, N.P., Niehues, J.M., Zhu, J., Foersch, S., Veldhuizen, G.P., Quirke, P., Grabsch, H.I., van den Brandt, P.A., et al.: Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell* (2023)
30. Wang, X., Du, Y., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Retccl: Clustering-guided contrastive learning for whole-slide image retrieval. *Medical Image Analysis* **83**, 102645 (2023)
31. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis* **81**, 102559 (2022)
32. Zheng, X., Wang, Y., Wang, G., Liu, J.: Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron* **107**, 55–71 (2018)

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Dec 17, 2024

This Agreement between Technical University Munich ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5931221004050
License date	Dec 17, 2024
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology
Licensed Content Author	Valentin Koch, Sophia J. Wagner, Salome Kazeminia et al
Licensed Content Date	Jan 1, 2024
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter

Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	Generalizable and Data- efficient Deep Learning for Computer-aided Cancer Diagnosis
Institution name	Technical University Munich
Expected presentation date	Feb 2025
The Requesting Person / Organization to Appear on the License	Technical University Munich
Requestor Location	Sophia Wagner Ingolstädter Landstraße 1 Neuherberg, 85764 Germany
Billing Type	Invoice
Billing Address	Sophia Wagner Ingolstädter Landstraße 1 Neuherberg, Germany 85764
Total	0.00 EUR
Terms and Conditions	

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed

Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at <https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/>.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a	

journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: *'Reproduced with permission from Springer Nature'*.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany's choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit <https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution>

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

A.6. Local attention graph-based transformer for multi-target genetic alteration prediction

This publication was published in the proceedings of the peer-reviewed international conference Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2022. The one-page summary can be found in Section 3.6.

Reisenbüchler, D., Wagner, S. J., Boxberg, M., & Peng, T. (2022, September). Local attention graph-based transformer for multi-target genetic alteration prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 377-386). Cham: Springer Nature Switzerland.

Copyright: Copyright © 2022, Springer Nature. Reproduced with permission from Springer Nature. The license is attached after the original publication.



Local Attention Graph-Based Transformer for Multi-target Genetic Alteration Prediction

Daniel Reisenbüchler^{1,2(✉)}, Sophia J. Wagner^{1,2}, Melanie Boxberg³,
and Tingying Peng²

¹ Technical University Munich, Munich, Germany

² Helmholtz AI, Neuherberg, Germany

reisenbuechler@helmholtz-muenchen.de

³ Institute of Pathology Munich-North, Munich, Germany

Abstract. Classical multiple instance learning (MIL) methods are often based on the identical and independent distributed assumption between instances, hence neglecting the potentially rich contextual information beyond individual entities. On the other hand, Transformers with global self-attention modules have been proposed to model the interdependencies among all instances. However, in this paper we question: Is global relation modeling using self-attention necessary, or can we appropriately restrict self-attention calculations to local regimes in large-scale whole slide images (WSIs)? We propose a general-purpose local attention graph-based Transformer for MIL (LA-MIL), introducing an inductive bias by explicitly contextualizing instances in adaptive local regimes of arbitrary size. Additionally, an efficiently adapted loss function enables our approach to learn expressive WSI embeddings for the joint analysis of multiple biomarkers. We demonstrate that LA-MIL achieves state-of-the-art results in mutation prediction for gastrointestinal cancer, outperforming existing models on important biomarkers such as microsatellite instability for colorectal cancer. Our findings suggest that local self-attention sufficiently models dependencies on par with global modules. Our LA-MIL implementation is available at https://github.com/agentdr1/LA_MIL.

Keywords: Multiple instance learning · Graph transformer · Local attention · Whole slide images · Mutation prediction

1 Introduction

Advances in slide-scanning microscopes and deep learning-based image analysis have significantly increased interest in computational pathology [2]. Whole slide images typically contain billions of pixels and reach up to several gigabytes in size. To mitigate the resulting computational burden, WSIs are commonly

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-16434-7_37.

tessellated into smaller tiles [16]. However, patient diagnosis is typically only available as weakly-supervised slide-level annotation, e.g., cancer vs. non-cancer classification, cancer subtyping, or genomic analyses.

In histopathological image analysis, this task is formulated as multiple instance learning (MIL), where a WSI is considered as a bag, and tiles as contained instances. Hence, efficiently learning representations and aggregating them from tiles to a bag label is crucial. One simple solution for this is to pass the bag label onto each tile, reducing MIL to supervised learning. This approach is particularly favored because of its ease of implementation, e.g., to predict microsatellite instability or tumor mutational burden [10, 22]. The final bag-level prediction is obtained by aggregating all instance-level predictions with average pooling. These methods have two drawbacks: (i) a fraction of instance labels may differ from the bag label and therefore form label noise in supervised learning, and (ii) no morphological or spatial correlation between tiles is taken into account.

To remedy (i), MIL can learn from bag-level annotation without assuming the same label for each tile. In particular, Ilse et al. [8] propose an attention-based pooling layer aiming to weight each tile individually for its relevance within the bag prediction task. To tackle (ii), recently, convolutional neural networks (CNNs) were combined with self-attention-based Transformers [21]. Here, the tiles are condensed into feature vectors and subsequently the resulting sequence is fed into a Transformer, where the interdependence between tiles is incorporated by self-attention mechanisms. For instance, Li et al. [11] propose a deformable Transformer-based encoder-decoder structure and evaluate it across encoder only based Transformer. Shao et al. [20] uses the Neystrom method to approximate self-attention, aiming to decrease the computational complexity. Myronenko et al. [17] suggest incorporating feature vectors of different scales into an encoder-based Transformer.

However, general Transformer approaches suffer from quadratic complexity with respect to the sequence length. This complexity is a general problem across computer vision and neural language processing (NLP) domains. To alleviate this concern, Transformer using local attention in the NLP domain [15] showed that it is sufficient for a token to restrict the attention calculation to a local neighborhood inside the sequence, i.e., the surrounding words. On the other hand, in computer vision self-attention can be modified by introducing local windowed attention [12]. However, this approach for WSIs is not as conveniently applicable as for domain areas where the images have the same size, such as in well-curated datasets like ImageNet [4]. Whole slide images come with varying geometrical shapes and the representation obtained by using tiles while excluding some entities (e.g. due to artifacts, pen marks, etc.) leads to holes within the visual representation. A handcrafted selection of which entities participate in the key-to-query product is not generally applicable for all WSIs, as this would not effectively adapt to the varying local neighborhoods.

Combining the local windowed attention idea with Graph Transformer [5], we propose a computationally light training pipeline consisting of a CNN and local attention-based Transformer with the following contributions:

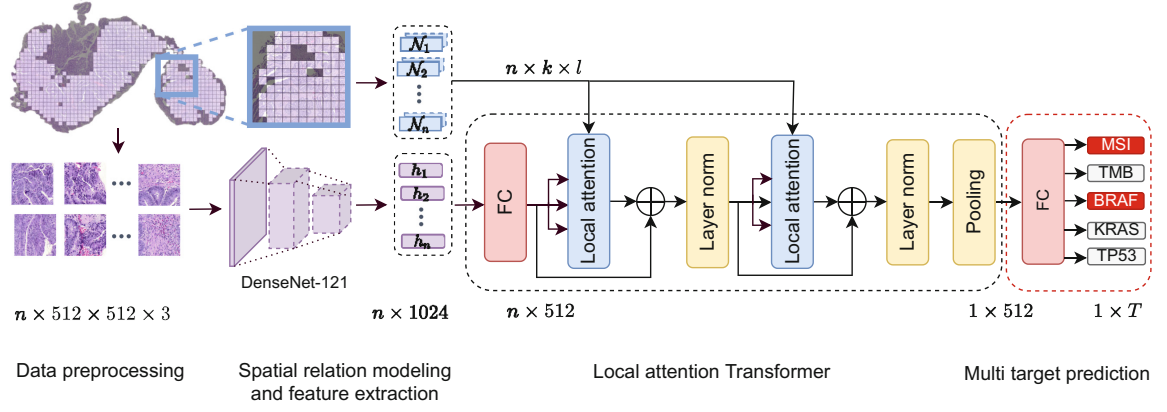


Fig. 1. LA-MIL overview: The pipeline consists of preprocessing, spatial relation modeling and feature extraction, and a local attention-based transformer.

- We present LA-MIL, a novel local attention graph-based Transformer that restricts self-attention calculations in Transformers by using k-nearest neighbor (kNN) graphs to model local regimes with respect to tiles inside the WSI.
- To our knowledge, LA-MIL is the first pipeline to predict microsatellite instability and tumor mutational burden jointly with genetic alterations as well as the first transformer-based approach for mutation prediction.
- An efficiently adapted loss function enables our approach to learn meaningful bag representations for a joint analysis of multiple imbalanced biomarkers.
- We evaluate our approach extensively on two datasets for gastrointestinal cancer and demonstrate that our local attention mechanism sufficiently leverages information on par with global self-attention modules.
- LA-MIL shows great modeling interpretability by visualizing local attention scores, consisting of spatial and morphological dependencies.

Our experiments indicate that LA-MIL outperforms state-of-the-art approaches or is on par for mutation prediction tasks in gastrointestinal cancer.

2 Method

The pipeline of our LA-MIL approach is visualized in Fig. 1. In the following, we introduce the key components of our algorithm.

2.1 LA-MIL Framework

Data Preprocessing. First, a given gigapixel WSI is tessellated into N tiles, where each tile $t_i \in \mathbb{R}^{H \times W \times C}$. Further, we extract the coordinates $c_i \in \mathbb{R}^2$ with respect to the WSI for each tile. Tiles containing background, artifacts, and non tumor-tissue are excluded using Otsu’s method [18] and region of interest (RoI) annotations, reducing the number of tiles for downstream processing to $n < N$.

Per-Tile Feature Extraction and Spatial Relation Modeling. We compress the visual information contained in each tile t_i by extracting features using KimiaNet [19], a pretrained DenseNet-121 [7]. Thus, the WSI is represented as a sequence of feature vectors $\{h_i\}_{i=1}^n \in \mathbb{R}^{n \times D}$, where the dimension D is the output size of the feature extraction CNN. For the spatial relation among the tiles of a WSI, we build l k-nearest neighbor (kNN) graphs $\mathcal{G}_{kNN}^l$ using the Euclidean distance of the coordinates c_i . A kNN graph can be represented by a matrix $A \in \mathbb{R}^{n \times k}$, indicating the k neighbors for all n tiles.

Transformer Architecture. Given a sequence of features $\{h_i\}_{i=1}^n$ and l graphs $\mathcal{G}_{kNN}^l$, we further downscale the feature vectors from D to d by using a fully-connected (FC) layer. Subsequently, a Transformer with l blocks of local attention layers is applied. These layers utilize the kNN graphs to update neighboring tiles, thereby modeling local morphological and spatial correlations. Note that a graph can also be shared between layers. Applying a residual connection and layer normalization [1] after each attention layer aims to improve the gradient flow and generalization performance. Finally, the sequence is aggregated into a bag-embedding vector $b \in \mathbb{R}^d$ by mean pooling as done in [21]. Another FC layer projects the bag-embedding vector into a target vector $t \in \mathbb{R}^T$, where T is the number of targets to predict. The sigmoid function is applied element-wise on the target vector t to obtain the scores for each target individually. In the context of mutation prediction, the thresholded scores indicate whether a particular gene occurs as wildtype or mutated, respectively.

Loss Function. Mutation prediction is a challenging task since the targets often only occur in small frequencies (see Table 1). Hence, we use a loss which penalizes the model for wrong decisions about the prediction of underrepresented classes by weighting each binary cross-entropy (BCE) term individually. We take the mean of T BCE losses, thus treating each target equally:

$$L(\mathbf{x}, \mathbf{y}) = -\frac{1}{T} \sum_{t=1}^T \frac{n_t^{\text{neg}}}{n_t^{\text{pos}}} (y_t \log(\sigma(x_t)) + (1 - y_t) \log(1 - \sigma(x_t))).$$

As mentioned in the introduction, most mutation prediction studies use tile-level supervised learning rather than bag-level MIL. Hence, a common strategy to tackle highly imbalanced classification is to apply downsampling to reach an equilibrium of classes in the dataset splits [9,10]. However, this may not be possible in the multi-target and bag-operating setup, as downsampling may exclude nearly all samples, depending on the individual class distributions.

2.2 Local Attention Layer

Self-attention is a key component in Transformer architectures, where each token h_i is updated with global information of the complete input sequence $\{h_1, \dots, h_n\}$. In contrast, our local attention modules constrain the updates for

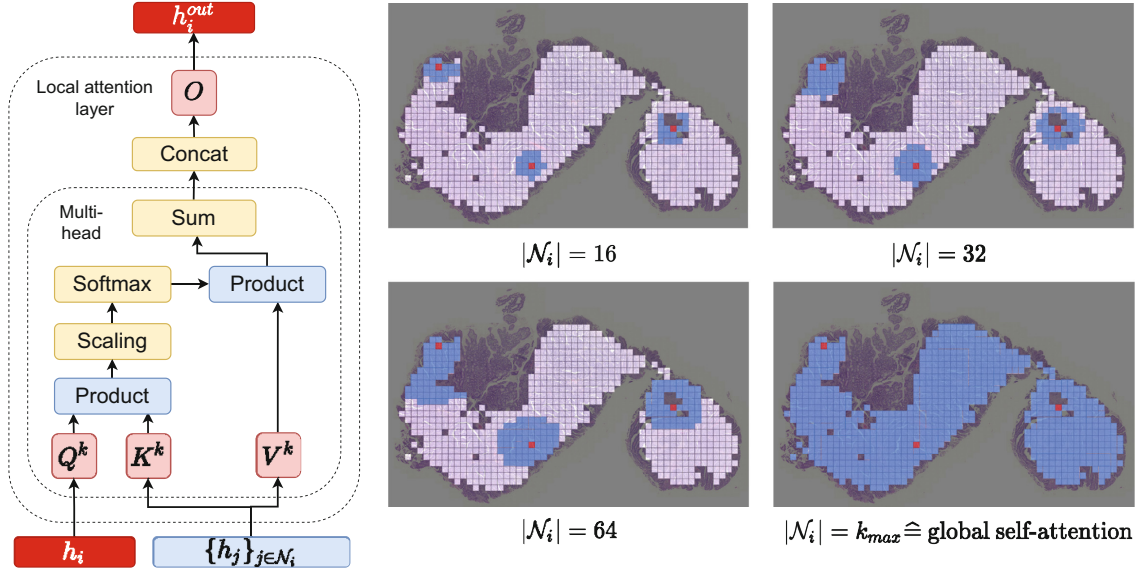


Fig. 2. Left: Computational block for a local attention layer. Right: Illustration of the spatial field of view for the attention calculations. Tiles in red visualize a query tile and blue colored tiles visualize the adaptive local neighborhoods of different sizes. (Color figure online)

each token h_i associated with node n_i to all tokens h_j with nodes $j \neq i$ that are connected with node n_i , as shown in Fig. 2. As input we consider a n -dimensional sequence of tokens h_i for $i = 1, \dots, n$ associated with n nodes of a graph $\mathcal{G}$ where the nodes are connected by $n \times k$ edges. The update equation for token h_i is

$$h_i = O \cdot \text{Concat}_k \left(\sum_{j \in \mathcal{N}_j} w_{ij}^k V^k h_j \right), \quad w_{i,j}^k = \text{softmax}_j \left(\frac{Q^k h_i \cdot K^k h_j}{\sqrt{d_k}} \right) \quad (1)$$

where $Q^k, K^k, V^k \in \mathbb{R}^{d_k \times d}$ and $O \in \mathbb{R}^{d \times d}$ with $k = 1, \dots, H$ denoting the number of the respective attention head. The notation $j \in \mathcal{N}_i$ refers to a set of indices j of nodes connected to the i th node by j edges. The cardinality of $\mathcal{N}_i$ is equal to the number of neighbors for all tiles. To calculate local attention scores a_i for each tile from local attention layers, we first cache the intermediate outputs $w_{i,j}^k$ from Eq. 1 and sum them across heads and local regimes, i.e.

$$a_i = \sum_{k=1}^H \sum_{j \in \mathcal{N}_j} w_{i,j}^k. \quad (2)$$

Subsequently, we normalize all n attention values a_i into the range $[0, 1]$ and denote the outcomes as local attention scores for each tile.

3 Experiments

We applied the proposed method on diagnostic formalin-fixed paraffin-embedded diagnostic slides for two cohorts in The Cancer Genome Atlas (TCGA)

Table 1. Distribution of genetic alterations in TCGA-CRC and TCGA-STAD. We denote the number of positive samples for each target.

Cohort	n	MSI	TMB	BRAF	ALK	ERBB4	FBXW7	KRAS	PIK3CA	SMAD4	TP53
CRC	594	78	85	66	40	62	107	223	178	81	332
STAD	440	75	86	13	19	62	38	40	86	36	225

dataset [24]. From tissue contained in WSIs, we followed recent works [3, 6, 9] and include only tumor-occupied tissue regions. All images were downsampled to $20\times$ magnification, corresponding to a resolution of $0.5 \frac{\mu m}{px}$. The task is to predict genetic alterations, the microsatellite status and the tumor mutational burden (TMB) as biomarkers [16] directly from WSIs.

TCGA Colorectal and TCGA Stomach Datasets. Our dataset TCGA-CRC consists of tiled WSIs from tumor regions of colorectal tissue. We used the preprocessed tumor tissue tiles from kather.ai. As a second dataset, we tiled WSIs of stomach tissue from TCGA, downloaded at portal.gdc.cancer.gov. After noise removal, we excluded all tiles which were not contained in the tumor region by using manual tumor annotations available at: kather.ai. We retrieved the genetic annotations matching the WSIs from xenabrowser.net. Annotations for the microsatellite stability/instability (MSS/MSI) and TMB are available at cbioportal.org. Following Luchini et al. [14], we binarized MSS and MSI-Low as MSS and MSI-High as MSI.

Implementation. Each tile was embedded into a 1024-dimensional feature space by a DenseNet-121 model that was pretrained on histopathological data. By using the coordinates of each tile, we built two kNN-Graphs with $k = 16$ and $k = 64$ for subsequent attention restriction in the first and second local attention module, respectively. In the training phase, each feature vector associated with the tiles was further compressed from 1024 to 512 by a FC layer. After a stack of two local attention layers, we averaged the feature vectors across all tiles. The resulting bag embedding was passed through a classification head, consisting of another fully connected layer from 512 to 10, to compute the logits. We applied the sigmoid activation function element-wise to calculate the probabilities for each individual target. For optimization, we employed the Lookahead optimizer [25] together with AdamW [13], and used a learning rate of $2e-05$ and $2e-04$ (for TCGA-CRC and TCGA-STAD, respectively) for 10 epochs, weight decay of $2e-05$, and batch size 1. The LA-MIL model with 2.1M parameter was implemented in PyTorch and DGL [23], and trained on a single Tesla V100 GPU.

Evaluation. To evaluate the mutation prediction task on both datasets TCGA-CRC and TCGA-STAD, we compared the performance of LA-MIL with state-of-the-art methods. The fact that most of the recent advances predict T biomarkers individually results in training, validating, and hyperparameter tuning for T

Table 2. Mean AUROC scores for mutation prediction on the datasets TCGA-CRC and TCGA-STAD. For the competitive methods, we report results from the original publications; for our methods, we report the mean over five folds (see supplementary material for results with standard deviation).

Dataset	Method	MSI	TMB	BRAF	ALK	ERBB4	FBXW7	KRAS	PIK3CA	SMAD4	TP53
TCGA-CRC	Kather et al. [10]	0.77	—	—	—	—	—	—	—	—	—
	Wang et al. [22]	—	0.82	—	—	—	—	—	—	—	—
	Kather et al. [9]	—	—	0.66	0.51	—	0.49	0.60	0.62	0.63	0.68
	Fu et al. [6]	—	—	0.57	—	—	0.66	0.55	0.59	0.58	0.68
	T-MIL (Ours)	0.85	0.82	0.73	0.61	0.57	0.64	0.61	0.60	0.60	0.64
	LA-MIL (Ours)	0.85	0.83	0.72	0.63	0.60	0.66	0.62	0.61	0.58	0.63
TCGA-STAD	Kather et al. [10]	0.81	—	—	—	—	—	—	—	—	—
	Wang et al. [22]	—	0.75	—	—	—	—	—	—	—	—
	Kather et al. [9]	—	—	0.37	0.45	—	0.74	0.64	0.67	0.61	0.60
	Fu et al. [6]	—	—	—	—	—	—	—	0.47	0.49	0.63
	T-MIL (Ours)	0.80	0.78	0.73	0.52	0.47	0.71	0.65	0.58	0.62	0.57
	LA-MIL (Ours)	0.78	0.77	0.67	0.52	0.47	0.72	0.70	0.61	0.64	0.58

separate models, while we used a single model to predict all biomarkers. Moreover, we implemented a Transformer MIL approach where we exchanged all local attention blocks with global self-attention [21], denoted as T-MIL. To stick with common evaluation procedures for mutation prediction in recent works, we evaluated our pipeline with a 5-fold cross validation (CV). We split the datasets into folds such that individual class distributions in each fold were approximately the same and ensured that no patient appeared in the training and validation set at the same time. We measured the performance using the area under the receiver operating characteristic curve (AUROC) for each target individually.

4 Results

Current methods, such as Kather et al. [10] for MSI and Wang et al. [22] for TMB, predict the biomarkers instance-wise as single targets for each tile with the corresponding label inherited from its parent WSI. Similarly, Kather et al. [9] and Fu et al. [6] train one model for each target gene when predicting mutations, but evaluate their results on WSI-level by average pooling of tile-wise predictions. In contrast, we train and evaluate only one model on WSI-level to predict multiple biomarkers using a MIL transformer that aggregates features from all tiles with global or local self-attention layers.

Table 2 shows the AUROC scores of state-of-the-art instance-wise methods compared to our methods on the dataset TCGA-CRC. The results suggest that our models can leverage information from multiple targets to achieve better overall performance. Interestingly, this holds especially for the prediction of MSI, where our approach improves the score by 8% from 0.77 to 0.85. The prediction of TMB and most of the gene mutations (except for SMAD4 and TP53) are on par or marginally better (+1%).

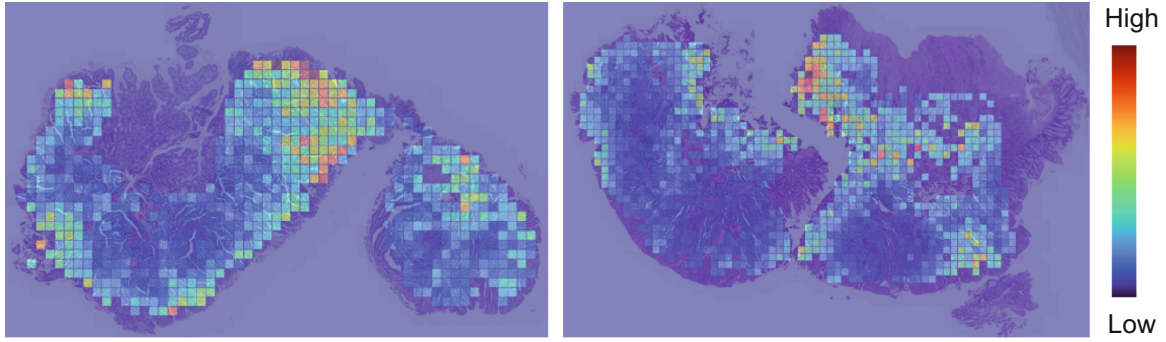


Fig. 3. Local attention scores visualization for the last local attention layer with restricted self-attention in a neighborhood of size 64.

The results for the TCGA-STAD dataset in MSI prediction are marginally worse (-1%). This could arise due to the fact that the compared work uses a slightly different label strategy for MSI-Low cases, which also affects the evaluation. The results for TMB improves by up to 3% . Similar as on the TCGA-CRC dataset, we observe an improvement of results using our Transformer-based approaches for the remaining targets except for a few genes.

Local Attention Visualization. As described in Eq. 2, we can calculate the local attention scores from the query-to-key product. In Fig. 3 we colorized the tiles according to their corresponding local attention score. The high attentive regions include the nuclear chromatin that seems to be hyperchromatic, as well as crowded glands or solid areas. Yet, to the best of our knowledge, there are no distinguishing patterns from WSI for mutated genes, which makes it difficult for a quantitative evaluation. Nevertheless, LA-MIL provides an interpretation based on the contribution of each tile for the bag-level prediction task and thus paves a way for a deeper investigation of highly scored tiles.

5 Conclusion

In this work, we proposed a novel MIL framework with local attention for WSI analysis. Local attention is achieved through a graph-based transformer that models region-wise inter-dependencies. As the size of the region can be set arbitrarily, our approach bridges the gap between instance-wise and global relation approaches by modeling local relations of arbitrary size. Moreover, an effective adapted loss enables us to learn multiple biomarkers at once, for low computational cost compared to CNN-only based methods.

However, there is often more than one WSI for a patient available in the TCGA database. Future work will investigate the strategy of combining all WSIs for a patient, while suitably scaling the coordinates for tiles of different WSIs. Thus, each tile will only be updated with local information from its direct parent WSI and transformed into a bag-embedding consisting of locally correlated

tiles of all WSIs from a patient. We believe that our approach provides a base for further applications in other WSI analysis tasks where structured relation modeling is crucial.

Acknowledgements. S.J.W. was supported by the Helmholtz Association under the joint research school “Munich School for Data Science - MUDS”.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016). <https://doi.org/10.48550/ARXIV.1607.06450>
2. Cooper, L.A., Demicco, E.G., Saltz, J.H., Powell, R.T., Rao, A., Lazar, A.J.: Pan-Cancer insights from the cancer genome atlas: the pathologist’s perspective. *J. Pathol.* **244**(5), 512–524 (2018)
3. Coudray, N., et al.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**(10), 1559–1567 (2018)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Dwivedi, V.P., Bresson, X.: A generalization of transformer networks to graphs. In: Methods and Applications, AAAI Workshop on Deep Learning on Graphs (2021)
6. Fu, Y., et al.: Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**(8), 800–810 (2020). <https://doi.org/10.1038/s43018-020-0085-8>
7. Huang, G., Liu, Z., Pleiss, G., Van Der Maaten, L., Weinberger, K.: Convolutional networks with dense connectivity. *IEEE Trans. Pattern Anal. Mach. Intell.* (2019)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (2018)
9. Kather, J.N., et al.: Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**(8), 789–799 (2020)
10. Kather, J.N., et al.: Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**(7), 1054–1056 (2019)
11. Li, H., et al.: DT-MIL: deformable transformer for multi-instance learning on histopathological image. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 206–216. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_20
12. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision (ICCV) (2021)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, 6–9 May 2019. OpenReview.net (2019)
14. Luchini, C., et al.: ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour mutational burden: a systematic review-based approach. *Ann. Oncol.* **30**(8), 1232–1243 (2019)

15. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421. Association for Computational Linguistics, Lisbon (2015). <https://doi.org/10.18653/v1/D15-1166>
16. Murchan, P., et al.: Deep learning of histopathological features for the prediction of tumour molecular genetics. *Diagnostics* **11**(8), 1406 (2021)
17. Myronenko, A., Xu, Z., Yang, D., Roth, H.R., Xu, D.: Accounting for Dependencies in Deep Learning Based Multiple Instance Learning for Whole Slide Imaging. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 329–338. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_32
18. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979). <https://doi.org/10.1109/TSMC.1979.4310076>
19. Riasatian, A., et al.: Fine-tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. *Med. Image Anal.* **70**, 102032 (2021)
20. Shao, Z., et al.: Transmil: transformer based correlated multiple instance learning for whole slide image classification (2021)
21. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. (2017)
22. Wang, L., Jiao, Y., Qiao, Y., Zeng, N., Yu, R.: A novel approach combined transfer learning and deep learning to predict TMB from histology image. *Pattern Recogn. Lett.* **135**, 244–248 (2020)
23. Wang, M., et al.: Deep graph library: a graph-centric, highly-performant package for graph neural networks (2020)
24. Weinstein, J.N., et al.: The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**(10), 1113–1120 (2013)
25. Zhang, M.R., Lucas, J., Hinton, G., Ba, J.: Lookahead optimizer: k steps forward, 1 step back (2019)

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Dec 16, 2024

This Agreement between Sophia J. Wagner ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5930651053928
License date	Dec 16, 2024
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Local Attention Graph-Based Transformer for Multi-target Genetic Alteration Prediction
Licensed Content Author	Daniel Reisenbüchler, Sophia J. Wagner, Melanie Boxberg et al
Licensed Content Date	Jan 1, 2022
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter

Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	Generalizable and Data- efficient Deep Learning for Computer-aided Cancer Diagnosis
Institution name	Technical University Munich
Expected presentation date	Feb 2025
The Requesting Person / Organization to Appear on the License	Sophia J. Wagner
Requestor Location	Sophia Wagner Ingolstädter Landstraße 1 Neuherberg, 85764 Germany
Billing Type	Invoice
Billing Address	Sophia Wagner Ingolstädter Landstraße 1 Neuherberg, Germany 85764
Total	0.00 EUR
Terms and Conditions	

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed

Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at <https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/>.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a	

journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: *'Reproduced with permission from Springer Nature'*.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licenser warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licenser and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licenser reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany's choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit <https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution>

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

A.7. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study

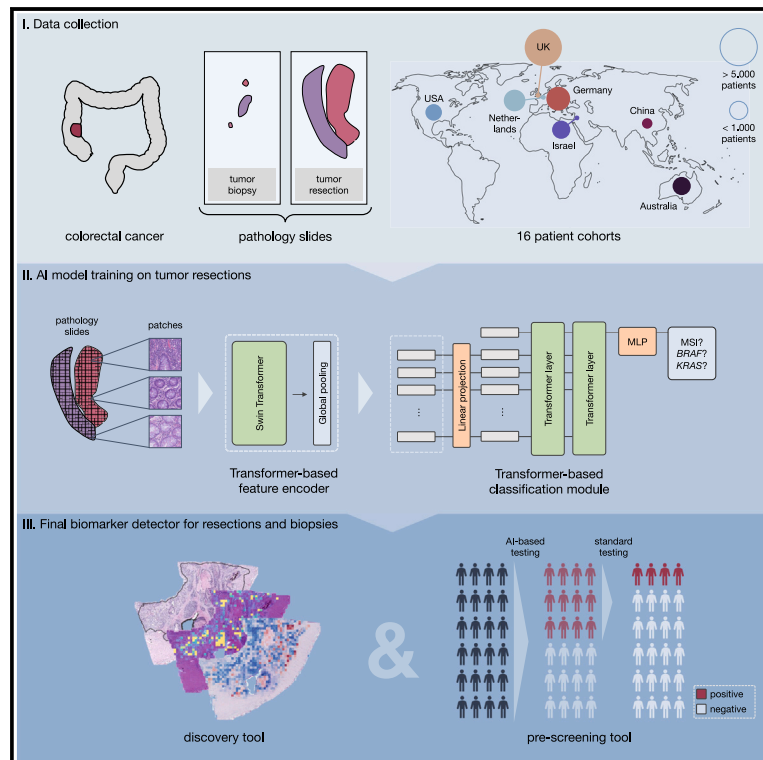
This publication was published in the peer-reviewed international journal *Cancer Cell* in 2023. The one-page summary can be found in Section 3.7.

Wagner, S. J., Reisenbüchler, D., West, N. P., Niehues, J. M., Zhu, J., Foersch, S., ... & Kather, J. N. (2023). Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell*, 41(9), 1650-1661.

Copyright: © 2023 The Author(s). Published by Elsevier Inc. This is an open access article distributed under the terms of the Creative Commons CC-BY license. The copyright clearance is attached after the original publication.

Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study

Graphical abstract



Authors

Sophia J. Wagner,
Daniel Reisenbüchler,
Nicholas P. West, ..., Melanie Boxberg,
Tingying Peng, Jakob Nikolas Kather

Correspondence

tingying.peng@helmholtz-munich.de
(T.P.),
jakob_nikolas.kather@
tu-dresden.de (J.N.K.)

In brief

Wagner et al. show that transformer-based prediction of biomarkers from histology substantially improves the performance, generalizability, data efficiency, and interpretability as compared with current state-of-the-art algorithms. The method significantly outperforms existing approaches for microsatellite instability detection in surgical resections and reaches clinical-grade performance on biopsies of colorectal cancer, solving a long-standing diagnostic problem.

Highlights

- AI-based prediction of biomarkers (MSI, *BRAF*, and *KRAS*) using transformers
- MSI prediction reaches clinical-grade performance on biopsies of colorectal cancer
- Transformer-based biomarker prediction generalizes better and is more data efficient
- Large-scale multi-cohort evaluation on over 13,000 patients from 16 cohorts



Article

Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study

Sophia J. Wagner,^{1,2,3} Daniel Reisenbüchler,¹ Nicholas P. West,⁴ Jan Moritz Niehues,³ Jiefu Zhu,³ Sebastian Foersch,⁴ Gregory Patrick Veldhuizen,³ Philip Quirke,⁵ Heike I. Grabsch,^{5,6} Piet A. van den Brandt,⁷ Gordon G.A. Hutchins,⁵ Susan D. Richman,⁵ Tanwei Yuan,⁸ Rupert Langer,⁹ Josien C.A. Jenniskens,⁷ Kelly Offermans,⁷ Wolfram Mueller,¹⁰ Richard Gray,¹¹ Stephen B. Gruber,¹² Joel K. Greenson,¹³ Gad Rennert,^{14,15} Joseph D. Bonner,¹⁴ Daniel Schmolze,¹² Jitendra Jonnagaddala,¹⁶ Nicholas J. Hawkins,¹⁷ Robyn L. Ward,^{17,18} Dion Morton,¹⁹

(Author list continued on next page)

¹Helmholtz Munich – German Research Center for Environment and Health, Munich, Germany

²School of Computation, Information and Technology, Technical University of Munich, Munich, Germany

³Else Kroener Fresenius Center for Digital Health (EFFZ), Technical University Dresden, Dresden, Germany

⁴Institute of Pathology, University Medical Center Mainz, Mainz, Germany

⁵Division of Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK

⁶Department of Pathology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, the Netherlands

⁷Department of Epidemiology, Maastricht University Medical Center+, Maastricht, the Netherlands

⁸Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany

⁹Institute of Pathology and Molecular Pathology, Johannes Kepler University Hospital Linz, Linz, Österreich

¹⁰Gemeinschaftspraxis Pathologie, Starnberg, Germany

¹¹Nuffield Department of Population Health, University of Oxford, Oxford, UK

¹²Center for Precision Medicine and Department of Medical Oncology, City of Hope National Medical Center, Duarte, CA, USA

¹³Department of Pathology, City of Hope Comprehensive Cancer Center, Duarte, CA, USA

¹⁴Department of Community Medicine & Epidemiology, Lady Davis Carmel Medical Center, Ruth & Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel

¹⁵Steve and Cindy Rasmussen Institute for Genomic Medicine, Lady Davis Carmel Medical Center and Technion Faculty of Medicine, Clalit National Cancer Control Center, Haifa, Israel

¹⁶School of Population Health, Faculty of Medicine and Health, UNSW Sydney, Sydney, NSW, Australia

¹⁷School of Medical Sciences, Faculty of Medicine and Health, UNSW Sydney, Sydney, NSW, Australia

¹⁸Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia

(Affiliations continued on next page)

SUMMARY

Deep learning (DL) can accelerate the prediction of prognostic biomarkers from routine pathology slides in colorectal cancer (CRC). However, current approaches rely on convolutional neural networks (CNNs) and have mostly been validated on small patient cohorts. Here, we develop a new transformer-based pipeline for end-to-end biomarker prediction from pathology slides by combining a pre-trained transformer encoder with a transformer network for patch aggregation. Our transformer-based approach substantially improves the performance, generalizability, data efficiency, and interpretability as compared with current state-of-the-art algorithms. After training and evaluating on a large multicenter cohort of over 13,000 patients from 16 colorectal cancer cohorts, we achieve a sensitivity of 0.99 with a negative predictive value of over 0.99 for prediction of microsatellite instability (MSI) on surgical resection specimens. We demonstrate that resection specimen-only training reaches clinical-grade performance on endoscopic biopsy tissue, solving a long-standing diagnostic problem.

INTRODUCTION

Precision oncology in colorectal cancer (CRC) requires the evaluation of genetic biomarkers, such as microsatellite instability

(MSI)^{1–8} and mutations in the *BRAF*^{4,7} and *NRAS/KRAS*⁹ genes. These biomarkers are typically assessed by polymerase chain reaction (PCR), sequencing, or immunohistochemical assays. Biomarker identification in patients with CRC is an important



Matthew Seymour,²⁰ Laura Magill,²¹ Marta Nowak,²² Jennifer Hay,²³ Viktor H. Koelzer,^{22,24,25} David N. Church,^{25,26} TransSCOT consortium, Christian Matek,^{1,27,28} Carol Geppert,^{27,28} Chaolong Peng,²⁹ Cheng Zhi,³⁰ Xiaoming Ouyang,³⁰ Jacqueline A. James,^{31,32,33} Maurice B. Loughrey,^{33,34,35} Manuel Salto-Tellez,^{31,32,36} Hermann Brenner,^{8,37,38} Michael Hoffmeister,⁸ Daniel Truhn,³⁹ Julia A. Schnabel,^{1,2,40} Melanie Boxberg,^{41,42} Tingying Peng,^{1,44,*} and Jakob Nikolas Kather^{3,5,43,44,45,*}

¹⁹University Hospital Birmingham, Birmingham, UK

²⁰St James's University Hospital, Leeds, UK

²¹University of Birmingham Clinical Trials Unit, Birmingham, UK

²²Department of Pathology and Molecular Pathology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

²³Glasgow Tissue Research Facility, University of Glasgow, Queen Elizabeth University Hospital, Glasgow, UK

²⁴Department of Oncology, University of Oxford, Oxford, UK

²⁵Nuffield Department of Medicine, University of Oxford, Roosevelt Drive, Oxford, UK

²⁶Oxford NIHR Comprehensive Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

²⁷Institute of Pathology, University Hospital Erlangen, FAU Erlangen-Nuremberg, Erlangen, Germany

²⁸Comprehensive Cancer Center Erlangen-EMN (CCC), University Hospital Erlangen, FAU Erlangen-Nuremberg, Erlangen, Germany

²⁹Medical School, Jianggang Shan University, Jiangxi, China

³⁰Department of Pathology, the Second Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

³¹Precision Medicine Centre of Excellence, Health Sciences Building, The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK

³²Regional Molecular Diagnostic Service, Belfast Health and Social Care Trust, Belfast, UK

³³The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK

³⁴Department of Cellular Pathology, Belfast Health and Social Care Trust, Belfast, UK

³⁵Centre for Public Health, Queen's University Belfast, Belfast, UK

³⁶Integrated Pathology Unit, Institute for Cancer Research and Royal Marsden Hospital, London, UK

³⁷Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany

³⁸German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

³⁹Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany

⁴⁰School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

⁴¹Institute of Pathology, Technical University Munich, Munich, Germany

⁴²Institute of Pathology Munich-North, Munich, Germany

⁴³Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg

⁴⁴These authors contributed equally

⁴⁵Lead contact

*Correspondence: tingying.peng@helmholtz-munich.de (T.P.), jakob_nikolas.kather@tu-dresden.de (J.N.K.)

<https://doi.org/10.1016/j.ccell.2023.08.002>

step in providing treatment as recommended by various medical guidelines, such as those in the USA (NCCN guideline),¹⁰ UK (NICE guideline),¹¹ and EU (ESMO guideline).¹² Increasingly, genetic biomarkers such as MSI are also used in earlier tumor stages of CRC.¹³ In the future, the importance of biomarker-stratified therapy will likely increase.¹⁴ The presence of MSI should also trigger additional diagnostic processes for a possible diagnosis of Lynch syndrome, one of the most prevalent hereditary cancer syndromes. However, genetic diagnostic assays have several disadvantages. For many patients in low- and middle-income countries, genetic biomarkers are not routinely available due to the prohibitive costs and complex infrastructure required for testing. Even in high-income countries with universal healthcare coverage where genetic biomarkers may be routinely available, their utilization is not without its drawbacks. In such contexts, biomarker assessment can take several days to weeks delaying therapy decisions.¹⁵

The diagnosis of CRC requires a pathologist's histopathological evaluation of tissue sections. Thus, tissue sections stained with hematoxylin and eosin (H&E) are routinely available for all patients with CRC. Since 2019, dozens of studies have shown that deep learning (DL) can predict genetic biomarkers directly from digitized H&E-stained CRC tissue sections.^{1,3,7,8,16,17} Based on these studies, the first commercial DL algorithm for

biomarker detection from H&E images has been approved for routine clinical use in Europe in 2022 (MSIntuit, Owkin, Paris/New York).¹⁸ When evaluated in external patient cohorts, the state-of-the-art approaches reach a sensitivity and specificity of 0.95 and 0.46, respectively.¹⁹ Increasing the specificity would be a way to improve these established approaches. Another clinically significant limitation of current approaches is the poor performance on endoscopic biopsy tissue. Recent clinical trials (NICHE²⁰ and NICHE-2¹³) show high efficacy of neoadjuvant immunotherapy for patients with MSI CRC. These findings imply that in the future every patient with CRC should be tested for MSI on the initial biopsy tissue, although not all current medical guidelines reflect this. Among previous DL-based studies for MSI detection, only Echle et al.³ determine the performance of DL-based biomarker prediction on CRC biopsy tissue in a multicentric setting. They report a much lower performance on biopsy tissue than on surgical resection tissue sections (biopsy AUROC: 0.78; resection AUROC of 0.96). Current clinically approved commercial products for MSI detection in CRC from histopathology are only applicable to surgical resection tissue. Therefore, DL-based MSI testing of biopsies is a clinical need.

The technology underlying these algorithms in literature is based on weakly supervised learning, consisting of two components: the *feature extractor* and the *aggregation module*.²¹ The

feature extractor is mostly based on a convolutional neural network (CNN), which processes multiple small tissue regions called tiles or patches.²² The CNN-based representations obtained from these tiles are subsequently aggregated to obtain a single prediction for the patient. Between 2019 and 2021, most studies used simple heuristics, such as taking the maximum (max pooling) or averaging (average pooling), as an aggregation module. Since then, variations of multiple instance learning (MIL) have become the new standard for this task, particularly for the prediction of genetic alterations from pathology slides.^{6,23,24} The most common approach replaces the pooling aggregation with a small two-layer network to learn the patch-level weighting of the embeddings.²³ However, current MIL approaches univariately consider a single tile during aggregation and do not place it in context with other tiles even though local and global contexts are crucial for medical diagnosis.

In many non-medical and medical image-processing tasks, transformer neural networks have recently been adopted for computer vision tasks,^{25–27} replacing CNNs because of their improved performance and robustness.²⁸ Originally proposed for sequencing tasks such as natural language processing, transformer networks show impressive capabilities of learning long-range dependencies and contextualizing concepts in long sequences. In computational pathology, transformers have therefore been proposed as potentially superior feature extractors²⁹ or aggregation models,^{30–33} though these proposals still lack empirical evidence from large-scale analyses.

In this work, we first aim to enhance the performance of DL-based biomarker detection from pathology slides. Thereafter, in order to provide large-scale evidence of the performance on clinically relevant tasks, we investigate the use of a fully transformer-based workflow in CRC. Here, we present a new method derived from a transformer-based feature extractor and a transformer-based aggregation model (Figure 1A–C), which we evaluate in a large multi-centric study of 15 cohorts with resection specimen slides from over 13,000 patients with CRC worldwide, as well as two cohorts of CRC biopsies from over 1,500 patients in total (Figure 1D–F).

RESULTS

Transformer-based MSI prediction outperforms the state-of-the-art

We tested our pipeline on MSI prediction in surgical resection cohorts of patients with CRC (Figure 1) in two ways: First, we trained the model on a single cohort and tested it on a held-out test set (in-domain) and on all other cohorts (external). We found that large cohorts, e.g., DACHS, QUASAR, or NLCS, achieved in-domain test AUROCs around 0.95 (Figure 2A). The model also achieved high performance close to 0.9 AUROC for early onset CRC, i.e., CRC in patients younger than 50 years (Figure S2B). We compared this performance to the work by Echle et al.³ which updated the CNN-based feature extractor during training and used mean pooling as their patch aggregation function. Our approach outperformed the CNN-based approach on all four cohorts. Further, we also evaluated AttentionMIL by Ilse et al.²³ with CTransPath as a feature extractor yielding higher performance than the CNN-based approach on the large cohorts but partly lower results on the external validation trained on the

smaller cohort TCGA. Overall, we observed the tendency of higher performance and better generalization for models trained on datasets with more than 1,000 patients. However, factors such as differing population genetics (e.g., for MECC) or the type of slide scanners (e.g., for ERLANGEN) influence the generalization capabilities beyond the training dataset size.

Second, we trained our model on all cohorts of CRC resections except YCR-BCIP and evaluated it on the external cohort YCR-BCIP (Figure 2B). In particular, we obtained a sensitivity of 0.99 with a negative predictive value of over 0.99 (Figure S2F, and Table S1). Analyzing the ROCs of patients with different clinicopathological properties showed that the model performs consistently well on all of these subgroups. Only on left-sided tumors the performance slightly dropped to 0.93 AUROC (Figure S2D). Moreover, a high-mean AUPRC score of 0.86 showed that the transformer-based model achieved high sensitivity with high precision despite a strong class imbalance of 12.9% MSI-high samples on average across all cohorts (Figure 2C). In parallel to our findings mentioned previously, we observed a generalization gap when intrinsic biological factors, such as ethnicity, change. However, the performance of our model on a cohort of MSI-high patients from Guangzhou, China, was still high with a sensitivity of 0.9. For a better comparison to state-of-the-art, we also mirrored the experimental setup of Echle et al.³ We trained AttentionMIL and our fully transformer-based model on the four large cohorts (DACHS, NLCS, QUASAR, and TCGA) using the same feature extractor CTransPath. The CNN-based approach from Echle et al. achieved an AUROC of 0.96, AttentionMIL yielded an AUROC of 0.96, and the fully transformer-based approach performed slightly better with an AUROC of 0.97.

The classical patch-based approach by Echle et al.³ suffered from severe losses in performance upon external testing. The largest performance drop in the AUROC of 0.21 was observed by a model trained on the DACHS and tested on the QUASAR cohort. Our transformer model, however, reduced the performance loss for external testing to a maximum of 0.09 for training on the NLCS and testing on the TCGA cohort (Figure 2). In addition, AttentionMIL trained with the same transformer-based feature extractor also demonstrated better generalization capabilities compared to the classical patch-based approach with mean pooling. This suggests that self-supervised pre-training on histology data contributes positively toward better generalization.

In summary, these results show that a fully transformer-based approach yields a higher performance for biomarker prediction both on large cohorts (DACHS, QUASAR, and NLCS) as well as on smaller cohorts (TCGA). Perhaps more importantly from a clinical perspective, the transformer-based approach resulted in a better generalization performance and more reliable results, as the deviation between the external cohorts was smaller. We published all trained models for reuse and further fine-tuning if needed.

The transformer-based model predicts multiple biomarkers in CRC

Next, we investigated whether the fully transformer-based model yields a similar high performance in other biomarker-prediction tasks. Following the experimental setup for MSI prediction, we

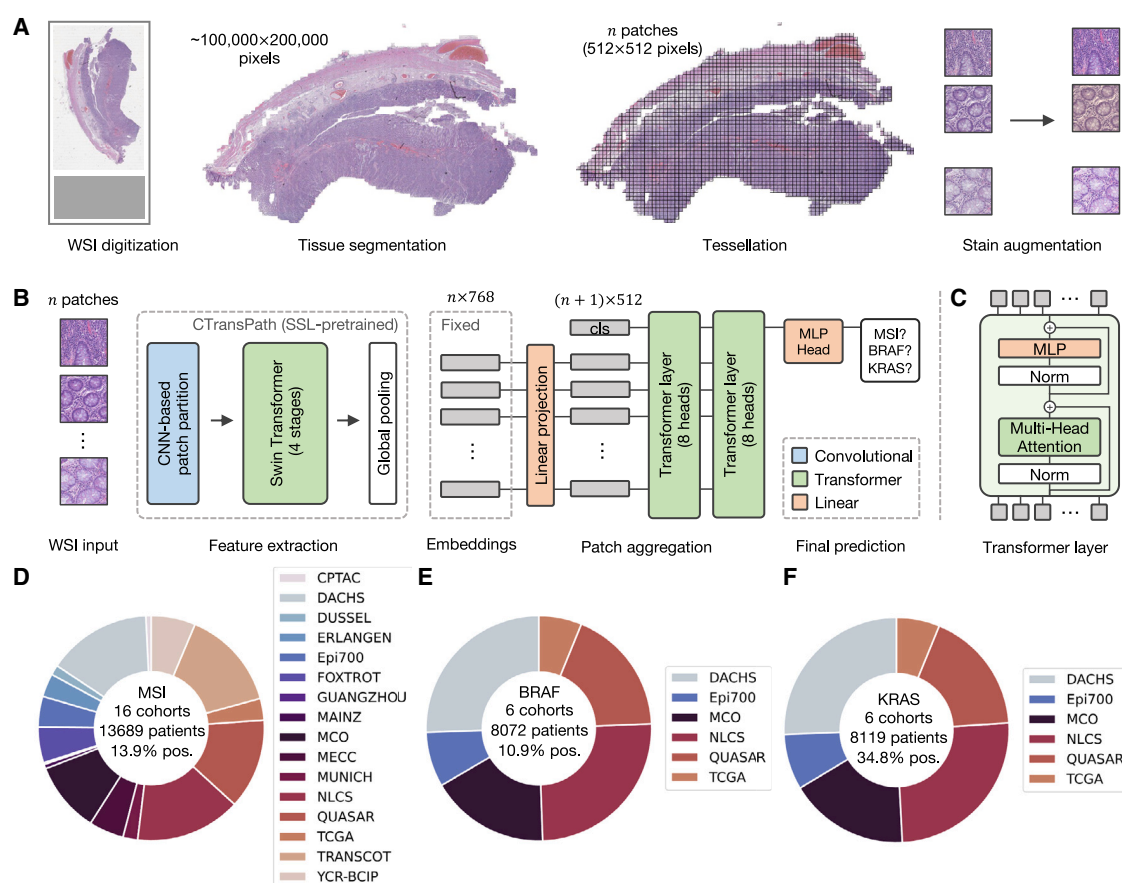


Figure 1. Workflow overview with pre-processing and model architecture and cohort overview

(A) The data pre-processing pipeline with the steps whole slide image (WSI) digitization, tissue segmentation, WSI tessellation into patches, and stain augmentation, (B) the model architecture including the pre-trained feature extractor CTransPath and our transformer-based aggregation module, and (C) details of the transformer layer architecture.

(D) Overview of the 16 cohorts of CRC resections and biopsies with MSI/dMMR status, which were used in this study and the subsets of six cohorts with (E) *BRAF* and (F) *KRAS* ground truth data, respectively.

trained the model first on single cohorts evaluated on all other external cohorts and second on one fully merged multi-center cohort excluding only one cohort from the training set to constitute an external test set. In clinical routine workup for CRC, the biomarkers *BRAF* and *KRAS* are determined in addition to MSI. We tested whether and how well these were predictable on the DACHS, QUASAR, MCO, NLCS, TCGA, and Epi700 cohorts, where the Epi700 cohort served as an external test set in the multi-centric run.

In the case of the largest cohorts, DACHS and NLCS, single cohort training was already capable of achieving good results, with AUROCs of 0.88 and 0.87, respectively (Figure 2D). The smaller cohorts achieved slightly poorer results with 0.83–0.85 AUROC and 0.78 for TCGA. Nonetheless, the AUROC for the in-domain test using TCGA by far outperformed previous approaches with AUROCs of 0.57,⁶³ 0.66,⁶⁴ and 0.73³³ in a more recent transformer-based method. The large multi-centric cohort yielded an AUROC of 0.88, almost reaching clinical-grade performance (Figure 2E). Furthermore, we observed that the generalization gap from the internal test set to external cohorts

was consistently small with the largest internal-to-external gap of 0.03 drop in AUROC. This was also the case in multi-centric evaluation, where the performance did not decrease from the internal to the external test cohort.

We observed similar results regarding the generalization when investigating *KRAS* as a target (Figures 2F and 2G), with an AUROC of 0.80 when trained on the multi-centric cohort outperforming state-of-the-art methods. The AUROCs of the single cohort training ranged from 0.53 to 0.77 for single cohorts, in line with or higher than state-of-the-art results.^{33,63,64} While DL-based prediction performance for *KRAS* is still relatively low compared to MSI or *BRAF*, the results show that performance profits substantially from multi-cohort training and larger training cohorts.

Overall, these findings show that our model can predict multiple biomarkers that are relevant for routine diagnostics in CRC while highlighting the importance of large training cohorts to reach clinically relevant performance even in biomarkers such as *KRAS* which are notoriously difficult to predict from pathology images alone.

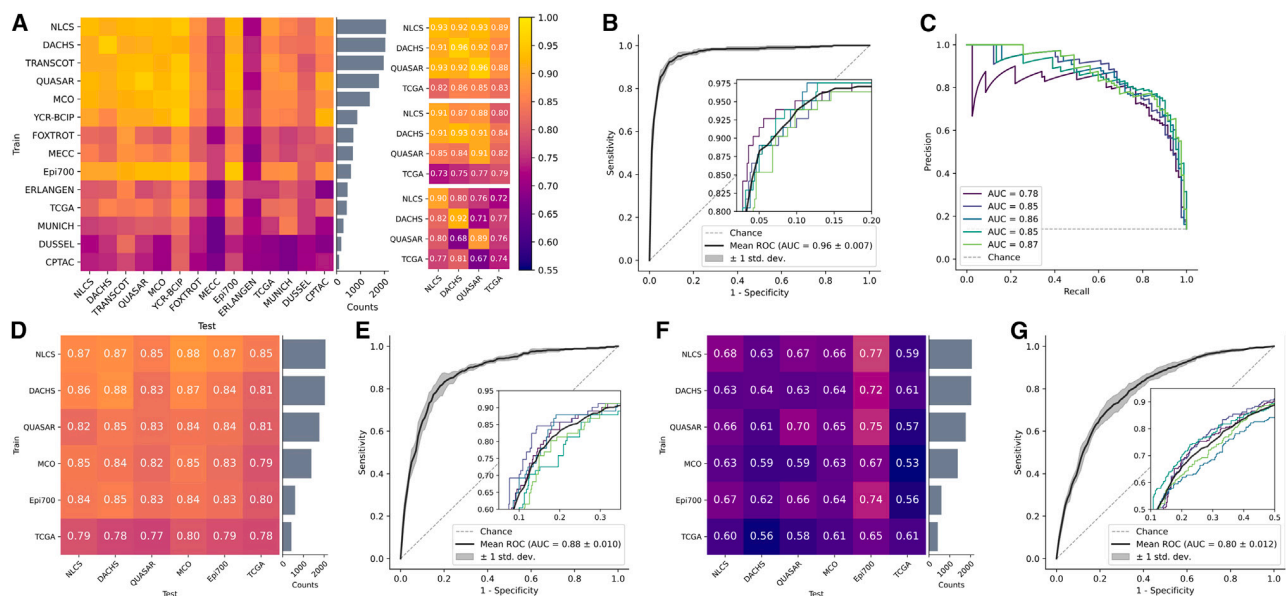


Figure 2. Evaluation of the prediction performance for the biomarkers MSI, BRAF, and KRAS in single cohort and large-scale multi-centric experiments

Experimental results for MSI-high (A–C), BRAF (D,E), and KRAS (F,G) prediction. All values represent the mean of 5-fold cross-validation: (A) AUROC scores for single cohort experiments for all CRC cohorts ordered by size of the cohort. Each row shows the test performance of training on one cohort with the in-domain test results in the diagonal. Results for our transformer approach, AttentionMIL, and CNN approach (results taken from Echle et al.) are visualized separately. Note that compared to AttentionMIL and CNN, our transformer not only shows higher overall prediction accuracy but also better model generalizability, demonstrated by a smaller gap between internal and external testing cohorts. Raw data for the heatmap in Table S5.

(B) Receiver operator curve (ROC) for the model trained on all resection cohorts except YCR-BCIP, tested on YCR-BCIP.

(C) Precision recall curve (PRC) for the model trained on all resection cohorts except YCR-BCIP, tested on YCR-BCIP.

(D) AUROC scores for single cohort experiments.

(E) ROC for the model trained on all BRAF cohorts except Epi700, tested on Epi700.

(F) AUROC scores for single cohort experiments.

(G) ROC for the model trained on all KRAS cohorts except Epi700, tested on Epi700.

Transformer-based workflows are explainable

Ideally, DL-based biomarker predictions should be explainable to domain experts. To this end, we visualized how much each patch contributed to the final classification via attention rollout as well as whether it contributes toward a positive or negative classification (Figures 3A–3C, and S3).

For better comparability, we used the same WSIs from the external cohort YCR-BCIP as had been used in a previous study³ for these visualizations (Figure 3A). For all three cases, the majority of highly contributing patches originate from tumor regions. In the MSI-high case, the mucinous region that is morphologically linked to MSI is correctly identified as important by high scores in the attention rollout as well as the patch-level classification (see the boxes in the first row of Figure 3). The MS-stable case in the second row of Figures 3A–3C attributes high contributions to the model's prediction to carcinoma regions. At the same time, these patches all receive low-classification scores yielding the correct classification result. Similarly, the second MS-stable case in the third row had highly contributing scores in the tumor region while having only low-classification scores for all patches. Further tissue details that are morphologically related to MSI, such as solid growth pattern, poor differentiation, or tumor-infiltrating lymphocytes (Figure S3A) are highlighted in the attention heads of the last layer together with healthy tissue structures,

such as the colon wall including muscle tissue or vessels (Figure 3D). The two cases with MSI-high ground truth predicted as MSS also show that relevant regions are identified and contribute to the prediction but the combination of potentially false attentions and associated classification scores of these attentive patches infer a wrong prediction (Figure S3B).

We quantified the morphological patterns occurring in high-attention regions in a small user study, where two pathologists annotated the patterns in 160 patches of 40 patients that the model attributes high attention to. We chose the 10 patients with the lowest and highest classification score for each MSS and MSI-high ground truth. For every patient, we chose the two patches associated with the highest and lowest class scores of the top 100 attention tiles. Our study showed that the majority of tiles belong to the tumor region (0.99% for high and 0.81% for low-classification scores) and cell types that are important for the prediction of MSI-high, such as lymphocytes occur in both tiles with low- and high-classification score in a similar ratio (0.28% vs. 0.2%, Figure 4A). Furthermore, morphological patterns related only to MSI-high, such as mucinous regions, occur more often in tiles with high-classification scores (0.4% vs. 0.1%). A chi-square test for independence shows that the underlying distributions of tiles with high- and low-classification scores are likely to be independent (p value = $5.6 \cdot 10^{-6} < 5 \cdot 10^{-5}$).

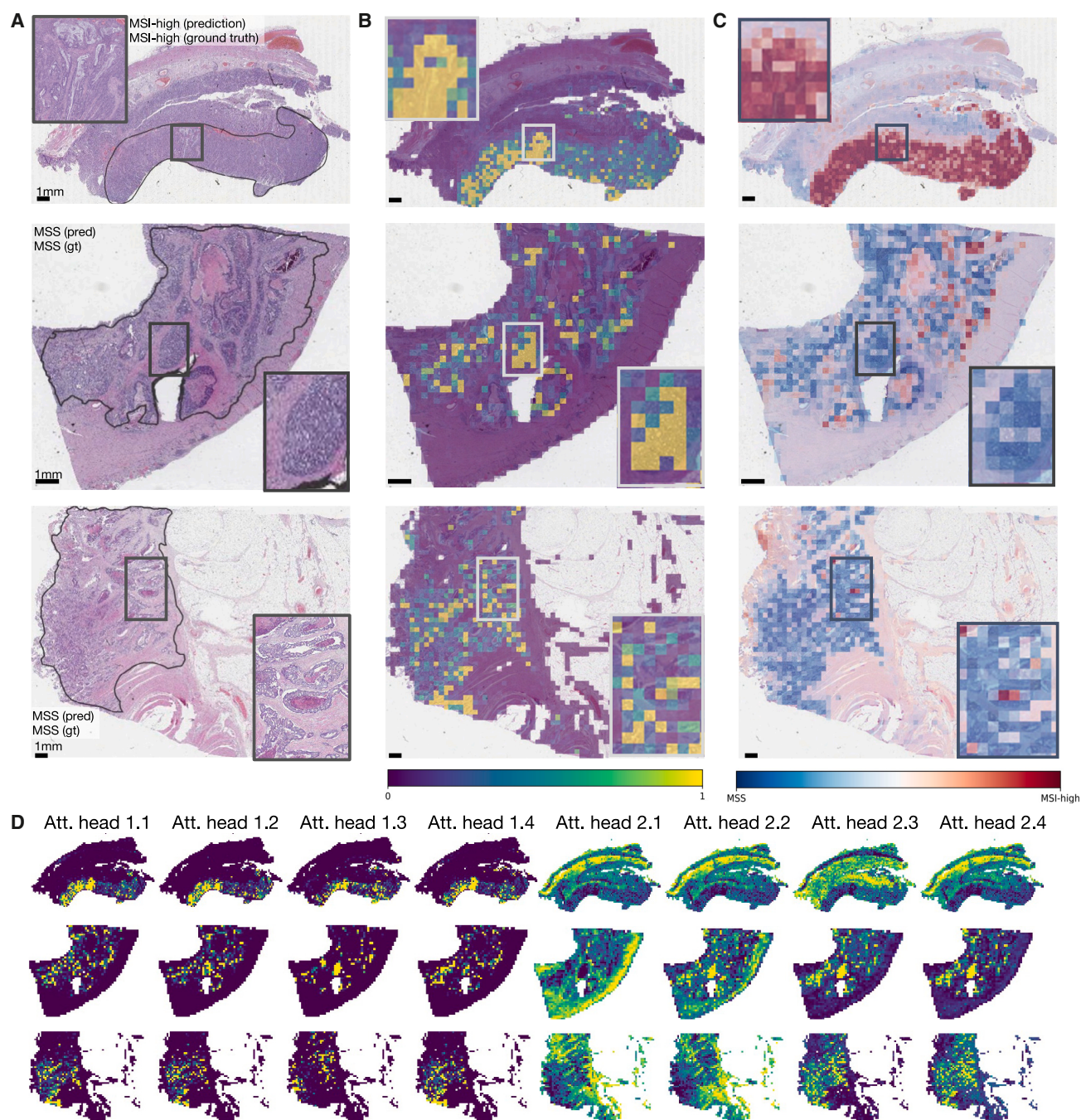


Figure 3. Attention and class score visualization for better model interpretability

(A) Resection specimen from the external cohort YCR-BCIP. The three depicted slides are the same as in Echle et al.³ Tumor regions are outlined in black. (B) Attention rollout per patch for our trained transformer-based feature aggregation model. Large values (yellow) signify a high contribution to the model's prediction, small values (purple) a low contribution. (C) MSI classification scores per patch, where MSI-high is the positive class and MS-stable is the negative class. (D) The attention heatmaps from eight heads, four of the first and four of the second layer. The model weights are taken from the best-performing fold of the multi-centric training on all cohorts except YCR-BCIP.

These examples showed that the model learns concepts relevant to MSI-high prediction and thus possesses a high degree of explainability. Visualizing the attention rollout together with the

classification scores demonstrates that relevant regions can receive high-attention scores while the model can learn to ignore non-relevant regions or give them low-classification scores.

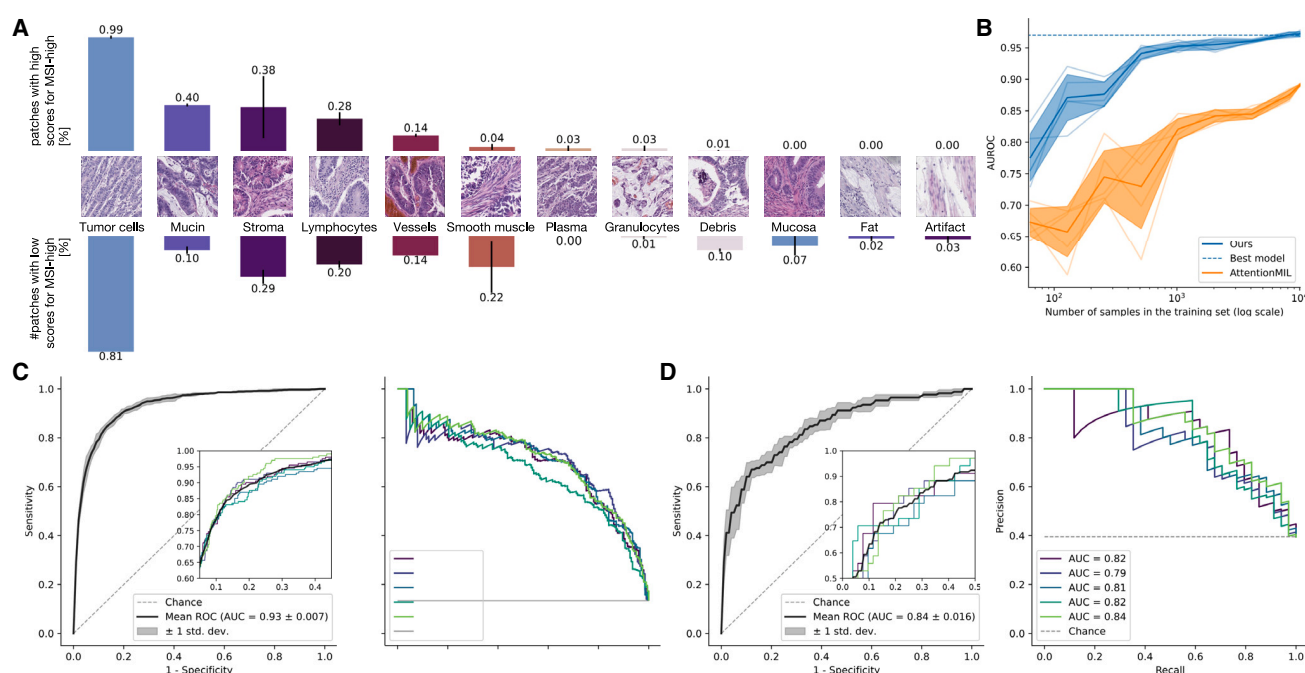


Figure 4. Analysis of the quantitative user study on high-attention tiles, data efficiency, and model generalization to biopsies

(A) Prevalence of 12 histological patterns in 160 patches of high attention regions from 40 patients split by low and high patch-wise classification scores.

(B) AUROC scores on YCR-BCIP depending on the number of patients available for training. The samples were randomly drawn from all resection cohorts except YCR-BCIP.

(C) ROC and PRC for testing our model on YCR-BCIP-biopsies, trained on resections from all cohorts except YCR-BCIP.

(D) ROC and PRC for testing our model on biopsies of the cohort MAINZ, trained on resections from all cohorts except YCR-BCIP.

Transformer-based workflows are more data efficient

A long-standing problem in computational pathology is to determine the number of samples required for a given prediction task. This is primarily due to two reasons. First, it is unclear what the minimum required sample size is, and second, it is unclear if adding more samples improves performance, and if so, up to what point. To investigate this, we varied the number of patients in the training set and analyzed its impact on the test performance. Specifically, we merged all cohorts except for an external validation cohort, YCR-BCIP, resulting in a training set with 8,181 patients from nine cohorts. We trained models using a fixed number of epochs and randomly sampled patients from the training set. All experiments were repeated five times, and we reported the means and standard deviations of the results.

Our fully transformer-based model architecture achieved a mean testing AUROC value above 0.9 with 250 patients (in particular, an AUROC value of 0.92), while the AttentionMIL model exceeded an AUROC of 0.9 only with 4,000 patients (Figure 4A). In a similar vein, our model surpassed the 0.95 mean testing AUROC with already 1,500 patients, while AttentionMIL did not reach this performance. Hence, the transformer-based aggregation module helped the model to learn from data in a more efficient way than the attention mechanism. This may be due to the attention mechanism not contextualizing information from all input patches. Of note, above 1,000 patients, the performance of the transformer-based model seems to slowly saturate, while the attention mechanism continues to increase in performance with more patients but on a lower level.

Our fully transformer-based approach yielded high performance with a small sample size. Compared to an AttentionMIL-based approach, our approach is more data efficient in the regime of small numbers of patients. Looking at larger training numbers, we observed that performance increase is directly proportional to the number of patients for both approaches, but the fully transformer-based approach reaches equivalent performance already with much smaller datasets.

Transformer-based workflows result in clinical-grade performance on biopsies

Virtually all previous studies on biomarker prediction in CRC were performed using surgical resection slides. For this reason, commercially available MSI detection algorithms are intended to be used only with resection slides. However, recent clinical evidence shows that MSI-positive patients with CRC require immunotherapy prior to surgery^{13,65} and hence need to be tested for MSI on biopsy material. We addressed this problem by training our model on resections from all cohorts except YCR-BCIP and evaluating it on biopsies from 1,592 patients with CRC of the YCIP-BCIP.

Our model yielded a mean AUROC score of 0.92 and 0.86 when validated on biopsies from two external cohorts, YCR-BCIP and MAINZ, respectively (Figure 4B). It is worth mentioning that the MSI-high ratio in the MAINZ biopsy cohort was higher than in the training cohorts. We outperformed existing approaches (0.78 by Echle et al.³) by far and achieved clinical-grade performance on biopsies after model training on resection

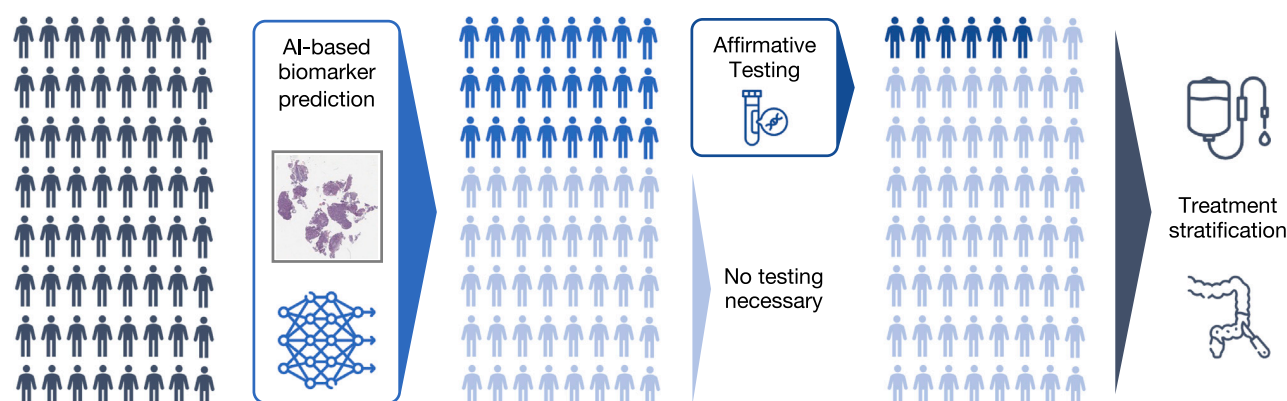


Figure 5. Envisioned clinical workflow for the proposed MSI-high classifier on biopsies

This assumes the system reaches a sufficient performance in additional external validation and is approved as a medical device. This workflow would only apply to non-metastatic disease. Neoadjuvant immunotherapy is not yet recommended by medical guidelines but is backed up by Phase-II clinical trials. Not shown: tissue preprocessing and scanning pipelines and confirmatory tests of MSI-high after a positive deep learning-based pre-screening.

specimens (Figure S4). The mean AUPRC score of 0.69 and 0.82, respectively, however, was lower than that of the external cohort YCR-BCIP for resections (0.86) (Figure 4C). Hence, choosing a classification threshold with high sensitivity, the ratio of correctly MSI positive predicted cases from all positive predicted cases was lower on biopsies compared to resections. Still, with a classification threshold fixed on the in-domain test set of resections, our model obtains sensitivity scores of 0.98 and 0.91, respectively, with negative predictive values of 0.99 and 0.9. Of note, these values are higher than (for the cohort YCR-BCIP) and close to (for the MAINZ cohort) the clinically approved DL algorithm for resections,¹⁹ suggesting that our algorithm has potential for clinical usage for biopsies.

Our intended clinical use of this workflow is as follows (Figure 5): First, a patient attends a clinic either with suspected CRC or for routine CRC screening. A colonoscopy shows a suspicious tumor, which is evaluated histologically and found to be an adenocarcinoma. In many countries, this biopsy will then be tested for MSI/MMR status and *BRAF* and/or *RAS* mutation status. In practice, these procedures may take several days to even weeks. However, in low- or middle-income countries, this might not happen at all. Based on the MSI, *BRAF*, and *RAS* status, the most suitable treatment approach will be chosen for the patient. For example, in patients with early (non-metastatic) CRC, the presence of MSI could qualify a patient for neoadjuvant immunotherapy followed by surgery with curative intent. Similarly, in the metastatic disease setting, the presence of MSI in the biopsy tissue would qualify a patient for palliative immunotherapy. Because of its high sensitivity, our algorithm could serve as a filtering step followed by affirmative testing for MSI-high predicted cases. Applying AI-based biomarker prediction would reduce the additional testing burden and therefore speed up the step between taking the biopsy and the molecular determination of MSI-high status, thus enabling an earlier treatment with immunotherapy if indicated.

In summary, to the best of our knowledge, we developed a DL-based MSI-high predictor for biopsies that achieves clinical-grade performance. In particular, this high performance was also observed for external tests and could therefore improve clinical routine and speed up treatment decisions.

DISCUSSION

The rollout of precision oncology to patients with CRC promises gains in life expectancy.⁶⁶ Unfortunately, however, its implementation still remains slow and patchy. One reason for this is that precision oncology biomarkers are complex, costly, and require intricate instrumentation and expertise. DL is emerging as a possible solution for this problem.^{22,67} DL can extract biomarker information directly from routinely available material, thereby potentially providing cost savings.¹ Using DL-based analysis of histopathology slides to extract biomarkers for oncology has become a common approach in the research setting in 2018.⁶⁸ In turn, this has recently led to regulatory approval of multiple algorithms for clinical use. Some of these examples include a breast cancer survival prediction algorithm by Paige (New York, NY, USA), a method to predict survival in CRC by DoMore Diagnostics (Oslo, Norway), a method to predict MSI status in CRC by Owkin (Paris, France, and New York, NY, USA), among others.^{18,69} However, existing DL biomarkers have some key limitations: it is debated whether or not their performance is sufficient for large-scale use, they do not necessarily generalize to any patient population, and finally, they are not approved for use on biopsy material, as the application of DL algorithms to biopsies typically results in much lower performance compared to application to surgical specimens.³

A key reason for the limited performance of existing DL systems could be the fundamental limitations of the technology employed. Most studies between 2018 and 2020 used convolutional neural networks (CNNs) as their DL backbone,³¹ using publicly available information. Commercial products in the DL biomarker space are based on the same technology.^{19,69,70} However, a new class of neural networks has recently started to replace CNNs: transformers. Originating from the field of natural language processing, transformers are a powerful tool to process sequences and leverage the potential of large amounts of data. Also in computer vision, transformers yield a higher accuracy for image classification in non-medical tasks,^{25,26} are more robust to distortions in the input data²⁸ and provide more detailed explainability.³⁰ These advantages of transformers

compared to CNNs have the potential to translate into more accurate and more generalizable clinical biomarkers, but there is currently no evidence to support this.

In the present study, we developed a transformer-based approach for biomarker prediction on whole-slide images of H&E-stained CRC tissue sections. Our model consists of a transformer-based feature extractor that was pretrained on histopathology images and a transformer-based aggregation module. In contrast to the state-of-the-art attention-based MIL approaches, the contribution of each patch was not only determined according to its feature embeddings but also contextualized with the feature embeddings of all other patches in the WSI via self-attention layers. Further, we presented a large-scale evaluation of transformers in biomarker prediction on WSIs. We demonstrated that transformer-based approaches learned better from small amounts of data and were therefore more data efficient than attention-based MIL approaches. At the same time, the performance increased proportionally with the number of training samples. Even though the performance seemed to plateau for MSI prediction, this suggests that larger training cohorts could lead to higher performance-approaching clinical application, also for more challenging tasks such as the prediction of the *BRAF* and *KRAS* mutational status. Our large-scale evaluation also showed that MIL and in particular transformer-based approaches generalize much better than the existing CNN approaches. We proved this by training the model on single cohorts and testing the generalization on all other cohorts. These experiments showed that the transformer-based approach reduced the drop in AUROC to under 0.09, while the CNN-based approach dropped by more than 0.21 in some cohorts. Most importantly, our approach trained on resections did not only generalize well to external cohorts of resections from geographically distinct regions but also to biopsies with a clinical-grade performance of 0.98 sensitivity on the YCR-BCIP biopsy cohort and 0.91 on the MAINZ cohort.

A caveat of our observations is that the ground truth might not be perfect. Potentially, the DL model is performing better than stated in the paper because dMMR and MSI only agree around 92% of the time and neither are 100% sensitive.^{57,71} Also, a small subset of CRCs have *POLD1* and *POLE* mutations with a high-mutation burden that behaves clinically similar to MSI and might have a similar phenotype but are not detected by established MSI detection assays.⁷² Similarly, gene sequencing does not detect all mutations in *KRAS/NRAS/BRAF* depending on the sensitivity of the technology used and the presence of smaller clonal mutations. Current DL tests are at such high levels of performance that these nuanced subpopulations may be important. Our study has additional limitations: The focus of this study was to investigate the effect of handling the data with fully transformer-based approaches, especially in the context of large-scale multi-institutional data. Therefore, we did not exhaustively optimize every single hyperparameter. Points for optimization in this direction would be finding a fitting positional encoding and tuning the architecture of the transformer network and attention mechanisms. Additionally, collecting biopsy samples from different hospitals, for multi-cohort training directly on biopsy data could potentially improve the performance of our model on biopsy material. This would also hold for the prediction of *BRAF* mutation status and, in particular, of *RAS* mutation status,

where we observed the largest potential for improvement. In both targets, the performance was higher on the larger cohorts with around 2,000 patients and increased dramatically by training on multiple cohorts. Further, we acknowledge that achieving an even higher specificity would be desirable. Choosing the final classification threshold is always a trade-off between sensitivity and specificity, where clinical application prefers a higher sensitivity, especially for pre-screening test as proposed in this study. Our method's performance on biopsies is in the same range as current clinically approved assays on resections, but unlike these assays, our method also works on biopsies.

In summary, to the best of our knowledge, we presented a fully transformer-based model to predict MSI on WSI from CRC with an AUROC of 0.97 on resections and 0.92 and 0.86 on biopsies on external validation cohorts. Our model generalized better to unseen cohorts and was more data efficient compared to existing state-of-the-art MIL or CNN approaches. By publishing all trained models, we enable researchers and clinicians to apply the automated MSI prediction tool for research purposes, which we expect to bring the field of DL-based biomarkers a step closer to large-scale integration in the clinical workflow.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
 - Lead contact
 - Materials availability
 - Data and code availability
- **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**
 - Ethics statement
 - Cohort description
- **METHOD DETAILS**
 - Model description
 - Experimental setup and implementation details
 - Visualization and explainability
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.ccell.2023.08.002>.

CONSORTIA

We acknowledge the support of the Rainbow-TMA Consortium, especially the project group: PA van den Brandt, A zur Hausen, HI Grabsch, M van Engeland, LJ Schouten, J Beckervordersandforth; PHM Peeters, PJ van Diest, HB Bueno de Mesquita; J van Krieken, I Nagtegaal, B Siebers, B Kiemeny; FJ van Kemenade, C Steegers, D Boomsma, GA Meijer; FJ van Kemenade, B Stricker; L Overbeek, A Gijsbers; and Rainbow-TMA collaborating pathologists, among others: A de Bruijne, JC Beckervordersandforth; J van Krieken, I Nagtegaal; W Timens; FJ van Kemenade; MCH Hogenes; PJ van Diest; RE Kibbelaar; AF Hamel; ATMG Tiebosch; C Meijers; R Natté; GA Meijer; JJTH Roelofs; RF Hoedemaeker; S Sastrowijoto; M Nap; HT Shirango; H Doornwaard; JE

Boers; JC van der Linden; G Burger; RW Rouse; PC de Bruin; P Drillenburger; C van Krimpen; JF Graadt van Roggen; SAJ Loyson; JD Rupa; H Kliffen; HM Hazzelbag; K Schelfout; J Stavast; I van Lijschoten; and K Duthoi.

ACKNOWLEDGMENTS

SJW and DR are supported by the Helmholtz Association under the joint research school “Munich School for Data Science - MUDS” and SJW is supported by the Add-on Fellowship of the Joachim Herz Foundation. JNK is supported by the German Federal Ministry of Health (DEEP LIVER, ZMV11-2520DAT111), the Max-Eder-Programme of the German Cancer Aid (grant #70113864), the German Federal Ministry of Education and Research (CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (Transplant.KI, 01VSF21048) and the European Union’s Horizon Europe and innovation program (ODELIA, 101057091; GENIAL, 101096312). JNK and MH are funded by the German Federal Ministry of Education and Research (PEARL, 01KD2104C). PQ, NW, SD, and GH are supported by Yorkshire Cancer Research grants L386 and L394. PQ, HG, NW, JNK, and SD are supported in part by the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Center. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. PQ is also supported by an NIHR Senior Investigator award. FOxTROT was funded by Cancer Research UK (grant reference: C551/A8283; recipient: D.M.). Additional support was provided by the Birmingham and Leeds ECMC network, the RCS Eng and Rosetrees Trust, and the Swedish Cancer Society. Panitumumab was provided free of charge by Amgen, who also supported RAS testing and additional CT scans (recipient: D.M.). P.Q., N.W., and M.S. are supported by Yorkshire Cancer Research, R.G. by the Medical Research Council. Tumor tissue collection in the NLCS was done in the Rainbow-TMA study, which was financially supported by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO 184.021.007 to PvdB). The analyses of MSI, *BRAF*, and *KRAS* in the NLCS were funded by The Dutch Cancer Society (KWF 11044 to PvdB). The DACHS study (HB, JCC, and MH) was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, HO 5117/2-1, HO 5117/2-2, HE 5998/2-1, HE 5998/2-2, KL 2354/3-1, KL 2354/3-2, RO 2270/8-1, RO 2270/8-2, BR 1704/17-1 and BR 1704/17-2), the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT; Germany) and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A and 01ER1505B). The study was further supported by project funding for the Pearl consortium from the German Federal Ministry of Education and Research (01KD2104A). SF is supported by the Deutsche Forschungsgemeinschaft (DFG) (FO 942/2-1), the German Federal Ministry of Education and Research (SWAG, 01KD2215A), the Mainz Research School of Translational Biomedicine (TransMed) and the Manfred-Stolte-Foundation. CM is supported by the Interdisciplinary Center for Clinical Research (IZKF) at the University Hospital of the University of Erlangen-Nuremberg (Junior Project J101). This work was supported in part by NIH R01 CA263318 (SG).

DACHS study: The authors thank the hospitals recruiting patients for the DACHS study and the cooperating pathology institutes. We thank the National Center for Tumor Diseases (NCT) Tissue Bank, Heidelberg, Germany, for managing, archiving, and processing tissue samples in the DACHS study.

The SCOT trial was funded by the Medical Research Council (transferred to NETSCC - Efficacy and Mechanism Evaluation) (Grant Ref: G0601705), NIHR Health Technology Assessment (Grant ref. 14/140/84), Cancer Research UK Core CTU Glasgow Funding (Funding Ref: C6716/A9894), and the Swedish Cancer Society. The TransSCOT sample collection was funded by a Cancer Research UK Clinical Trials Awards and Advisory Committee - Sample Collection (Grant Ref: C6716/A13941).

Molecular analysis of the SCOT samples were funded by the Oxford NIHR Comprehensive Biomedical Research Centre (BRC), a Cancer Research UK (CRUK) Advanced Clinician Scientist Fellowship (C26642/A27963) to DNC, CRUK award A25142 to the CRUK Glasgow Center. V.H.K. acknowledges funding by the Preeclampsia Foundation (F-87701-41-01). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health.

We furthermore acknowledge all collaborators in the YCR-BCIP and FOxTROT studies, as well as all other collaborators for all other cohorts included in this study.

AUTHOR CONTRIBUTIONS

S.J.W., M.B., T.P., and J.N.K. designed the concept of the study, S.J.W., J.M.N., and J.Z. prepared the data for this study, S.J.W. implemented the methods and ran all experiments and evaluations; S.F., C.M., C.P., M.B., and G.P.V. supported with domain-related advice; D.R., S.J.W., M.B., and T.P. developed and evaluated a preliminary study, where D.R. implemented the methods; all authors provided clinical and histopathological data and expertise; all authors provided clinical expertise and contributed to the interpretation of the results. S.J.W. wrote the manuscript with J.N.K. and input from all other authors.

DECLARATION OF INTERESTS

J.N.K. reports consulting services for Owkin, France, Panakeia, UK, and Do-More Diagnostics, Norway and has received honoraria for lectures by M.S.D., Eisai, and Fresenius. N.W. has received fees for advisory board activities with BMS, Astellas, GSK, and Amgen, not related to this study. N.W. has received fees for advisory board activities with BMS, Astellas, and Amgen, not related to this study. P.Q. has received fees for advisory board activities with Roche and AMGEN and research funding from Roche through an Innovate UK National Pathology Imaging Consortium grant. H.I.G. has received fees for advisory board activities by AstraZeneca and BMS, not related to this study. M.S.T. is a scientific advisor to Mindpeak and Sonrai Analytics, and has received honoraria recently from BMS, MSD, Roche, Sanofi, and Incyte. He has received grant support from Phillips, Roche, MSD, and Akoya. None of these disclosures are related to this work. D.N.C. has participated in advisory boards for MSD and has received research funding on behalf of the TransSCOT consortium from HalioDx for analyses independent of this study. V.H.K. has served as an invited speaker on behalf of Indica Labs and has received project-based research funding from The Image Analysis Group and Roche outside of the submitted work. No other potential disclosures are reported by any of the authors.

Received: January 4, 2023

Revised: June 18, 2023

Accepted: August 7, 2023

Published: August 30, 2023

REFERENCES

- Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., et al. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* 25, 1054–1056.
- Cao, R., Yang, F., Ma, S.-C., Liu, L., Zhao, Y., Li, Y., Wu, D.-H., Wang, T., Lu, W.-J., Cai, W.-J., et al. (2020). Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer. *Theranostics* 10, 11080–11091.
- Echle, A., Grabsch, H.I., Quirke, P., van den Brandt, P.A., West, N.P., Hutchins, G.G.A., Heij, L.R., Tan, X., Richman, S.D., Krause, J., et al. (2020). Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. *Gastroenterology* 159, 1406–1416.e11.
- Bilal, M., Ahmed Raza, S.E., Azam, A., Graham, S., Ilyas, M., Cree, I.A., Snead, D., Minhas, F., and Rajpoot, N.M. Novel Deep Learning Algorithm Predicts the Status of Molecular Pathways and Key Mutations in Colorectal Cancer from Routine Histology Images. 10.1101/2021.01.19.21250122
- Lee, S.H., Song, I.H., and Jang, H.-J. (2021). Feasibility of deep learning-based fully automated classification of microsatellite instability in tissue slides of colorectal cancer. *Int. J. Cancer* 149, 728–740.
- Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., and Teuwen, J. (2022). DeepSMILE: Contrastive self-supervised pre-training benefits MSI and

- HRD classification directly from H&E whole-slide images in colorectal and breast cancer. *Med. Image Anal.* 79, 102464.
7. Schrammen, P.L., Ghaffari Laleh, N., Echle, A., Truhn, D., Schulz, V., Brinker, T.J., Brenner, H., Chang-Claude, J., Alwers, E., Brobeil, A., et al. (2022). Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *J. Pathol.* 256, 50–60.
8. Yamashita, R., Long, J., Longacre, T., Peng, L., Berry, G., Martin, B., Higgins, J., Rubin, D.L., and Shen, J. (2021). Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol.* 22, 132–141.
9. Jang, H.-J., Lee, A., Kang, J., Song, I.H., and Lee, S.H. (2020). Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. *World J. Gastroenterol.* 26, 6207–6223.
10. Benson, A.B., Venook, A.P., Al-Hawary, M.M., Cederquist, L., Chen, Y.-J., Ciombor, K.K., Cohen, S., Cooper, H.S., Deming, D., Engstrom, P.F., et al. (2018). NCCN Guidelines Insights: Colon Cancer, Version 2.2018. *J. Natl. Compr. Cancer Netw.* 16, 359–369.
11. National Institute for Health and Care Excellence (2020). Colorectal cancer [NICE Guideline No. 151]. <https://www.nice.org.uk/guidance/ng151>.
12. Schmoll, H.J., Van Cutsem, E., Stein, A., Valentini, V., Glimelius, B., Haustermans, K., Nordlinger, B., van de Velde, C.J., Balmana, J., Regula, J., et al. (2012). ESMO Consensus Guidelines for management of patients with colon and rectal cancer: a personalized approach to clinical decision making. *Ann. Oncol.* 23, 2479–2516.
13. Chalabi, M., Verschoor, Y.L., van den Berg, J., Sikorska, K., Beets, G., Lent, A.V., Grootsholten, M.C., Aalbers, A., Buller, N., Marsman, H., et al. (2022). LBA7 Neoadjuvant immune checkpoint inhibition in locally advanced MMR-deficient colon cancer: The NICHE-2 study. *Ann. Oncol.* 33, S1389.
14. Vacante, M., Borzi, A.M., Basile, F., and Biondi, A. (2018). Biomarkers in colorectal cancer: Current clinical utility and future perspectives. *World J. Clin. Cases* 6, 869–881.
15. Lim, C., Tsao, M.S., Le, L.W., Shepherd, F.A., Feld, R., Burkes, R.L., Liu, G., Kamel-Reid, S., Hwang, D., Tanguay, J., et al. (2015). Biomarker testing and time to treatment decision in patients with advanced non-small-cell lung cancer. *Ann. Oncol.* 26, 1415–1421.
16. Niehues, J.M., Quirke, P., West, N.P., Grabsch, H.I., van Treeck, M., Schirris, Y., Veldhuizen, G.P., Hutchins, G.G.A., Richman, S.D., Foersch, S., et al. (2023). Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: A retrospective multi-centric study. *Cell Rep. Med.* 4, 100980.
17. Bilal, M., Raza, S.E.A., Azam, A., Graham, S., and Ilyas, M. (2021). Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal *The Lancet Digital*.
18. Saillard, C., Dubois, R., Tchita, O., Loiseau, N., Garcia, T., Adriansen, A., Carpentier, S., Reyre, J., Enea, D., Kamoun, A., et al. (2022). Blind validation of MSIntuit, an AI-based pre-screening tool for MSI detection from histology slides of colorectal cancer. Preprint at medRxiv.
19. Svrcek, M., Saillard, C., Dubois, R., Loiseau, N., Mespoulhe, P., Brulport, F., Guillon, J., Auffret, M., Sefta, M., Kamoun, A., et al. (2022). 920P Blind validation of MSIntuit, an AI-based pre-screening tool for MSI detection from colorectal cancer H&E slides. *Ann. Oncol.* 33, S967.
20. Chalabi, M., Fanchi, L.F., Dijkstra, K.K., Van den Berg, J.G., Aalbers, A.G., Sikorska, K., Lopez-Yurda, M., Grootsholten, C., Beets, G.L., Snaebjornsson, P., et al. (2020). Neoadjuvant immunotherapy leads to pathological responses in MMR-proficient and MMR-deficient early-stage colon cancers. *Nat. Med.* 26, 566–576.
21. Bilal, M., Jewsbury, R., Wang, R., AlGhamdi, H.M., Asif, A., Eastwood, M., and Rajpoot, N. (2022). An Aggregation of Aggregation Methods in Computational Pathology. Preprint at arXiv. [cs.CV].
22. Shmatko, A., Ghaffari Laleh, N., Gerstung, M., and Kather, J.N. (2022). Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat. Can. (Ott.)* 3, 1026–1038.
23. Ilse, M., Tomczak, J., and Welling, M. (10–15 Jul 2018) Attention-based Deep Multiple Instance Learning. In *Proceedings of the 35th International Conference on Machine Learning Proceedings of Machine Learning Research*, J. Dy and A. Krause, eds. (PMLR), pp. 2127–2136.
24. Saldanha, O.L., Loeffler, C.M.L., Niehues, J.M., van Treeck, M., Seraphin, T.P., Hewitt, K.J., Cifci, D., Veldhuizen, G.P., Ramesh, S., Pearson, A.T., and Kather, J.N. (2023). Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. *npj. Precis. Oncol.* 7, 35.
25. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at arXiv. [cs.CV].
26. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022.
27. He, K., Gan, C., Li, Z., Reik, I., Yin, Z., Ji, W., and Gao, Y. (2022). Transformers in medical image analysis: A review. Preprint at arXiv.
28. Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P., Han, T., van Treeck, M., Buelow, R.D., Langer, R., Dislich, B., Boor, P., Schulz, V., and Kather, J.N. (2022). Adversarial attacks and adversarial robustness in computational pathology. *Nat. Commun.* 13, 5711.
29. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., and Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* 81, 102559.
30. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., and Mahmood, F. (2022). Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16144–16155.
31. Ghaffari Laleh, N., Muti, H.S., Loeffler, C.M.L., Echle, A., Saldanha, O.L., Mahmood, F., Lu, M.Y., Trautwein, C., Langer, R., Dislich, B., et al. (2022). Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med. Image Anal.* 79, 102474.
32. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., and Zhang, Y. (2021). TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, pp. 2136–2147.
33. Reisenbuechler, D., Wagner, S.J., Boxberg, M., and Peng, T. (2022). Local Attention Graph-based Transformer for Multi-target Genetic Alteration Prediction. Preprint at arXiv. [cs.CV].
34. Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 8, 679–698.
35. Wagner, S.J., Khalili, N., Sharma, R., Boxberg, M., Marr, C., de Back, W., and Peng, T. (2021). Structure-Preserving Multi-domain Stain Color Augmentation Using Style-Transfer with Disentangled Representations. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (Springer International Publishing), pp. 257–266.
36. Kim, Y.J., Jang, H., Lee, K., Park, S., Min, S.-G., Hong, C., Park, J.H., Lee, K., Kim, J., Hong, W., et al. (2021). PAIP 2019: Liver cancer segmentation challenge. *Med. Image Anal.* 67, 101854.
37. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., and Ketchum, K.A. (2015). The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* 14, 2707–2713.
38. Vasaiakar, S., Huang, C., Wang, X., Petyuk, V.A., Savage, S.R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O.A., et al. (2019). Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* 177, 1035–1049.e19.

39. Hoffmeister, M., Jansen, L., Rudolph, A., Toth, C., Kloor, M., Roth, W., Bläker, H., Chang-Claude, J., and Brenner, H. (2015). Statin use and survival after colorectal cancer: the importance of comprehensive confounder adjustment. *J. Natl. Cancer Inst.* *107*, djv045.
40. Brenner, H., Chang-Claude, J., Seiler, C.M., and Hoffmeister, M. (2011). Long-term risk of colorectal cancer after negative colonoscopy. *J. Clin. Oncol.* *29*, 3761–3767.
41. Grabsch, H., Dattani, M., Barker, L., Maughan, N., Maude, K., Hansen, O., Gabbert, H.E., Quirke, P., and Mueller, W. (2006). Expression of DNA double-strand break repair proteins ATM and BRCA1 predicts survival in colorectal cancer. *Clin. Cancer Res.* *12*, 1494–1500.
42. Gray, R.T., Loughrey, M.B., Bankhead, P., Cardwell, C.R., McQuaid, S., O'Neill, R.F., Arthur, K., Bingham, V., McGready, C., Gavin, A.T., et al. (2017). Statin use, candidate mevalonate pathway biomarkers, and colon cancer survival in a population-based cohort study. *Br. J. Cancer* *116*, 1652–1659.
43. Gray, R.T., Cantwell, M.M., Coleman, H.G., Loughrey, M.B., Bankhead, P., McQuaid, S., O'Neill, R.F., Arthur, K., Bingham, V., McGready, C., et al. (2017). Evaluation of PTGS2 Expression, PIK3CA Mutation, Aspirin Use and Colon Cancer Survival in a Population-Based Cohort Study. *Clin. Transl. Gastroenterol.* *8*, e91.
44. Morton, D., Seymour, M., Magill, L., Handley, K., Glasbey, J., Glimelius, B., Palmer, A., Seligmann, J., Laurberg, S., Murakami, K., et al. (2023). Preoperative Chemotherapy for Operable Colon Cancer: Mature Results of an International Randomized Controlled Trial. *J. Clin. Oncol.* *41*, 1541–1552.
45. Hawkins, N. (2011). MCO Study Tumour Collection.
46. Jonnagaddala, J., Croucher, J.L., Jue, T.R., Meagher, N.S., Caruso, L., Ward, R., and Hawkins, N.J. (2016). Integration and Analysis of Heterogeneous Colorectal Cancer Data for Translational Research. *Stud. Health Technol. Inf.* *225*, 387–391.
47. Ward, R., and Hawkins, N. Molecular and Cellular Oncology (MCO) Study Data. UNSW Australia. doi
48. (2015). MCO Study Whole Slide Image Collection.
49. Shulman, K., Barnett-Griness, O., Friedman, V., Greenston, J.K., Gruber, S.B., Lejbkowitz, F., and Rennert, G. (2018). Outcomes of Chemotherapy for Microsatellite Instable-High Metastatic Colorectal Cancers. *JCO Precis. Oncol.* *2*, 1–10.
50. van den Brandt, P.A., Goldbohm, R.A., van 't Veer, P., Volovics, A., Hermus, R.J., and Sturmans, F. (1990). A large-scale prospective cohort study on diet and cancer in The Netherlands. *J. Clin. Epidemiol.* *43*, 285–295.
51. Offermans, K., Jenniskens, J.C., Simons, C.C., Samarska, I., Fazzi, G.E., Smits, K.M., Schouten, L.J., Weijenberg, M.P., Grabsch, H.I., and van den Brandt, P.A. (2022). Expression of proteins associated with the Warburg-effect and survival in colorectal cancer. *J. Pathol. Clin. Res.* *8*, 169–180.
52. Quirke, P., and Morris, E. (2007). Reporting colorectal cancer. *Histopathology* *50*, 103–112.
53. Quasar Collaborative Group, Gray, R., Barnwell, J., McConkey, C., Hills, R.K., Williams, N.S., and Kerr, D.J. (2007). Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. *Lancet* *370*, 2020–2029.
54. Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* *487*, 330–337.
55. Isella, C., Cantini, L., Bellomo, S.E., and Medico, E. (2014). TCGA CRC 450 Dataset.
56. Iveson, T.J., Kerr, R.S., Saunders, M.P., Cassidy, J., Hollander, N.H., Tabernero, J., Haydon, A., Glimelius, B., Harkin, A., Allan, K., et al. (2018). 3 versus 6 months of adjuvant oxaliplatin-fluoropyrimidine combination therapy for colorectal cancer (SCOT): an international, randomised, phase 3, non-inferiority trial. *Lancet Oncol.* *19*, 562–578.
57. West, N.P., Gallop, N., Kaye, D., Glover, A., Young, C., Hutchins, G.G.A., Brockmoeller, S.F., Westwood, A.C., Rossington, H., and Quirke, P.; Yorkshire Cancer Research Bowel Cancer Improvement Programme Group (2021). Lynch syndrome screening in colorectal cancer: results of a prospective two-year regional programme validating the NICE diagnostics guidance pathway across a 5.2 million population. *Histopathology* *79*, 690–699.
58. Taylor, J., Wright, P., Rossington, H., Mara, J., Glover, A., West, N., Morris, E., and Quirke, P.; YCR BCIP study group (2019). Regional multidisciplinary team intervention programme to improve colorectal cancer outcomes: study protocol for the Yorkshire Cancer Research Bowel Cancer Improvement Programme (YCR BCIP). *BMJ Open* *9*, e030618.
59. Loshchilov, I., and Hutter, F. (2017). Decoupled Weight Decay Regularization. Preprint at arXiv. [cs.LG].
60. Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. Preprint at arXiv. [cs.LG].
61. Smith, L.N., and Topin, N. (2019). Super-convergence: very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications (SPIE)*, pp. 369–386.
62. Abnar, S., and Zuidema, W. (2020). Quantifying Attention Flow in Transformers. Preprint at arXiv. [cs.LG].
63. Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L., Jimenez-Linan, M., Moore, L., and Gerstung, M. (2020). Pan-cancer Computational Histopathology Reveals Mutations, Tumor Composition and Prognosis.
64. 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., et al. (2020). Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Can. (Ott.)* *1*, 789–799.
65. Cercek, A., Lumish, M., Sinopoli, J., Weiss, J., Shia, J., Lamendola-Essel, M., El Dika, I.H., Segal, N., Shcherba, M., Sugarman, R., et al. (2022). PD-1 Blockade in Mismatch Repair–Deficient, Locally Advanced Rectal Cancer. *N. Engl. J. Med.* *386*, 2363–2376.
66. Hendricks, A., Amallraja, A., Meißner, T., Forster, P., Rosenstiel, P., Burmeister, G., Schafmayer, C., Franke, A., Hinz, S., Forster, M., and Williams, C.B. (2020). Stage IV Colorectal Cancer Patients with High Risk Mutation Profiles Survived 16 Months Longer with Individualized Therapies. *Cancers* *12*, 393.
67. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., and Madabhushi, A. (2019). Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* *16*, 703–715.
68. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* *24*, 1559–1567.
69. Kleppe, A., Skrede, O.-J., De Raedt, S., Hveem, T.S., Askautrud, H.A., Jacobsen, J.E., Church, D.N., Nesbakken, A., Shepherd, N.A., Novelli, M., et al. (2022). A clinical decision support system optimising adjuvant chemotherapy for colorectal cancers by integrating deep learning and pathological staging markers: a development and validation study. *Lancet Oncol.* *23*, 1221–1232.
70. Campanella, G., Hanna, M.G., Geneslaw, L., Miralor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* *25*, 1301–1309.
71. National Institute for Health and Care Excellence (NICE) (2020). Molecular Testing Strategies for Lynch Syndrome in People with Colorectal Cancer (NICE Guidance). <https://www.nice.org.uk/guidance/ng151>.
72. Liu, Y., Sethi, N.S., Hinoue, T., Schneider, B.G., Cherniack, A.D., Sanchez-Vega, F., Seoane, J.A., Farshidfar, F., Bowlby, R., Islam, M., et al. (2018). Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell* *33*, 721–735.e8.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
The Cancer Genome Archive (TCGA)	https://portal.gdc.cancer.gov/	RRID:SCR_003193
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	https://proteomic.datacommons.cancer.gov/pdc/	RRID:SCR_017135
Software and algorithms		
Framework for experiments and model implementation	This manuscript; https://github.com/peng-lab/HistoBistro	https://zenodo.org/badge/latestdoi/613444008
The model has also been implemented in this framework	https://github.com/KatherLab/marugoto	

RESOURCE AVAILABILITY

Lead contact

Further information and requests regarding this manuscript should be sent to and will be fulfilled by the lead investigator, Jakob Nikolas Kather (jakob_nikolas.kather@tu-dresden.de).

Materials availability

We release all multi-cohort model weights created in this study under an open-source license. More specifically, the model for MSI high, *BRAF*, and *KRAS* detection.

Data and code availability

Some of the data that support the findings of this study are publicly available, and some are proprietary datasets provided under collaboration agreements. All data (including histological images) from the TCGA database are available at <https://portal.gdc.cancer.gov/>. All data from the CPTAC cohort are available at <https://proteomic.datacommons.cancer.gov/>. All molecular data for patients in the TCGA and CPTAC cohorts are available at <https://cbioportal.org/>. Data access for the Northern Ireland Biobank can be requested at <http://www.nibiobank.org/for-researchers>. Data access for the MCO cohort can be requested at <https://researchdata.edu.au/mco-study-tumour-collection/1957427>. All other data are under controlled access according to the local ethical guidelines and can only be requested directly from the respective study groups that independently manage data access for their study cohorts.

All code was implemented in Python using the DL framework PyTorch. All source codes to reproduce the experiments of this paper are available under an open-source license at <https://github.com/peng-lab/HistoBistro> (<https://doi.org/10.5281/zenodo.8208791>). The model is also implemented in the DL pipeline <https://github.com/KatherLab/marugoto/tree/transformer>. We release all model weights under an open-source license.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ethics statement

In this study, we retrospectively analyzed anonymized patient samples from multiple academic institutions. At each of the following sites, the respective ethics board has given consent to this analysis: DACHS, Epi700, ERLANGEN, MAINZ, MECC, MUNICH, NLCS, QUASAR, FOxTROT, TRANSCOT, MCO. At the following sites, specific ethics approval was not required for a retrospective analysis of anonymized samples: CPTAC, DUSSEL, TCGA, GUANGZHOU, and YCR-BCIP. Our study adheres to STARD (Table S3).

Cohort description

Through coordination by the MSIDECT consortium (www.msitect.eu), we have collected over 20,000 H&E tissue sections of 13,689 patients with CRC from 16 patient cohorts in total, including two public databases (Figures 1D–1F). The cohorts obtained are as follows:

1. The public database “The Clinical Proteomic Tumor Analysis Consortium”, CPTAC (publicly available at <https://pdc.cancer.gov/pdc/>, USA)^{37,38} which includes tumors of any stage;

2. DACHS (Darmkrebs: Chancen der Verhütung durch Screening, Southwest Germany),^{39,40} a large population-based case-control and patient cohort study on CRC, including samples of patients with stages I-IV from different laboratories in south-western Germany coordinated by the German Cancer Research Center (Heidelberg, Germany);
3. The DUSSEL (DUSSEldorf, Germany) cohort, a case series of CRC tumors resected with curative intent and collected at the Marien-Hospital in Duesseldorf, Germany, between January 1990 and December 1995⁴¹;
4. Epi700 (Belfast, N. Ireland, UK),^{42,43} a population-based cohort of stage II and III colon cancers treated by surgical resection between 2003 and 2008;
5. The ERLANGEN cohort, a CRC cohort collected at the Uniklinikum Erlangen in Germany between 2002 and 2010.
6. The “Fluoropyrimidine, Oxaliplatin, and Targeted Receptor pre-Operative Therapy for colon cancer cohort” (FOxTROT)⁴⁴ including pre-therapeutic biopsy and post-therapeutic resection tumors from UK sites;
7. The GUANGZHOU cohort, a small CRC case series of MSI-high cases collected in The Second Affiliated Hospital of Guangzhou Medical University, China;
8. The MAINZ cohort, a small CRC case series of biopsies collected in the University Medical Center Mainz in Germany.
9. The Molecular and Cellular Oncology Study (MCO) cohort^{45–48} from the University of New South Wales, Australia;
10. MECC (Molecular Epidemiology of Colorectal Cancer, Israel),⁴⁹ a population-based case-control study in northern Israel;
11. The MUNICH (Munich, Germany) CRC series, a case series collected at the Technical University of Munich in Germany.
12. The NLCS (Netherlands Cohort Study, The Netherlands)^{50,51} cohort, which contains tissue samples obtained from patients with any tumor stage as part of the Rainbow-TMA consortium study;
13. QUASAR, the “Quick and Simple and Reliable” trial investigating survival benefit of adjuvant chemotherapy in patients from the United Kingdom with mostly stage II tumors^{52,53};
14. The public repository “The Cancer Genome Atlas”, TCGA (publicly available at <https://portal.gdc.cancer.gov/>, USA)^{54,55} which includes tumors of any stage;
15. The TransSCOT cohort, the translational arm of the SCOT trial, an “international, randomised, phase 3, non-inferiority trial” involving adult patients with high-risk stage II or stage III CRC⁵⁶;
16. The YCR-BCIP (Yorkshire Cancer Research Bowel Cancer Improvement Program, Yorkshire, United Kingdom [UK]), a population-based register of bowel cancer patients in Yorkshire, UK,^{57,58} for which surgical resections and biopsies were available as separate cohorts.

Detailed clinicopathological variables are shown in Table S4. In all cohorts, formalin-fixed paraffin-embedded (FFPE) tissue was used. Slides have been scanned at their respective centers. For each patient, either an MSI status or an MMR status, obtained by PCR or IHC, respectively, is available. Although MSI status and MMR status are not fully concordant,⁵⁷ they are used interchangeably in clinical routine and grouped as a single category in this study. *KRAS* and *BRAF* mutational status are available for the cohorts DACHS, Epi700, NLCS, QUASAR, and TCGA.

METHOD DETAILS

Model description

Our biomarker prediction pipeline consists of three steps (Figure 1): i) the data pre-processing pipeline (Figure 1A), ii) the transformer-based feature extractor, and iii) the transformer-based aggregation module that yields the final prediction from the embeddings of all patches of a whole-slide image (WSI) (Figure 1B).

In the pre-processing pipeline, tissue regions are segmented using RGB thresholding and Canny edge detection³⁴ to detect background and blurry regions. We include all tiles from a WSI, i.e., both tumor and healthy tissue tiles, thus reducing the burden of manual annotations when applying the algorithm. Subsequently, the WSI is tessellated into tiles of size 512 × 512 pixels at 20× magnification with a resolution of 0.5 microns per pixel. To reduce the impact of the staining color on the model generalization, the tiles are stain-color augmented using a structure-preserving GAN trained on TCGA.³⁵

We extract feature representations of dimension 768 for every tile using the CTransPath model.²⁹ (Figure 1B). The model architecture is based on a Swin Transformer²⁶ that combines the hierarchical structure of CNNs with the global self-attention modules of transformers by computing self-attention in a sliding-window fashion. Similar to CNNs, these are stacked to increase the receptive field in every stage. CTransPath consists of three convolutional layers at the beginning to facilitate local feature extraction and improve training stability,²⁹ followed by four Swin Transformer stages. Wang et al. trained the network using an unsupervised contrastive loss on data from TCGA and PAIP³⁶ from multiple organs and provided the weights for public use. The embeddings for each tile are stored for the subsequent training procedure.

The final part of the model takes all patches of a WSI as input and predicts one biomarker for all input patches in a weakly supervised manner (Figure 1B). Common attention-based MIL approaches²³ use a small neural network, which mostly consists of two layers, to compute patch importance based on the embeddings. Each weight is computed based on one patch and finally, all weights are normalized over the input elements. In contrast to this, in our model, the patch embeddings are passed into a transformer network using multi-headed self-attention that considers the patch embeddings as a sequence and relates each element to every other element. In particular, assuming that $x \in \mathbb{R}^{n \times d}$ is the input sequence representing a WSI with n patch embeddings of dimension d , the self-attention layer computes a query-key product in the following way

$$SA(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where queries $Q \in \mathbb{R}^{n \times d_k}$, keys $K \in \mathbb{R}^{n \times d_k}$, and values $V \in \mathbb{R}^{n \times d_v}$. These are computed from the input sequence x by

$$Q = W_Q \cdot x, K = W_K \cdot x \text{ and } V = W_V \cdot x,$$

where $W_Q \in \mathbb{R}^{d \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$, and $W_V \in \mathbb{R}^{d \times d_v}$ are learnable parameters. Multi-headed self-attention applies self-attention in every head and concatenates the heads in a weighted manner:

$$MSA(x) = \text{concat}(\text{head}_1, \dots, \text{head}_h) \cdot W_O$$

where $\text{head}_i = SA(Q^{(i)}, K^{(i)}, V^{(i)})$ for $i \in \{1, \dots, h\}$

and $W_O \in \mathbb{R}^{hd_v \times d}$ is learnable. We choose a small transformer network architecture consisting of two layers with each eight heads ($h = 8$), a latent dimension of 512, and the same dimension for query, keys, and values. Therefore, the latent dimension of each head is $d_v = d_k = 64$, such that $hd_v = 8 \cdot 64 = 512$.

Assuming that n is the number of patches per WSI, the embeddings of each patch $i \in \{1, \dots, n\}$ are stacked to a sequence of dimension $n \times 768$ and are passed through a linear projection layer followed by the non-linear activation ReLU to reduce the dimension from 768 to 512. Subsequently, a class token is concatenated to the input, similar to the usage in vision transformers,²⁵ yielding an input of dimension $(n + 1) \times 512$ that is passed to the transformer layer. In each transformer layer, a block of layer normalization and multi-headed self-attention is followed by a block of layer normalization and a multi-layer perceptron (MLP), with skip connections applied across each block (Figure 1C).

After the two transformer layers, the class token of size 1×512 is passed into an MLP head. Depending on the number of class tokens used, this enables single-target or multi-target binary prediction. Instead of attaching a class token, all n sequence elements could be averaged to a single sequence element of size 1×512 and passed into the MLP head. The averaging approach achieves similar performance to the class token version (Table S2), but we decided to use the class token for better interpretability of the attention heads. We also compared our model architecture to the existing transformer-based aggregation module TransMIL³² (Table S2).

Experimental setup and implementation details

We performed all experiments using 5-fold cross-validation with in-domain validation and testing. In this cross-validation variant, in-domain validation and test set are split off the full dataset on patient level, leaving 3-folds for training. By also cycling the in-domain test set through the complete dataset, we evaluated our model on more representative test sets than when fixing one smaller set for the dataset. During training, the validation set was used to determine the best model, which was finally evaluated on the test set. We further evaluated our models on external cohorts outside the dataset for out-of-domain testing.

The transformer models were trained with the AdamW⁵⁹ optimizer using weight decay and learning rate both of 2×10^{-5} . All models were trained for eight epochs with batch size of one for two reasons: first, the sequences of embeddings had different lengths due to the variable number of tiles per WSI and could thus not be stacked to mini-batches of equal length inputs; second, limits in GPU capacity (32GB) because of the quadratic complexity of the self-attention mechanism and the large number of tiles per slide (up to 12,000, Figure S1). To account for the varying cohort size, we evaluated the models every 500 iterations for runs on single cohorts, and every 1000 iterations for runs on multiple cohorts.

For comparison, we implemented the attention-based MIL approach from Ilse et al.,²³ referred to as AttentionMIL. It provided the best results with Adam⁶⁰ optimizer, 1×10^{-2} as weight decay value, along with the fit-one-cycle learning rate scheduling policy⁶¹ with a maximum learning rate of 1×10^{-4} , trained over 32 epochs, and the first 25% of the cycle with increasing learning rate.

Visualization and explainability

The final prediction is retrieved via the class token that is attached to the input sequence. To visualize the contribution of each input patch, we employed attention rollout as introduced by Abnar and Zuidema.⁶² To obtain the attention at the class token in the final layer, the attention maps of the preceding layers are multiplied recursively. Attention rollout thus quantifies to which extent each patch contributes to the final prediction in the class score. Additionally, we visualized the attention scores for each head in the transformer by taking the class token's self-attention, i.e., the query and key product. All presented attention scores were normalized to the range $[0, 1]$ and clamped to the lower and higher 5%-quantiles, respectively, for better visual interpretability.

To visualize whether a patch contributed toward a positive or negative classification outcome, we fed the patches one-by-one through the transformer and visualized the resulting classification scores of the model. These scores were naturally in the range $[0, 1]$ and can thus be directly visualized without further normalizing or clamping of values.

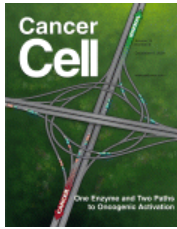
QUANTIFICATION AND STATISTICAL ANALYSIS

We used the area under the receiver operator curve (AUROC) as our main evaluation metric. Since our data are naturally highly imbalanced with respect to the target variables MSI, BRAF, and KRAS (Figures 1D and 1E), we further used the area under the

precision-recall curve (AUPRC) as a metric as this metric accounts better for class imbalances than the AUROC metric. The precision-recall curve relates the recall or specificity, i.e., the ratio of correctly positive predicted samples to all positive samples, to the precision, i.e., the ratio of correctly positive predicted samples to all positive predicted samples. For every experiment, we reported the mean and the standard deviation of respective 5-fold cross-validation's model in-domain and external test performances. We split the dataset into patient-wise training, validation, and internal test sets stratified by the target label, thus ensuring that every patient can only occur in one of these sets. The external test sets always consisted of different cohorts to better quantify the generalization properties of our algorithms.



RightsLink



Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study

Author:

Sophia J. Wagner, Daniel Reisenbüchler, Nicholas P. West, Jan Moritz Niehues, Jiefu Zhu, Sebastian Foersch, Gregory Patrick Veldhuizen, Philip Quirke, Heike I. Grabsch, Piet A. van den Brandt, Gordon G.A. Hutchins, Susan D. Richman, Tanwei Yuan, Rupert Langer et al.

Publication: Cancer Cell**Publisher:** Elsevier**Date:** 11 September 2023

© 2023 The Author(s). Published by Elsevier Inc.

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC-BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact [Elsevier](#) Global Rights Department.

Are you the [author](#) of this Elsevier journal article?

© 2024 Copyright - All Rights Reserved | [Copyright Clearance Center, Inc.](#) | [Privacy statement](#) | [Data Security and Privacy](#)
| [For California Residents](#) | [Terms and Conditions](#) Comments? We would like to hear from you. E-mail us at customercare@copyright.com

A.8. Built to last? Reproducibility and reusability of deep learning algorithms in computational pathology

This publication was published in the peer-reviewed international journal *Modern Pathology* in 2024. The one-page summary can be found in Section 3.8.

Wagner, S. J.*, Matek, C.*, Boushehri, S. S., Boxberg, M., Lamm, L., Sadafi, A., ... & Peng, T. (2024). Built to last? Reproducibility and reusability of deep learning algorithms in computational pathology. *Modern Pathology*, 37(1), 100350. *Contributed equally

Copyright: © 2023 United States & Canadian Academy of Pathology. Published by Elsevier Inc. The copyright clearance is attached after the original publication.

Review Article

Built to Last? Reproducibility and Reusability of Deep Learning Algorithms in Computational Pathology

Sophia J. Wagner^{a,b}, Christian Matek^{c,d}, Sayedali Shetab Boushehri^{b,c,e}, Melanie Boxberg^{f,g}, Lorenz Lamm^{a,h}, Ario Sadafi^{b,c}, Dominik J.E. Winter^{c,i}, Carsten Marr^{c,*}, Tingying Peng^{a,*}

^a Helmholtz AI, Helmholtz Munich—German Research Center for Environmental Health, Neuherberg, Germany; ^b School of Computation, Information and Technology, Technical University of Munich, Garching, Germany; ^c Institute of AI for Health, Helmholtz Munich—German Research Center for Environmental Health, Neuherberg, Germany; ^d Institute of Pathology, University Hospital Erlangen, Erlangen, Germany; ^e Data & Analytics (D&A), Roche Pharma Research and Early Development (pRED), Roche Innovation Center Munich, Germany; ^f Institute of Pathology, Technical University Munich, Munich, Germany; ^g Institute of Pathology Munich-North, Munich, Germany; ^h Helmholtz Pioneer Campus, Helmholtz Munich—German Research Center for Environmental Health, Neuherberg, Germany; ⁱ School of Life Sciences, Technical University of Munich, Weihenstephan, Germany

ARTICLE INFO

Article history:

Received 10 February 2023

Revised 2 October 2023

Accepted 3 October 2023

Available online 10 October 2023

Keywords:

artificial intelligence
computational pathology
deep learning
histology/histopathology
reproducibility
reusability

ABSTRACT

Recent progress in computational pathology has been driven by deep learning. While code and data availability are essential to reproduce findings from preceding publications, ensuring a deep learning model's reusability is more challenging. For that, the codebase should be well-documented and easy to integrate into existing workflows and models should be robust toward noise and generalizable toward data from different sources. Strikingly, only a few computational pathology algorithms have been reused by other researchers so far, let alone employed in a clinical setting. To assess the current state of reproducibility and reusability of computational pathology algorithms, we evaluated peer-reviewed articles available in PubMed, published between January 2019 and March 2021, in 5 use cases: stain normalization; tissue type segmentation; evaluation of cell-level features; genetic alteration prediction; and inference of grading, staging, and prognostic information. We compiled criteria for data and code availability and statistical result analysis and assessed them in 160 publications. We found that only one-quarter (41 of 160 publications) made code publicly available. Among these 41 studies, three-quarters (30 of 41) analyzed their results statistically, half of them (20 of 41) released their trained model weights, and approximately a third (16 of 41) used an independent cohort for evaluation. Our review is intended for both pathologists interested in deep learning and researchers applying algorithms to computational pathology challenges. We provide a detailed overview of publications with published code in the field, list reusable data handling tools, and provide criteria for reproducibility and reusability.

© 2023 United States & Canadian Academy of Pathology. Published by Elsevier Inc. All rights reserved.

Introduction

Technical progress has been driving digitization in pathology over the past decade. Coupled with advances in deep learning (DL) methods, computational approaches help to localize, segment, and classify single cells and tissue types in an automated manner

These authors contributed equally: Sophia J. Wagner and Christian Matek.

* Corresponding authors.

E-mail addresses: tingying.peng@helmholtz-munich.de (T. Peng), carsten.marr@helmholtz-munich.de (C. Marr).



ELSEVIER

0893-3952/\$ - see front matter © 2023 United States & Canadian Academy of Pathology. Published by Elsevier Inc. All rights reserved.
<https://doi.org/10.1016/j.modpat.2023.100350>

Table 1

Glossary for commonly used terms in computational pathology

Digital pathology	Histologic slides are scanned and digitized so pathologists can examine the patient material on a screen instead of working on optical microscopes. Digitized slides can be stored and processed, enabling the use of computational methods in the diagnostic process.
Computational pathology	The analysis of digitized histologic slides with computational methods. ¹
Specimen	A tissue sample, for example, obtained during a biopsy or other surgical procedures, typically fixed in formalin and embedded in paraffin.
Section	A thin slice (with a typical thickness of 3–15 μm) of a specimen mounted on a microscopic slide. WSI: The digitized image of a tissue section on a microscope slide. Slides can be scanned in very high magnification resulting in images of sizes up to several gigapixels.
Patches and tiles	WSIs are split up into smaller images (eg, 512×512 pixels), also called patches or tiles, that can be processed by neural networks. Unlike WSIs, these smaller image data units allow for easier and parallelized image processing. The terms “tiles” and “patches” are often inconsistently used. Overall, “patches” is the more common term, but if the smaller images are used for subsequent WSI rendering by stitching them back together, they are sometimes called tiles.
Annotations	Diagnostic information on pixel or patch levels is obtained from manual expert pathologist labeling at different levels of resolution (eg, tumor regions are outlined at $2\times$ to $5\times$ while cell-level annotations are performed at $20\times$ to $40\times$). WSI-level annotations can be all diagnostic information about the patient (eg, age, survival, staging, and grading) mainly obtained without additional expert pathologist interaction.
Supervised learning	Training procedure of a neural network, where the ground truth, that is, the correct label for the task, is available for each data point. However, in medical imaging and especially computational pathology, full expert annotations at pixel or patch level are time-consuming and, hence, rare. Pixel-level annotations are used to localize tissues in segmentation tasks, where each pixel is assigned a tissue label. Patch-level annotations are used for classification tasks, where one label is predicted for the entire input patch.
Weakly supervised learning	Due to the rareness of fully annotated WSIs, weakly supervised learning approaches, like MIL, are often used to train neural networks. With MIL, only WSI-level annotations, such as diagnostic information on the cancer type or survival, are required for classification.
CNN	A neural network that can be trained to extract features by sliding trainable filters across the image. This makes CNNs translationally invariant and, therefore, well suited for histologic data because relevant features can be found anywhere on a tile.
U-Net	A CNN with an encoder-decoder architecture for segmenting biomedical images. ² It was adapted in many ways and ranks among the most common architectures for segmentation tasks.
Mask R-CNN	A CNN architecture for instance segmentation in object detection. ³ In contrast to U-Net, which does not distinguish between instances of a class, Mask region-CNN (Mask R-CNN) outputs a segmentation mask for each instance on the image, which makes it useful for tasks such as nuclei segmentation and cell counting.

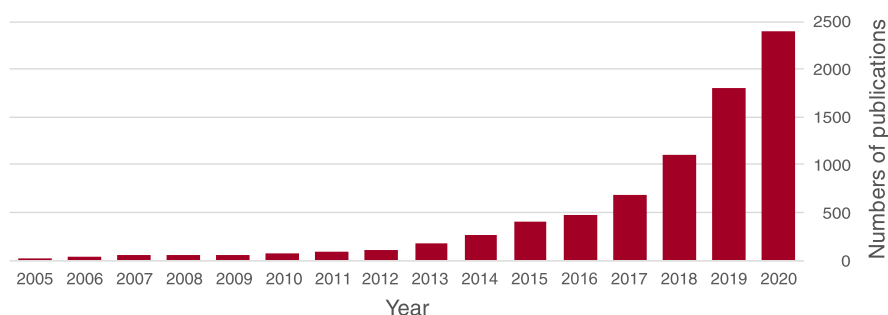
CNN, convolutional neural network; MIL, multiple instance learning; WSI, whole slide image.

and form the research field of computational pathology¹ (see Table 1 for a glossary^{1–3}). In particular, deep neural networks have recently been shown to reach the performance level of medical experts on well-defined tasks such as skin cancer diagnosis,⁴ lung cancer subtype classification,⁵ or the recognition of malignant white blood cells.⁶

However, despite the steady increase in publications in this field (Fig. 1) and their promising results, only a few have reached clinical implementation.^{7,8} This is due to several reasons: for DL-based methods, code availability is a natural requirement for reproducibility and, in particular, reusability, which, unfortunately, is not yet current practice for most publications. Even when code is available, reproducing the original results can be challenging and requires the assistance of the original author.⁹ In particular, ready-to-use scripts with sufficient instructions or intuitive demo examples are rarely published. This makes the reuse of current methods difficult for non-DL experts, especially for pathologists who are not supported by computational experts.

Another reason, particularly relevant to clinical implementation, is the generalization gap of algorithms in computational pathology. Often, the published performance of DL algorithms cannot be transferred to other data sets due to differences in staining or scanner settings. Therefore, external validation of algorithms and statistical robustness analysis are essential to assess generalizability. Finally, algorithms marketed for use in a clinical setting must additionally be approved by national or international authorities such as the US Food and Drugs Administration or the European Medicines Agency, an often lengthy and complicated process involving business markets and legal issues, which is beyond the scope of this review.

Here, we focus on DL algorithms for computational pathology and their reproducibility and reusability. For the ultimate goal of *reusing* DL algorithms, the algorithms must be *reproducible* and generalizable to similar data sets (ie, *robust*) and external data sets (ie, *replicable*) (see Table 2^{9–13} and Fig. 2 for definitions of reproducibility, reusability, and related terms). Because hematoxylin

**Figure 1.**

The number of articles published in the field of computational pathology in PubMed (retrieved on July 22, 2021) has markedly increased in the past 15 years.

Table 2

Definitions of reproducibility, reusability, and related terms

Reproducibility	Using identical materials and procedures, the results of a study can be duplicated, and ultimately, identical conclusions can be drawn. ¹⁰ In the context of algorithms, the same result can be obtained from the same data, code, and analysis methods. ^{9,11}
Robustness	The same results are obtained from an algorithm despite small perturbations in the input. ^{12,13}
Replicability	Conclusions are stable based on independently acquired data, ^{9,11} that is, code and analysis methods can be employed to external data with similar results and performance. For deep learning algorithms, replicability is equivalent to model generalizability, a key requirement for the clinical application of new algorithms.
Reusability	A piece of software is considered reusable if it can be included in an existing computational pathology setup with minor efforts (eg, without the need for extensive rearrangements of the workflow or software development).

and eosin (H&E) staining is the most commonly used routine staining for cancer diagnosis,¹⁴ we restricted this review to methods for H&E-stained whole slide image (WSI) analysis. We considered 5 computational pathology use cases and assembled a systematic overview of publications published between January 2019 and March 2021. For this, we compiled criteria for reproducibility in a practical context and examined each work with respect to these. We additionally provide an overview of current data handling tools.

Use Cases

We collected 160 studies between January 2019 and March 2021 on the automated analysis of histologic slides for cancer diagnosis and treatment (Supplementary Tables S1–S5; flowchart for the selection is presented in Supplementary Fig. S1). We split this body of literature into the following use cases: (1) stain normalization; (2) tissue type segmentation; (3) evaluation of cell-level features; (4) genetic alteration prediction; and (5) inference of grading, staging, and prognostic information (Fig. 3).

In this technical chapter, we first briefly introduce every use case, followed by an overview of the latest DL methods, focusing on works that provide code along with the publication. At the end of each section, we wrap up with an analysis of the reproducibility in the specific context.

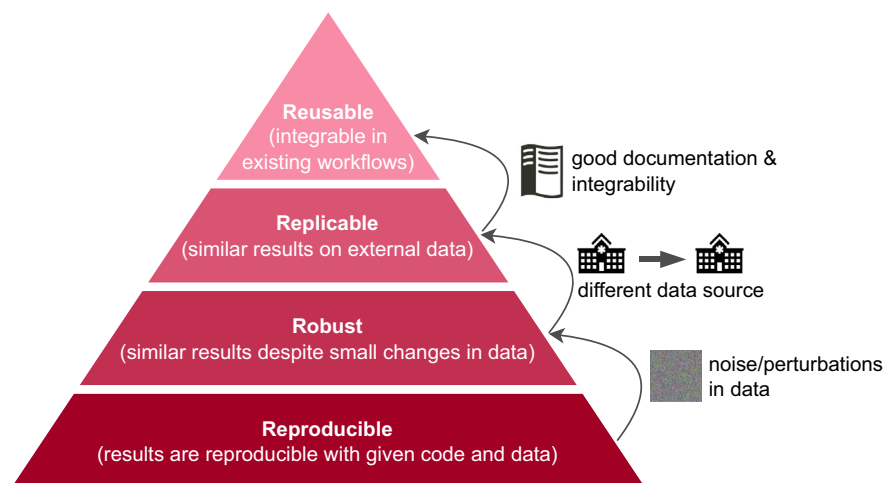
As a prerequisite, open and publicly available data handling tools for reading, annotating, and sharing histopathological data are essential. We compiled the most common tools in Table 3^{15–20} and provided a more detailed overview in Supplementary Table S6

on software features, requirements, and the possibility of extending the tool by its own code.

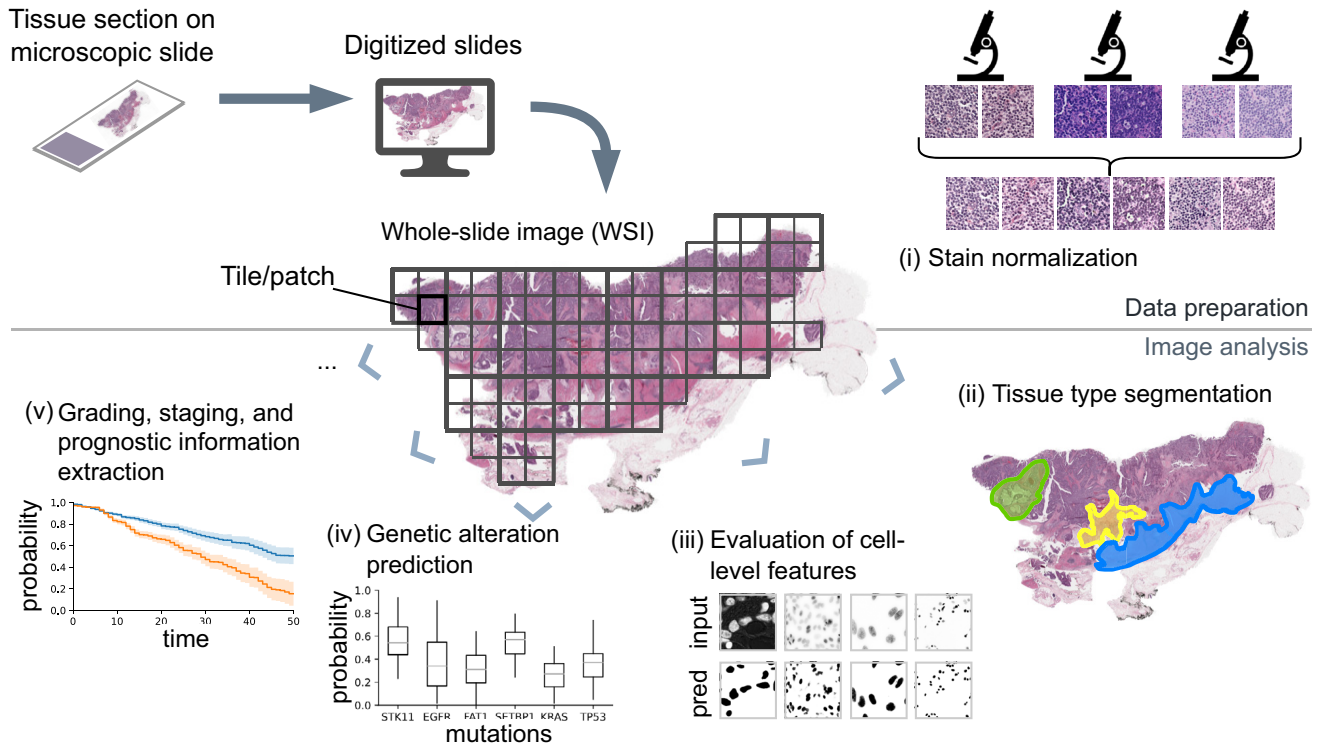
Stain Normalization

Most work in computational histopathology focuses on H&E-stained routine sections, with hematoxylin staining nuclei in blue-purple and eosin staining extracellular material in pink.²¹ Digitized sections are prone to many image variations caused by differences in the tissue preparation technique (eg, the thickness and flatness of the sample cut), staining protocols, handling, and storage conditions. Moreover, slide scanners differ in microscope illumination, image postprocessing, or noise handling. These factors lead to significant variability in the visual appearance of WSIs, which affects subsequent analysis and may lead to poor generalizability of algorithms. Computational methods aim at reducing the effects of these variations,²² for example, by normalizing the stain color from a predefined source domain to 1 or more target domains. These methods include improved analytical approaches, such as color deconvolution and, more recently, DL-based methods.

Color Space Methods. Color deconvolution separates the hematoxylin from the eosin component in optical density space based on a reference tile.²³ This approach has been developed further recently: adaptive color deconvolution²⁴ incorporated the underlying stain distribution of the target WSI instead of only a single tile. Alternatively, nonnegative matrix factorization has been used to obtain a color deconvolution matrix²⁵ and was optimized for graphics processing unit (GPU) usage.²⁶ (This approach is not DL-based. We decided to keep it in the article because non-DL-based color space methods are heavily used in DL applications of computational pathology.)

**Figure 2.**

Reproducibility, robustness, replicability, and reusability in the context of deep learning algorithms for computational pathology.

**Figure 3.**

Overview of the use of deep learning in computational pathology, including data handling tools for reading, annotating, and sharing whole slide images (WSIs) (Table 3), and 5 applications of deep learning methods, which are covered in the Use Cases.

Generative Models. The increasing popularity of generative adversarial networks (GANs) has led to the development of style transfer methods for stain normalization,²⁷ which can be trained on all target WSI tiles instead of expert-picked reference tiles.²⁸ StainGAN was trained with a cycle consistency loss between the source and target domains, and the generator of the target domain was used to normalize all images in that domain.²⁹ However, GANs may not always preserve the tissue structure.³⁰ To overcome this, StainNet trained a convolutional neural network (CNN) consisting of 1×1 convolutions, transforming the source image from its original color space to the target color space without losing structural information.³¹ Alternatively, additional loss functions that compare images before and after normalization can be used to preserve the histopathological information, including texture, structure, and color features.²⁸

Stain Augmentation. In contrast to normalization methods, GANs can also simulate stain variability by generating synthetic images.

This renders neural networks on downstream tasks more robust and avoids loss of relevant information due to the limitations of normalization methods. Yamashita et al.³² propose data augmentation based on style transfer from artistic paintings, and Wagner et al.³³ used a GAN architecture for multiple domains to synthesize realistic histologic images while preserving the tissue structure.

Stain-Aware Models. Unlike the above stain normalization methods that project the external test data to the original training domain as a preprocessing step, stain information can be incorporated directly into the model, for example, for nuclei segmentation by creating a hematoxylin-aware CNN.³⁴

Tissue Type Segmentation

Accurate segmentation of a WSI into tissue types (eg, epithelial vs stromal vs lymphatic tissue) allows for quantitative follow-up analysis. Depending on the kind of available annotations, we discriminate between the following categories: methods for pixel-

Table 3

Overview of commonly used data handling tools in computational pathology

An essential prerequisite for implementing, transferring, and reusing computational pathology algorithms between researchers and different laboratories or institutions is a software structure for exchanging image data, annotations, and meta-information. With the progress of computational pathology, numerous data handling tools have been developed. In many tools, image data handling is based on system-level libraries, such as OpenSlide¹⁵ or Open Microscopy Environment Remote Objects (OMERO).¹⁶ They enable data to be interoperable between different vendor-specific image formats. Most of these tools also provide user interfaces for pathologists to analyze and annotate images. Annotations include class labeling or point flags, geometric shapes, and image-level labeling. While many popular image data handling and annotation tools were developed as standalone packages (eg, SlideRunner,¹⁷ QuPath,¹⁸ and Automated Slide Analysis Platform [ASAP]), an increasing number of recently developed packages, such as Cytomine¹⁹ or EXpert Algorithm Collaboration Tool (EXACT),²⁰ allow for web-based, collaborative data handling, which is essential for distributing, exchanging, and annotating data as well as evaluating models in a multi-institutional setting.

In addition to image annotation and exchange, data handling packages allow integration with independently developed analysis algorithms at different levels. Some tools offer integrated scripting for the automation of tasks, for example, using Groovy in QuPath. Additionally, programming interfaces to popular machine learning languages, such as Python, have been developed, for example, for OMERO. Several tools, such as EXACT and CaMicroscope, offer integrated, server-side evaluation of deep learning models. A detailed overview of open and publicly available data handling tools and their respective functionalities is provided in Supplementary Table S1.

wise or patch-wise segmentation, hierarchical architectures that imitate a pathologist's workflow, and methods that use WSI-level annotations and are therefore weakly supervised.

Pixel-Wise Segmentation. Vellal et al³⁵ assessed the risk of breast cancer from image features, such as the percentage of fibrous stroma, epithelium, and fatty tissue. Graham et al³⁶ developed a rotation-invariant CNN to account for the inherent rotational symmetry of histology images and validated their application on pixel-wise gland segmentation. Jayapandian et al³⁷ used pixel-wise segmentation for identifying 6 tissue types in kidney biopsies, applying the same U-Net architecture for segmenting patches with different magnifications.

Patch-Wise Segmentation. Image patches can be classified separately and subsequently stitched together to create a coarse segmentation map of the entire WSI. Zhao et al³⁸ computed a WSI's tumor-to-stroma ratio from such segmented patches, a prognostic factor for colorectal cancer. Rączkowski et al³⁹ proposed an active learning framework to train a CNN for patch classification in colorectal cancer. The network's uncertainty, estimated via Monte-Carlo dropout sampling,⁴⁰ was used to detect outlier tiles in the training set and to select them afterward for reconsideration. Wang et al⁴¹ generated spatial tissue maps by classifying single cells into tumor, stroma, and lymphocytes.

Hierarchical Segmentation. Hierarchical segmentation approaches mimic the workflow of pathologists by aggregating information from multiple scales of magnification. Schmitz et al⁴² created a family of U-Net–based encoder-decoder architectures that process high- and low-resolution image tiles in separate branches, fusing them with learnable gates from 3 publicly available data sets with liver, breast, and lymph node tissues. Alternatively, HookNet⁴³ fused the hidden space of multiple U-Net–based models that operate on different scales to deal with high-resolution and contextual information in breast and lung cancers.

Weakly Supervised Methods. Weakly supervised methods typically require slide-level annotations only. One recent approach was training CNNs directly on the entire WSI of lung cancer sections.⁴⁴ Subsequently, class activation maps⁴⁵ highlight relevant cancerous regions that were also identified by pathologists, which can be interpreted as a confidence measure of the algorithm. Silva-Rodríguez et al⁴⁶ trained a feature extraction network on the entire downsampled WSI to classify global Gleason grades on prostate cancer. During inference, semantic segmentations are upsampled from the feature maps and achieve similar performance as fully supervised approaches while only using the global labels. Alternatively, the WSI can be split into patches and patch-wise features for WSI classification.⁴⁷ Using this approach, attention scores produce interpretable heatmaps to visualize which regions contribute to the network's prediction.

Evaluation of Cell-Level Features

The evaluation of cell-level properties is a standard task in histopathology. For example, cell density and the abundance of dividing cells in tumor tissue are critical features for tumor grading. Here, we focus on 2 of the most widely studied cell-level tasks: segmentation of nuclei and detection of mitoses.

Nuclei Segmentation. Segmentation challenges have been introduced to benchmark efforts in this field, such as the Multi-Organ Nucleus Segmentation Challenge,⁴⁸ which provided 30 images and approximately 22,000 nuclear boundary

annotations in a public data set. Of the top 6 participants, 3 used U-Net–based semantic segmentation,² 2 used Mask region (R)-CNN–based instance segmentation,³ and one group used stacked U-Net and R-CNN models. The 2 dominating algorithms have been further tailored toward nuclear segmentation: Cui et al⁴⁹ predicted a boundary map additionally to object segmentations to separate touching nuclei efficiently. Jin et al⁵⁰ incorporated a U-Net into a pipeline to detect lymph node metastasis in patients with breast cancer. Mask R-CNN has been combined with a deep convolutional Gaussian mixture color normalization model, which clusters pixels according to nucleus morphology.⁵¹ Recently, other approaches, such as GANs, have been proposed, where the network is trained on unpaired data to map segmentation masks to nuclei images.⁵² To ease the annotation process for nuclei segmentation, Qu et al⁵³ provided a DL framework trained on incomplete annotations, which are much easier to generate.

Mitosis Detection. Identifying cells in the cell cycle's mitotic phases is a diagnostically relevant task, for example, for breast cancer grading and prognosis.⁵⁴ Several challenges released public data sets and benchmarked competing approaches for mitosis detection, for example, ICPR MITOS-2012,⁵⁵ ICPR MITOS-ATYPIA-2014,⁵⁶ or Tumor Proliferation Assessment Challenge (TUPAC) 16.⁵⁴ Most recently developed mitosis detection methods can be grouped into classification, segmentation, and detection. Pati et al⁵⁷ combined a classification task with metric learning to reduce the necessary amount of labeled data for more efficient network training for patch-level classification. Another approach for mitosis detection is pixel-wise semantic segmentation. Jiménez and Racocanu⁵⁸ showed that a U-Net–based semantic segmentation approach led to higher accuracy than that with previous classification approaches. Lafarge et al⁵⁹ proposed a special Euclidean motion group convolution to achieve translation and rotation invariance, which was integrated into a U-Net architecture and improved the model's robustness. Many other recent publications on mitosis detection were based on object detection,^{60–63} where only a weak centroid annotation that marks the center of the mitotic figure is required, compared to pixel-wise annotations for segmentation approaches. An alternative approach applied a cascade network, combining a first-stage object detection to identify mitosis candidates and a second-stage classification network for refinement.⁶⁴

Genetic Alteration Prediction

As genetic alterations can carry crucial predictive and prognostic information for cancer diagnosis, they have become increasingly relevant to the diagnostic workup and the selection of therapeutic pathways.⁶⁵ Therefore, patients are profiled for genetic alterations or other biomarkers that characterize their disease, for example, colorectal cancer,⁶⁶ to obtain better-targeted therapies. However, using molecular assays to determine the mutational spectrum of malignant cells is expensive and time-consuming. Furthermore, DNA or RNA extracted from small samples may not suffice quantitatively for a comprehensive analysis. RNA in older samples may already be degraded and, hence, not qualified for analysis. Techniques such as whole-genome sequencing require fresh tissue and are thus not applicable to formalin-fixed paraffin-embedded tissue usually used for histologic sample preparation. Therefore, algorithm-based prognostic stratification and mutation prediction from H&E-stained WSIs offer an attractive, cost- and time-effective, as well as tissue-sparing addition to existing molecular characterization methods.

Image-Based Mutation Prediction. Although some genetic alterations, such as mutations, copy number variations, and translocations, can be relevant for disease characterization, most work has so far focused on mutation prediction from imaging data. Coudray et al⁵ found that 6 commonly mutated genes in lung adenocarcinoma (*STK11*, *EGFR*, *FAT1*, *SETBP1*, *KRAS*, and *TP53*) can be predicted from WSI images. Since then, image-based mutation prediction has been applied to various types of cancers, such as melanoma,^{67,68} breast cancer,⁶⁹⁻⁷¹ lung cancer,^{72,73} colorectal cancer,⁷⁴⁻⁷⁶ bladder cancer,⁷⁷ and thyroid carcinoma.⁷⁸ Several recent studies attempt a pan-cancer approach that predicts genetic alteration across multiple tissue types from WSIs directly.^{79,80}

Modeling Strategies for Mutation Prediction. Most approaches rely on standardized processing pipelines from preprocessing ([Stain Normalization](#)) and region of interest extraction ([Tissue Type Segmentation](#)) to model training and evaluation. Common CNN or transformer-based models are used as network architectures for per-tile prediction, where all tiles from a patient's WSI inherit the same label. This label can be both continuous (eg, tumor mutational burden) or categorical (eg, the mutation status of a selected gene or the microsatellite status).^{5,79} The final prediction at WSI level is an aggregation of tile labels using, in the simplest case, majority voting (for categorical targets) or averaging (for continuous targets). Alternatively, Cao et al⁷⁶ employed multiple instance learning to weight the tile-level prediction by importance, which achieved better accuracy than standard supervised learning methods. Fu et al⁸¹ classified each tile into different malignant and nonmalignant tissue types in a pan-cancer fashion and, subsequently, predicted driver gene mutations. Almost all studies mentioned above, except for Bychkov et al,⁷¹ trained their networks on publicly available data from The Cancer Genome Atlas (TCGA).⁸²

Grading, Staging, and Prognostic Information Extraction

A typical goal of histopathology is to recognize and evaluate primary lesions, determine their histopathologic subtype and grade (as defined by respective World Health Organization classifications for different tumor entities), and derive therapeutically relevant information from these features. In the context of computational pathology, this set of tasks can be addressed by determining features known to possess prognostic or predictive value. Alternatively, extracting prognostic or predictive information directly from imaging data, molecular properties, or clinical data can be attempted. Both methodologies have recently been applied across a variety of entities.

Inference of Known Factors and Biomarkers. Computational pathology approaches to extract known markers and scores include determination of the Gleason score in prostate cancer,⁸³⁻⁸⁵ grading of gliomas,^{86,87} and automated evaluation of mitoses^{88,89} or tumor-infiltrating lymphocytes in breast⁹⁰ and head and neck cancer.⁹¹ Machine learning methods have also been applied in disease staging, for example, to assess the degree of spread to the lymph nodes, either by highlighting areas suspicious for lymphovascular invasion, as in the case of testicular cancer,⁹² or by predicting the risk of lymph node metastasis from the primary lesion, as in the case of bladder cancer.⁹³ Several studies inferred molecular properties with a prognostic value from H&E, such as microsatellite instability in gastrointestinal cancer⁹⁴ or the molecular subtype of invasive bladder cancer.⁷⁷

Image-Based Biomarkers. Finally, prognostic factors can be derived directly from histopathology images³⁸ or in combination with other clinical or molecular data, for example, from the genome or

transcriptome.^{38,95,96} This route can be followed without referring to previously known prognostic factors. Hence, although these approaches may first rely on computational predictions alone, they may also lead to identifying novel prognostic or predictive factors that lend themselves to direct human evaluation, which can be identified through explainability methods.⁹⁷ Examples of this approach include automated quantification of intratumoral stroma in rectal cancer,⁹⁸ evaluation of nuclear morphology for survival prediction in lung cancer,⁹⁹ DL-based prognosis in nasopharyngeal cancer,^{98,100} and survival prediction in colorectal cancer.^{101,102}

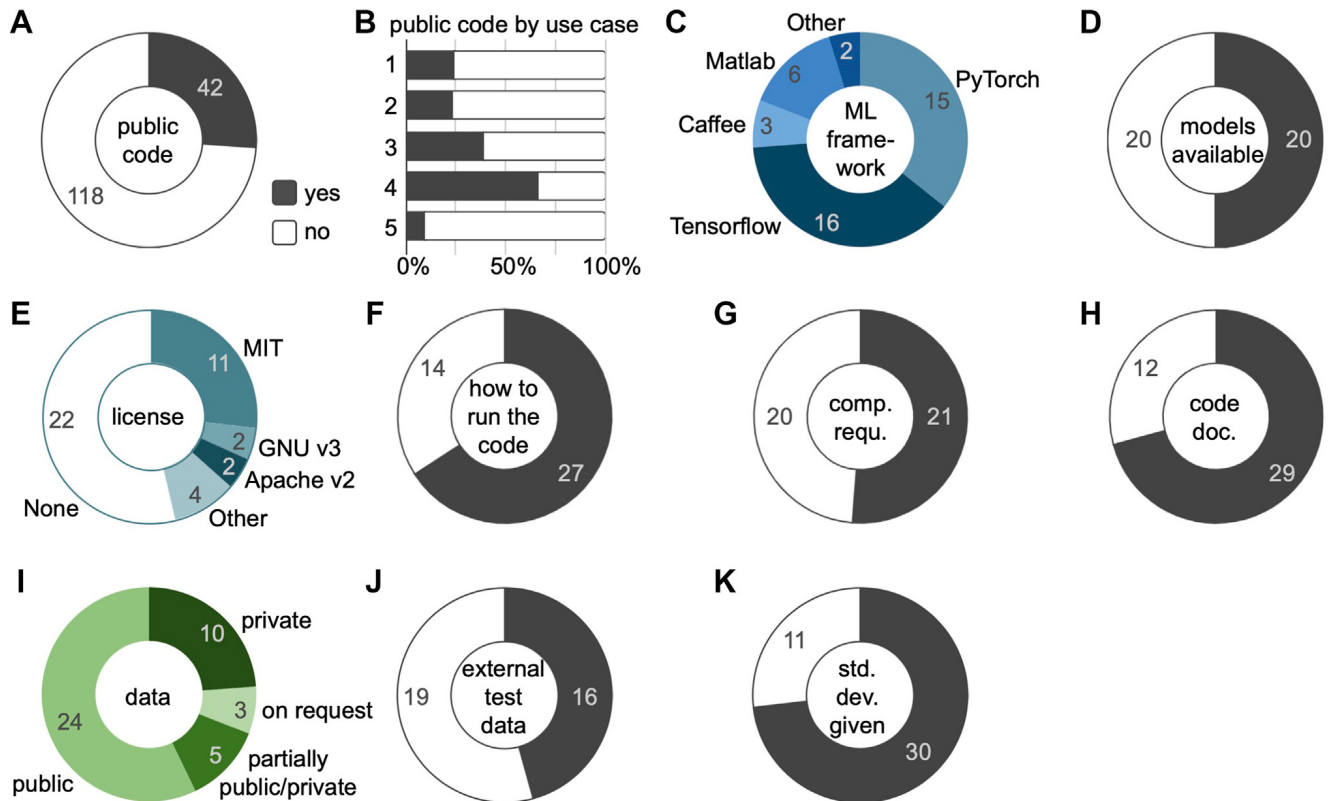
Materials and Methods

To assess reproducibility and reusability in computational pathology, we scanned whether and how code was publicly available, evaluated criteria for data access, and checked if the statistical variance of the reported findings was provided. In [Supplementary Table S2](#), we list all 41 publications (out of 160) together with the following evaluation criteria that we used for code, data, and statistical variance. (1) Inspired by the FAIR principles demanding that data should be Findable, Accessible, Interoperable, and Reusable,¹⁰³ we surveyed whether the code was made publicly available. Additionally, we noted the platform used for sharing and the programming language and machine learning frameworks employed. Furthermore, we checked for instructions for running the code, whether the code was minimally documented, and if a pretrained model was available for direct application. (2) We evaluated data access for the 41 publications with available code and checked whether the data set and required annotations were publicly available. Additionally, we recorded what kind of data had been used (eg, tiled WSIs vs entire WSIs). We also reported whether the appropriate preprocessing steps were provided and the training-validation-test split used for model development and evaluation. In terms of replicability, we specified what kind of test set had been used, whether it was similar to the training set or covered an independent cohort. (3) Finally, we checked if any measure of statistical variance of the reported findings was provided. This is one way to tackle the difficulties concerned with reproducing the results, which can be introduced on multiple levels: computer-level inaccuracies, such as floating-point numbers that can be rounded differently on different machines, architectures, or execution environments,¹⁰⁴ or algorithm-level stochasticity due to the stochastic behavior of optimization techniques. One way of dealing with this is to statistically analyze the experimental results (eg, by determining the SD of the results) and perform a sufficient exploration of hyperparameters.⁹ A straightforward evaluation approach is to repeat an experiment multiple times or over several folds of cross-validation and report mean and SD across experiments.

Results

Code Availability

In our study, 41 of the 160 publications (26%) made their code publicly available ([Fig. 4A](#)). Interestingly, the ratio of publications with code differs across the 5 use cases ([Fig. 4B](#)). For stain normalization, we retrieved 29 research articles, where only 7 (24%) of them provided code with their method. In the field of tissue type segmentation and localization, only 12 of the total 51 investigated articles (24%) had their code publicly available, and

**Figure 4.**

Analysis of our systematic literature search of computational pathology articles. (A) A quarter of the methods are published with code (26%). (B) The ratio differs across the use cases. (C) Most works (74%) used PyTorch or TensorFlow as machine learning (ML) frameworks. (D) Half of the publications with code release their final model weights. (E) Half of the code is published under a license. (F) More than half of the code includes instructions on how to run the code. (G) The computational requirements are mentioned in half of the repositories. (H) Almost three-quarters documented their code inline. (I) Large public data sets are mainly used and sometimes complemented by private cohorts. (J) Almost half of the publications used an independent cohort for evaluation. (K) The majority analyzes their results statistically.

only 3 publications provided the pretrained model weights. Among the 28 research articles that we screened for the evaluation of cell-level features, 10 articles (39%) provided code with the publications. The code of 5 of these 10 could be run in Google Colab and thus was directly applicable. For genetic alteration prediction, 8 of the 13 articles (61%) have provided their code along with their method. In survival analysis, only 4 of the 38 studies (11%) published their code. Interestingly, genetic alteration prediction has the highest ratio of published code. One reason could be that a key publication included a well-documented codebase⁵; most subsequent publications were aware of this work and, therefore, matched this standard by publishing their code, including documentations. Hence, the level of reusability in one field may depend on the preceding publications. This also strengthens the role of the publisher in the context of reproducibility and reusability in computational pathology. For the 41 articles that published their code, we checked the evaluation criteria for code, data, and statistical variance detailed in [Materials and Methods](#) and detailed them in [Supplementary Table S2](#).

Machine Learning Frameworks, Model Weights, and License

Almost 75% of the methods were implemented in Python using the open and publicly available machine learning frameworks TensorFlow (<https://www.tensorflow.org/>) (38%) and PyTorch (<https://pytorch.org/>) (36%; [Fig. 4C](#)). Pretrained model weights were only available for half of the publications that also provided

code ([Fig. 4D](#)); this makes reproducing experimental results difficult, particularly because the preprocessing steps are rarely available. Also, without model weights, a direct application without retraining the model is not possible, hampering its use by pathologists not specialized in DL. Less than half of the repositories were published under a license (46%), which is the legal requirement for reusing code since code without any license falls under the default copyright laws. This shows that the importance of licensing is not yet widely recognized in the research field representing a hurdle toward clinical implementation.

Documentation

Detailed documentation is an essential prerequisite for any step toward reproducing or reusing a published codebase. In the repositories analyzed in our study, 27 repositories (66%) contained instructions on how to run the code or train the DL model. However, only 21 repositories (51%) defined the computational requirements or provided a computational environment for running the code ([Fig. 4G](#)). Examining the main code files, 29 codebases (71%) provided at least a minimal level of documentation in the form of inline comments ([Fig. 4H](#)).

Data Sets

More than half of all methods (57%) were evaluated on publicly available data sets ([Fig. 4I](#)). Most studies (eg, all 13 studies

reviewed for genetic alteration prediction) developed their methods based on TCGA⁸²; it contains data from multiple institutions; thus, it can be split into training and test sets by cohort level. Complementing TCGA with external, mostly private, data as an independent evaluation cohort was also a common practice (Fig. 4J). Nevertheless, TCGA, with a few hundred slides for each cancer type, is insufficient to represent all cancer heterogeneities. The reliance of computational pathology on relatively few publicly available data sets renders their selection strategy and processing critical. Batch effects can be detected by DL models and lead to overestimation of the model's performance.¹⁰⁵ Therefore, we strongly encourage the development of more publicly available multi-institutional data sets.

Statistical Variance

We believe that a thorough evaluation of sources and magnitude of variability, both on an algorithmic and a data level, is essential to making modern computational pathology algorithms more reusable and generalizable. Almost three-quarters (73%) of the methods analyzed their results statistically, in which we considered all kinds of statistical notions to be statistical analysis (Fig. 4K). Many different sources of variability can be relevant to the performance of computational pathology algorithms. It is, therefore, difficult to devise a single strategy for quantifying all sources.

Discussion

It has been increasingly recognized that computational reproducibility and reusability are an essential part of good scientific practice for DL applications.¹⁰⁶⁻¹⁰⁹ Especially for the interdisciplinary field of computational pathology, both are critical requirements for enabling wider use of algorithms and the translation to clinical application.

In our survey of recently published computational pathology DL approaches, however, we found that there is still a long way to go. For stain color normalization, for example, techniques to reduce the color and intensity variations in histologic images from different laboratories can render a downstream task algorithm more generalizable. Although neural network-based stain normalization techniques have evolved considerably in recent years (*Stain Normalization*), their use in downstream applications is still limited, probably because pretrained stain normalization models are rarely available and, in most cases, code is not shared. Instead, we observe that easy-to-use algorithms not based on DL are typically applied. The lack of reusability hinders the practical application of innovative network-based methods.

Even if the code is shared, supporting documentation is necessary to reuse the code but is often missing. Especially for researchers without a computational background, even a well-documented codebase might not lower the hurdle sufficiently to adopt existing DL algorithms. Build-in plugins or models integrated into commonly used data handling tools are desirable from a clinical perspective.

Furthermore, convincing experiments on external cohorts are often missing, lowering the chances of successful reuse and translation. Most state-of-the-art methods in computational pathology are based on DL algorithms and typically require large amounts of labeled training data. Making these data available is as crucial as providing a well-documented code. We acknowledge that data and appropriate annotations cannot be publicly shared

in some cases, for example, due to legal or ethical constraints. Here, reasonable compromises, such as partial data sharing or evaluation using public data sets,¹⁰⁸ should be considered.

Finally, reusable code is likely to have an impact on the adoption of computational methods not only in a research context but also in the diagnostic routine. Although the full translation of research algorithms to a diagnostic setting is a complex process beyond the scope of the present discussion, shared and well-documented code can be of key importance for generating initial results on routine data. These could serve as an initial test for users from the diagnostic routine to evaluate whether a given algorithm might be usable and helpful in their specific setting, hence providing motivation for users to initiate the path toward integration of algorithms into their diagnostic routine. An increase of reusable and reproducible algorithms in a research context is likely to foster the adoption of DL algorithms in a routine setting.

Despite the lack of reproducibility and reusability in many computational pathology approaches, we hope the field will profit from the surging discussions, for example, in computer vision.⁹ As a step in this direction, large conferences, such as Medical Image Computing and Computer-Assisted Intervention since 2021, started to employ reproducibility checklists for authors in their submission form that will be publicly available upon acceptance of the article. We encourage the scientific community to recognize the long-term value of reproducibility and reusability and to foster their realization in computational pathology.

Author Contributions

S.J.W., C.Matek, T.P., and C.Marr designed the concept of the review. M.B. and C.Matek supported with domain-related advice. S.J.W., L.L., and D.J.E.W. compiled the evaluation criteria for reproducibility and reusability. S.J.W. and C.Matek assessed current data handling tools. For the use cases, S.J.W., D.J.E.W., and S.S.B. reviewed the publications for stain normalization, L.L. and S.J.W. for tissue type segmentation, D.J.E.W. and T.P. for evaluation of cell-level features, T.P. for genetic alteration prediction, and A.S. and C.Matek for grading, staging, and prognostic information inference. S.J.W. analyzed the results. All authors contributed to constructive discussions. S.J.W. wrote the manuscript with C.Matek, C.Marr, and T.P. and input from all other authors.

Data Availability

Not applicable.

Funding

S.J.W., L.L., and S.S.B. are supported by the Helmholtz Association under the joint research school "Munich School for Data Science—MUDS." S.J.W. is supported by the Add-on Fellowship of the Joachim Herz Foundation. C.Marr has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement No. 866411).

Declaration of Competing Interest

None reported.

Ethics Approval and Consent to Participate

Not applicable.

Supplementary Material

The online version contains supplementary material available at <https://doi.org/10.1016/j.modpat.2023.100350>

References

- Fuchs TJ, Buhmann JM. Computational pathology: challenges and promises for tissue analysis. *Comput Med Imaging Graph.* 2011;35(7–8):515–530.
- Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Springer International Publishing; 2015:234–241.
- He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017:2961–2969.
- Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115–118. <https://doi.org/10.1038/nature21056>
- Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med.* 2018;24(10):1559–1567.
- Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nat Mach Intell.* 2019;1(11):538–544.
- Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer.* 2021;124(4):686–696. <https://doi.org/10.1038/s41416-020-01122-x>
- van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med.* 2021;27(5):775–784.
- Pineau J, Vincent-Lamarre P, Sinha K, et al. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). Preprint. Posted online March 27, 2020. arXiv 2003.12206. <https://doi.org/10.48550/arXiv.2003.12206>
- Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? *Sci Transl Med.* 2016;8(341):341ps12.
- Artner R, Verliefe T, Steegen S, et al. The reproducibility of statistical results in psychological research: an investigation using unpublished raw data. *Psychol Methods.* 2021;26(5):527–546. <https://doi.org/10.1037/met0000365>
- Oala L, Fehr J, Gilli L, et al. ML4H auditing: from paper to practice. In: Alsentzer E, McDermott MBA, Falck F, Sarkar SK, Roy S, Hyland SL, eds. *Proceedings of the Machine Learning for Health NeurIPS Workshop*. Vol 136. Proceedings of Machine Learning Research; 2020:280–317.
- Li JZ. Principled approaches to robust machine learning and beyond. Massachusetts Institute of Technology. Accessed June 28, 2021. <https://dspace.mit.edu/handle/1721.1/120382?show=full>
- Rosai J. Why microscopy will remain a cornerstone of surgical pathology. *Lab Invest.* 2007;87(5):403–408.
- Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: a vendor-neutral software foundation for digital pathology. *J Pathol Inform.* 2013;4:27.
- Allan C, Burel JM, Moore J, et al. OMERO: flexible, model-driven data management for experimental biology. *Nat Methods.* 2012;9(3):245–253.
- Aubreville M, Bertram C, Klopffleisch R, Maier A. SlideRunner. In: *Bildverarbeitung Für Die Medizin 2018*. Springer Berlin Heidelberg; 2018:309–314.
- Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep.* 2017;7(1), 16878.
- Marée R, Rollus L, Stévens B, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinformatics.* 2016;32(9):1395–1401.
- Marzahl C, Aubreville M, Bertram CA, et al. EXACT: a collaboration toolset for algorithm-aided annotation of images with annotation version control. *Sci Rep.* 2021;11(1):4343.
- Chan JKC. The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. *Int J Surg Pathol.* 2014;22(1):12–32.
- Chen JM, Li Y, Xu J, et al. Computer-aided prognosis on breast cancer with hematoxylin and eosin histopathology images: a review. *Tumour Biol.* 2017;39(3), 1010428317694550.
- Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE; 2009. <https://doi.org/10.1109/ISBI.2009.5193250>
- Zheng Y, Jiang Z, Zhang H, Xie F, Shi J, Xue C. Adaptive color deconvolution for histological WSI normalization. *Comput Methods Programs Biomed.* 2019;170:107–120.
- Vahadane A, Peng T, Sethi A, et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans Med Imaging.* 2016;35(8):1962–1971.
- Anand D, Ramakrishnan G, Sethi A. Fast GPU-enabled color normalization for digital pathology. In: *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*. 2019:219–224.
- Tschuchnig ME, Oostingh GJ, Gadermayr M. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Patterns (N Y).* 2020;1(6), 100089.
- Liang H, Plataniotis KN, Li X. Stain Style Transfer of histopathology images via structure-preserved generative learning. In: *Machine Learning for Medical Image Reconstruction*. Springer International Publishing; 2020:153–162.
- Shaban MT, Baur C, Navab N, Albarqouni S. Staingan: stain style transfer for digital histological images. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019:953–956.
- Cohen JP, Luck M, Honari S. Distribution matching losses can hallucinate features in medical image translation. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*. Springer International Publishing; 2018:529–536.
- Kang H, Luo D, Feng W, et al. StainNet: a fast and robust stain normalization network. *Front Med (Lausanne)*. 2021;8:746307.
- Yamashita R, Long J, Banda S, Shen J, Rubin DL. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Trans Med Imaging.* 2021;40(12):3945–3954. <https://doi.org/10.1109/TMI.2021.3101985>
- Wagner SJ, Khalili N, Sharma R, et al. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference*. 2021. https://doi.org/10.1007/978-3-030-87237-3_25
- Zhao B, Chen X, Li Z, et al. Triple U-net: hematoxylin-aware nuclei segmentation with progressive dense feature aggregation. *Med Image Anal.* 2020;65, 101786.
- Vellal AD, Sirinukunwattan K, Kensler KH, et al. Deep learning image analysis of benign breast disease to identify subsequent risk of breast cancer. *JNCI Cancer Spectr.* 2021;5(1):kaa119.
- Graham S, Epstein D, Rajpoot N. Dense steerable filter CNNs for exploiting rotational symmetry in histology images. *IEEE Trans Med Imaging.* 2020;39(12):4124–4136.
- Jayapandian CP, Chen Y, Janowczyk AR, et al. Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains. *Kidney Int.* 2021;99(1):86–101.
- Zhao K, Li Z, Yao S, et al. Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer. *EBioMedicine.* 2020;61:103054.
- Rączkowski A, Możejko M, Zambonelli J, Szczurek E. ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. *Sci Rep.* 2019;9(1):14347.
- Gal Y, Ghahramani Z. Bayesian convolutional neural networks with Bernoulli approximate variational inference. Preprint. Posted online June 6, 2015. arXiv 1506.02158. <https://doi.org/10.48550/arXiv.1506.02158>
- Wang S, Wang T, Yang L, et al. ConvPath: a software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. *EBioMedicine.* 2019;50:103–110.
- Schmitz R, Madesta F, Nielsen M, et al. Multi-scale fully convolutional neural networks for histopathology image segmentation: from nuclear aberrations to the global tissue architecture. *Med Image Anal.* 2021;70:101996.
- van Rijthoven M, Balkenhol M, Siliņa K, van der Laak J, Ciompi F. HookNet: multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Med Image Anal.* 2021;68:101890.
- Chen CL, Chen CC, Yu WH, et al. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nat Commun.* 2021;12(1):1193.
- Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016:2921–2929.
- Silva-Rodríguez J, Colomer A, Naranjo V. WeGleNet: a weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images. *Comput Med Imaging Graph.* 2021;88, 101846.
- Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng.* 2021;5(6):555–570. <https://doi.org/10.1038/s41551-020-00682-w>
- Kumar N, Verma R, Anand D, et al. A multi-organ nucleus segmentation challenge. *IEEE Trans Med Imaging.* 2020;39(5):1380–1391.
- Cui Y, Zhang G, Liu Z, Xiong Z, Hu J. A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images. *Med Biol Eng Comput.* 2019;57(9):2027–2043.
- Jin YW, Jia S, Ashraf AB, Hu P. Integrative data augmentation with U-net segmentation masks improves detection of lymph node metastases in

- breast cancer patients. *Cancers*. 2020;12(10). <https://doi.org/10.3390/cancers12102934>
51. Jung H, Lodhi B, Kang J. An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images. *BMC Biomed Eng*. 2019;1:24.
 52. Mahmood F, Borders D, Chen RJ, et al. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Trans Med Imaging*. 2020;39(11):3257–3267.
 53. Qu H, Wu P, Huang Q, et al. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE Trans Med Imaging*. 2020;39(11):3655–3666.
 54. Veta M, Heng YJ, Stathonikos N, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Med Image Anal*. 2019;54:111–121.
 55. Roux L, Racoceanu D, Loménié N, et al. Mitosis detection in breast cancer histological images an ICPR 2012 contest. *J Pathol Inform*. 2013;4(1):8. <https://doi.org/10.4103/2153-3539.112693>
 56. Roux L, Racoceanu D, Capron F, et al. MITOS & ATYPIA—detection of mitosis and evaluation of nuclear atypia score in breast cancer histological images. An ICPR 2014 Contest. 22nd International Conference on Pattern Recognition. Image Pervasive Access Lab(IPAL) Lab; June 27, 2014; Singapore.
 57. Pati P, Foncubierta-Rodríguez A, Goksel O, Gabrani M. Reducing annotation effort in digital pathology: a co-representation learning framework for classification tasks. *Med Image Anal*. 2021;67:101859.
 58. Jiménez G, Racoceanu D. Deep learning for semantic segmentation vs. classification in computational pathology: application to mitosis analysis in breast cancer grading. *Front Bioeng Biotechnol*. 2019;7:145.
 59. Lafarge MW, Bekkers EJ, Pluim JPW, Duits R, Veta M. Roto-translation equivariant convolutional networks: application to histopathology image analysis. *Med Image Anal*. 2021;68:101849.
 60. Lei H, Liu S, Xie H, Kuo JY, Lei B. An improved object detection method for mitosis detection. *Conf Proc IEEE Eng Med Biol Soc*. 2019;2019:130–133.
 61. Sohail A, Khan A, Wahab N, Zameer A, Khan S. A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images. *Sci Rep*. 2021;11(1):6215.
 62. Wollmann T, Rohr K. Deep consensus network: aggregating predictions to improve object detection in microscopy images. *Med Image Anal*. 2021;70:102019.
 63. Lei H, Liu S, Elazab A, Gong X, Lei B. Attention-guided multi-branch convolutional neural network for mitosis detection from histopathological images. *IEEE J Biomed Health Inform*. 2021;25(2):358–370.
 64. Mahmood T, Arsalan M, Owais M, Lee MB, Park KR. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs. *J Clin Med*. 2020;9(3):749. <https://doi.org/10.3390/jcm9030749>
 65. Ashley EA. Towards precision medicine. *Nat Rev Genet*. 2016;17(9):507–522.
 66. Singh MP, Rai S, Pandey A, Singh NK, Srivastava S. Molecular subtypes of colorectal cancer: an emerging therapeutic opportunity for personalized medicine. *Genes Dis*. 2021;8(2):133–145.
 67. Zhang H, Kalirai H, Acha-Sagredo A, Yang X, Zheng Y, Coupland SE. Piloting a deep learning model for predicting nuclear BAP1 immunohistochemical expression of uveal melanoma from hematoxylin-and-eosin sections. *Transl Vis Sci Technol*. 2020;9(2):50.
 68. Kim RH, Nomikou S, Dawood Z, et al. *A Deep Learning Approach for Rapid Mutational Screening in Melanoma*. Cold Spring Harbor Laboratory; 2019: 610311. <https://doi.org/10.1101/610311>
 69. Lu Z, Xu S, Shao W, et al. Deep-learning-based characterization of tumor-infiltrating lymphocytes in breast cancers from histopathology images and multiomics data. *JCO Clin Cancer Inform*. 2020;4:480–490.
 70. Anand D, Kurian NC, Dhage S, et al. Deep learning to estimate human epidermal growth factor receptor 2 status from hematoxylin and eosin-stained breast tissue images. *J Pathol Inform*. 2020;11:19.
 71. Bychkov D, Linder N, Tiulpin A, et al. Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. *Sci Rep*. 2021;11(1):4037. <https://doi.org/10.1038/s41598-021-83102-6>
 72. Wang S, Rong R, Yang DM, et al. Computational staining of pathology images to study the tumor microenvironment in lung cancer. *Cancer Res*. 2020;80(10):2056–2066.
 73. Yu KH, Wang F, Berry GJ, et al. Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks. *J Am Med Inform Assoc*. 2020;27(5):757–769.
 74. Jang HJ, Lee A, Kang J, Song IH, Lee SH. Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. *World J Gastroenterol*. 2020;26(40):6207–6223.
 75. Echle A, Grabsch HI, Quirke P, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology*. 2020;159(4):1406–1416.e11.
 76. Cao R, Yang F, Ma SC, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer. *Theranostics*. 2020;10(24):11080–11091.
 77. Woerl AC, Eckstein M, Geiger J, et al. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *Eur Urol*. 2020;78(2):256–264.
 78. Tsou P, Wu CJ. Mapping driver mutations to histopathological subtypes in papillary thyroid carcinoma: applying a deep convolutional neural network. *J Clin Med*. 2019;8(10):1675. <https://doi.org/10.3390/jcm8101675>
 79. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer*. 2020;1(8):789–799.
 80. Noorbakhsh J, Farahmand S, Foroughi Pour A, et al. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat Commun*. 2020;11(1):6367.
 81. Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer*. 2020;1(8):800–810. <https://doi.org/10.1038/s43018-020-0085-8>
 82. Gutman DA, Cobb J, Somanna D, et al. Cancer Digital slide archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Am Med Inform Assoc*. 2013;20(6):1091–1098.
 83. Steiner DF, Nagpal K, Sayres R, et al. Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies. *JAMA Netw Open*. 2020;3(11):e2023267.
 84. Bulten W, Balkenhol M, Belinga JA, et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod Pathol*. 2021;34(3):660–671.
 85. Nagpal K, Foote D, Liu Y, et al. Erratum: publisher correction: development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med*. 2019;2:113.
 86. Truong AH, Sharmanska V, Limbäck-Stanic C, Grech-Sollars M. Optimization of deep learning methods for visualization of tumor heterogeneity and brain tumor grading through digital pathology. *Neurooncol Adv*. 2020;2(1):vdaa110.
 87. Rathore S, Niazi T, Iftikhar MA, Chaddad A. Glioma grading via analysis of digital pathology images using machine learning. *Cancers*. 2020;12(3):578. <https://doi.org/10.3390/cancers12030578>
 88. Chang MC, Mrkonjic M. Review of the current state of digital image analysis in breast pathology. *Breast J*. 2020;26(6):1208–1212.
 89. Pantanowitz L, Hartman D, Qi Y, et al. Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. *Diagn Pathol*. 2020;15(1):80.
 90. Balkenhol MC, Ciompi F, Swiderska-Chadaj Z, et al. Optimized tumour infiltrating lymphocyte assessment for triple negative breast cancer prognostics. *Breast*. 2021;56:78–87.
 91. Shaban M, Khurram SA, Fraz MM, et al. A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma. *Sci Rep*. 2019;9(1), 13341.
 92. Ghosh A, Sirinukunwattana K, Khalid Alham N, et al. The potential of artificial intelligence to detect lymphovascular invasion in testicular cancer. *Cancers*. 2021;13(6):1325. <https://doi.org/10.3390/cancers13061325>
 93. Harmon SA, Sanford TH, Brown GT, et al. Multiresolution application of artificial intelligence in digital pathology for prediction of positive lymph nodes from primary tumors in bladder cancer. *JCO Clin Cancer Inform*. 2020;4:367–382.
 94. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*. 2019;25(7):1054–1056.
 95. Hao J, Kosaraju SC, Tsaku NZ, Song DH, Kang M. PAGE-Net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. *Pac Symp Biocomput*. 2020;25:355–366.
 96. Failmezger H, Muralidhar S, Rullan A, de Andrea CE, Sahai E, Yuan Y. Topological tumor graphs: a graph-based spatial model to infer stromal recruitment for immunosuppression in melanoma histology. *Cancer Res*. 2020;80(5):1199–1209.
 97. Tosun AB, Pullara F, Becich MJ, Taylor DL, Fine JL, Chennubhotla SC. Explainable AI (xAI) for anatomic pathology. *Adv Anat Pathol*. 2020;27(4):241–250.
 98. Geessink OGF, Baidoshvili A, Klaase JM, et al. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. *Cell Oncol (Dordr)*. 2019;42(3):331–341.
 99. Alsubaie NM, Snead D, Rajpoot NM. Tumour nuclear morphometrics predict survival in lung adenocarcinoma. *IEEE Access*. 2021;9:12322–12331. <https://doi.org/10.1109/ACCESS.2021.3049582>
 100. Zhang F, Zhong LZ, Zhao X, et al. A deep-learning-based prognostic nomogram integrating microscopic digital pathology and macroscopic magnetic resonance images in nasopharyngeal carcinoma: a multi-cohort study. *Ther Adv Med Oncol*. 2020;12, 1758835920971416.
 101. Kather JN, Krisam J, Charoentong P, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med*. 2019;16(1), e1002730.
 102. Abbet C, Zlobec I, Bozorgtabar B, Thiran JP. Divide-and-rule: self-supervised learning for survival analysis in colorectal cancer. *Medical Image Computing and Computer Assisted Intervention—MICCAI*. 2020;2020:480–489. https://doi.org/10.1007/978-3-030-59722-1_46
 103. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.

104. Hill DRC. Repeatability, Reproducibility, Computer Science and High Performance Computing: stochastic simulations can be reproducible too. In: *2019 International Conference on High Performance Computing Simulation*. HPCS; 2019:322–323.
105. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun.* 2021;12(1):4423. <https://doi.org/10.1038/s41467-021-24698-1>
106. Stodden V, McNutt M, Bailey DH, et al. Enhancing reproducibility for computational methods. *Science.* 2016;354(6317):1240–1241.
107. Hutson M. Artificial intelligence faces reproducibility crisis. *Science.* 2018;359(6377):725–726.
108. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. *Nature.* 2020;586(7829):E14–E16.
109. Carpenter AE, Kametsky L, Eliceiri KW. A call for bioimaging software usability. *Nat Methods.* 2012;9(7):666–670.



Built to Last? Reproducibility and Reusability of Deep Learning Algorithms in Computational Pathology

Author:

Sophia J. Wagner, Christian Matek, Sayedali Shetab Boushehri, Melanie Boxberg, Lorenz Lamm, Ario Sadafi, Dominik J.E. Winter, Carsten Marr, Tingying Peng

Publication: Modern Pathology

Publisher: Elsevier

Date: January 2024

© 2023 United States & Canadian Academy of Pathology. Published by Elsevier Inc.

Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: <https://www.elsevier.com/about/our-business/policies/copyright#Author-rights>

BACK

CLOSE WINDOW

© 2024 Copyright - All Rights Reserved | [Copyright Clearance Center, Inc.](#) | [Privacy statement](#) | [Data Security and Privacy](#)
| [For California Residents](#) | [Terms and Conditions](#) Comments? We would like to hear from you. E-mail us at customercare@copyright.com

Privacy - Terms

A.9. Make deep learning algorithms in computational pathology more reproducible and reusable

This publication was published as a peer-reviewed comment in the international journal *Nature Medicine* in 2022. The one-page summary can be found in Section 3.9.

Wagner, S. J.*, Matek, C.*, Shetab Boushehri, S., Boxberg, M., Lamm, L., Sadafi, A., ... & Peng, T. (2022). Make deep learning algorithms in computational pathology more reproducible and reusable. *Nature Medicine*, 28(9), 1744-1746. *Contributed equally

Copyright: Copyright © 2022, Springer Nature America, Inc. Reproduced with permission from Springer Nature. The copyright clearance is attached after the original publication.

Make deep learning algorithms in computational pathology more reproducible and reusable

Greater emphasis on reproducibility and reusability will advance computational pathology quickly and sustainably, ultimately optimizing clinical workflows and benefiting patient health.

Sophia J. Wagner, Christian Matek, Sayedali Shetab Boushehri, Melanie Boxberg, Lorenz Lamm, Ario Sadafi, Dominik J. E. Waibel, Carsten Marr and Tingying Peng

For over a century, trained pathologists have visually assessed stained tissue sections for diagnosis under a microscope to classify diseases and derive prognostic information. Recently, the digitization of histological slides has enabled computational methods from artificial intelligence (AI; Table 1), in particular deep learning algorithms, to aid in the assessment of microscopic image data in pathology^{1,2} (see Fig. 1 for a typical workflow). Algorithms in computational pathology can achieve the performance level of human experts on specific, well-defined tasks^{3–5} or predict information that humans would not be able to diagnose from microscopic images alone^{6,7}.

However, from the thousands of publications in computational pathology in the past decade, only a small percentage of algorithms have been reused by other researchers, and even fewer have entered the routine clinical workflow^{8,9}. There are several reasons that prevent algorithms from being used in a routine clinical setting, the most important of which is the lack of reproducibility and reusability^{10–12}. Limited code accessibility and documentation make it difficult even for machine learning specialists to reproduce results; this is particularly relevant for the interdisciplinary field of computational pathology, in which potential users from medicine may not have access to sufficient resources for the re-implementation of algorithms.

Another reason for the limited use of deep learning algorithms is the generalization gap; it is common for the performance of an algorithm to be higher on the training data but lower on external test data from other sources. In part, the generalization gap arises from the intrinsic heterogeneity of histopathology data, including variabilities in staining protocols and different scanners being used in different pathology labs¹³.

Table 1 | Common use cases of AI in computational pathology

Use of AI	Aim
Stain normalization	Reduce the variability in the visual appearance of whole-slide images that results from differences in the slide preparation and the digitization process, such as staining protocols, scanner settings, or post-processing.
Tissue type segmentation	Segment whole-slide images into the tissue types present for further quantitative analysis.
Evaluation of cell-level features	Quantitatively analyze cell-level features, such as cell density and the abundance of dividing cells in tumor tissue (mitosis detection).
Genetic alteration prediction	Predict genetic alterations from whole-slide images as an effective and tissue-sparing alternative to the evaluation of molecular assays, which has become increasingly relevant to the diagnostic workup and selection of therapeutic pathways.
Tumor grading, staging and prognostic information extraction	Establish a medical diagnosis to derive therapeutically relevant information based on imaging, molecular, or clinical data.

Lack of reproducibility and reusability

To assess the reproducibility, replicability and reusability of computational pathology algorithms, we recently evaluated 161 peer-reviewed articles published between January 2019 and March 2021¹⁴. In line with findings from other machine learning application domains¹⁵, in particular in health¹⁶, only one-quarter (26%) of all reviewed publications made their code publicly available and fulfilled our minimum requirement for reproducibility and reusability. Among these 42 papers, three-quarters (30 out of 42) analyzed their results statistically, less than half (20 out of 42) released their trained model weights, and only about one-third (16 out of 42) used an independent cohort for evaluation.

The proportion of publications providing code varied drastically between the use cases mentioned in Table 1, with only 11% of code shared for publications involving grading, staging and prognostic information extraction, compared with more than half of publications (61%) in genetic alteration prediction providing code¹⁴. On the basis of these findings,

recommendations are proposed to address the lack of reproducibility and reusability in computational pathology (Table 2).

Involve clinicians

Domain experts, such as clinicians or pathologists, should be included early in the model development process to define clinically relevant tasks and identify modifications that are required by the use cases. A key parameter that should be determined at the outset is the data basis required for a particular task, such as relevant sub-cohorts that should be defined and represented in the training data. Only through collaboration with domain experts can biases in model training be avoided.

Data publishing and splitting

Deep learning algorithms typically require large amounts of labeled training data. Multi-institutional datasets should ideally be publicly shared to facilitate future research. However, in many use cases, data and appropriate annotations cannot be publicly shared, in part because of legal or ethical constraints. Partial data sharing, evaluation

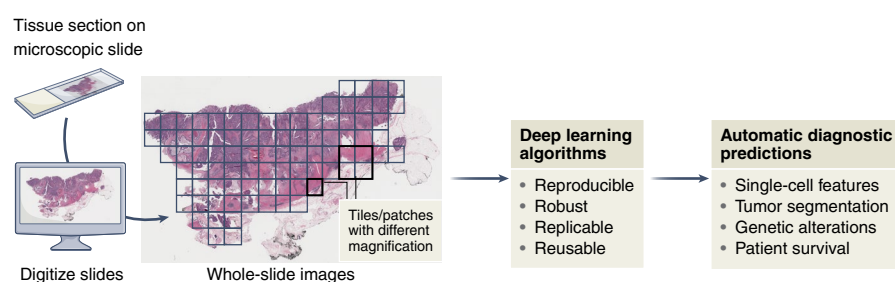


Fig. 1 | Workflow in computational pathology. Deep learning algorithms can aid pathologists in the assessment of whole-slide images, by segmenting data into tissue types or by predicting genetic alterations such as mutations, or patient survival. To be used in a routine pathological workflow, a deep learning algorithm needs to be reproducible (same results can be obtained with given code and data), robust (similar results despite small perturbations in data), replicable (similar performance on external data from an independent cohort) and reusable (well-documented and integrable in existing pathology workflows).

Table 2 | Recommendations to achieve better reproducibility and reusability in computational pathology

Stage	Recommendation
Data	Share the data publicly
	Split the dataset according to patients, and, if possible, also according to cohort
	Publish the data split
Pre-processing	Publish the data handling tools or pipelines that were used for: segmenting the tissue area; tiling the whole-slide images; stain normalization; and data augmentation
	Publish algorithm code together with required specifications to reproduce the results
	Document the code and provide instructions for running it
Model development	Create ready-to-use scripts for direct application and publish the model weights
	Repeat the experiments several times
	Publish the statistical analysis of the variance
Final evaluation	Journals and conferences should use reproducibility checklists to evaluate publications and check that the codebase is well-documented
	Authors should be required to publish their code

of the algorithm using public datasets, or publishing the model predictions without the test data to enable the statistical analysis of the results may allow the algorithm to be tested and verified, while maintaining data publishing restrictions¹².

Good scientific practice in machine learning requires the available data to be split into training and test sets, in order to ensure that the test data are kept separate and only used for the evaluation of the final algorithm. Both the training and test sets should be composed of a similar distribution of the target diagnoses or mutations. Although this is used in most cases, researchers still need to be careful in data preparation and splitting to avoid data

leakage and overestimation of the model performance.

Data should usually be split at a patient level, such that all data from one patient are in the same set, to prevent the algorithm from learning patient-specific features that are then recognized in the test set. We also strongly encourage splitting the data at the cohort level (that is, not using all available data sources for training), whenever data from several institutions are available because this further enables the measurement of the algorithm's generalizability¹³.

Data preprocessing and model training

Whole-slide images are often preprocessed into small patches to ensure efficient

handling for neural network training. In a data-constrained setting, characteristic of many digital pathology applications, the use of appropriate normalization and augmentation methods is a key step to obtaining generalizable and robust results. To ensure the reusability of models, the data handling tools or pipelines that were used for preprocessing, and their parameters, should be published.

The code should be well documented and shared together with the final obtained model weights, to allow for independent external validation of the results. For simplified usage, ready-to-use scripts should be created. Information on the algorithm hyperparameters (including their selection) along with the hardware specifications and timings are crucial and must be included.

Evaluation and publication

The final evaluation of the algorithm on an external test set should provide appropriate measures of the variation of performance on different data, together with the test metric. The focus of model evaluation has increasingly shifted away from the isolated consideration of predictive accuracy toward the evaluation of multiple aspects, including model calibration, robustness, simplicity, and interpretability¹⁷. Publishing a thorough evaluation of the sources and magnitude of variability, both at the algorithmic and data level, is an important step towards making modern computational pathology algorithms more reusable and generalizable.

Once the model has passed the evaluation stage, it should be made publicly available through code archives, such as GitHub or Zenodo, or dedicated repositories for deep learning models, such as ModelZoo. Deep learning models often rely on complex dependencies, and so container systems such as Docker or CodeOcean can facilitate the reuse of published algorithms and considerably speed up model evaluation among users and developers from different institutions. Publishers should require authors to provide a well-documented codebase, which can increase reusability in the field, as with genetic alteration prediction. Checklists can be used to evaluate the reproducibility of a publication.

The use of machine learning in digital pathology is likely to increase greatly in the near future. Algorithms should therefore be built to last, satisfying the clinicians' needs for interpretability, usability, and robustness. Computational pathology can benefit from the discussions on reproducibility that are happening in the machine learning community, allowing the field to fulfill its full potential for diagnostic medicine. □

Sophia J. Wagner ^{1,2,11}, Christian Matek^{3,4,11},
Sayedali Shetab Boushehri ^{3,5,6},
Melanie Boxberg^{7,8}, Lorenz Lamm ^{1,2,9},
Ario Sadafi ^{2,3}, Dominik J. E. Waibel ^{3,10},
Carsten Marr ^{1,3,12} ✉ and
Tingying Peng ^{1,3,12} ✉

¹Helmholtz AI, Helmholtz Munich – German Research Center for Environmental Health, Neuherberg, Germany. ²Department of Informatics, Technical University of Munich, Garching, Germany. ³Institute of AI for Health, Helmholtz Munich – German Research Center for Environmental Health, Neuherberg, Germany. ⁴Institute of Pathology, University Hospital Erlangen, Erlangen, Germany. ⁵Department of Mathematics, Technical University of Munich, Garching, Germany. ⁶Data Science, Pharmaceutical Research and Early Development Informatics (pREDi), Roche Innovation Center Munich (RICM), Penzberg, Germany. ⁷Institute of Pathology, Technical University Munich, Munich, Germany. ⁸Institute of Pathology Munich-North, Munich, Germany. ⁹Helmholtz Pioneer Campus, Helmholtz Munich – German Research Center

for Environmental Health, Neuherberg, Germany.

¹⁰School of Life Sciences, Technical University of Munich, Weihenstephan, Germany. ¹¹These authors contributed equally: Sophia J. Wagner, Christian Matek. ¹²These authors jointly supervised this work: Carsten Marr, Tingying Peng.

✉e-mail: carsten.marr@helmholtz-munich.de; tingying.peng@helmholtz-munich.de

Published online: 8 August 2022
<https://doi.org/10.1038/s41591-022-01905-0>

References

1. Fuchs, T. J. & Buhmann, J. M. *Comput. Med. Imaging Graph.* **35**, 515–530 (2011).
2. Förtsch, S. et al. *Dtsch. Arztebl. Int.* **118**, 194–204 (2021).
3. Esteva, A. et al. *Nature* **542**, 115–118 (2017).
4. Coudray, N. et al. *Nat. Med.* **24**, 1559–1567 (2018).
5. Matek, C. et al. *Nat. Mach. Intell.* **1**, 538–544 (2019).
6. Lu, M. Y. et al. *Nature* **594**, 106–110 (2021).
7. Fu, Y. et al. *Nat. Cancer* **1**, 800–810 (2020).
8. Echle, A. et al. *Br. J. Cancer* **124**, 686–696 (2021).
9. van der Laak, J. et al. *Nat. Med.* **27**, 775–784 (2021).
10. Hutson, M. *Science* **359**, 725–726 (2018).
11. Stodden, V. et al. *Science* **354**, 1240–1241 (2016).
12. Haibe-Kains, B. et al. *Nature* **586**, E14–E16 (2020).

13. Howard, F. M. et al. *Nat. Commun.* **12**, 4423 (2021).
14. Wagner, S. J. et al. Preprint at *medRxiv* <https://doi.org/10.1101/2022.05.15.22275108> (2022).
15. Pineau, J. et al. *J. Mach. Learn.* **22**, 1–20 (2021).
16. McDermott, M. B. A. et al. *Sci. Transl. Med.* **13**, eabb1655 (2021).
17. Wiens, J. et al. *Nat. Med.* **25**, 1337–1340 (2019).

Acknowledgements

We thank P. Schüffler for feedback. S.J.W., L.L. and S.S.B. are supported by the Helmholtz Association under the joint research school “Munich School for Data Science - MUDS”. S.J.W., L.L. and T.P. were funded by Helmholtz Association’s Initiative and Networking Fund through Helmholtz AI. S.S.B. has received funding by F. Hoffmann-la Roche LTD (No grant number is applicable). L.L. acknowledges a fellowship from the Boehringer Ingelheim Fonds. C.M. has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant no. 866411).

Author contributions

S.J.W., Ch.M., T.P. and C.M. wrote the manuscript. All authors contributed to the review¹⁴ in preparation for this Comment.

Competing interests

The authors declare no competing interests.



RightsLink



Make deep learning algorithms in computational pathology more reproducible and reusable

SPRINGER NATURE

Author: Sophia J. Wagner et al

Publication: Nature Medicine

Publisher: Springer Nature

Date: Aug 8, 2022

Copyright © 2022, Springer Nature America, Inc.

Author Request

If you are the author of this content (or his/her designated agent) please read the following. If you are not the author of this content, please click the Back button and select no to the question "Are you the Author of this Springer Nature content?".

Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:

To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).

The author and any academic institution, where they work, at the time may reproduce the contribution for the purpose of course teaching.

To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them.

To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website.

Authors wishing to use the published version of their article for promotional use or on a web site must request in the normal way.

If you require further assistance please read Springer Nature's online [author reuse guidelines](#).

For full paper portion: Authors of original research papers published by Springer Nature are encouraged to submit the author's version of the accepted, peer-reviewed manuscript to their relevant funding body's archive, for release six months after publication. In addition, authors are encouraged to archive their version of the manuscript in their institution's repositories (as well as their personal Web sites), also six months after original publication.

v1.0

BACK

CLOSE WINDOW