Decoding the Black Box : Enhancing Interpretability and Trust in Artificial Intelligence for Biomedical Imaging - a Step Toward Responsible Artificial Intelligence

Mehdi Ounissi

Sorbonne Université

Doctoral Thesis

Decoding the Black Box: Enhancing Interpretability and Trust in AI for Biomedical Imaging—A Step Toward Responsible Artificial Intelligence

Author:
Mehdi Ounissi

Supervisor:
Prof. Daniel Racoceanu

Jury members:

Prof. Charles Kervrann	Jury president
Prof. Olivier Lézoray	Reviewer
Prof. Thomas Walter	Reviewer
Prof. Dominique Berrebi	Examiner
Prof. Xavier Descombes	Examiner
Prof. Vlad Popovici	Examiner

A thesis submitted in fulfillment of the requirements for the degree of Ph.D. in Computer Science

in the

ARAMIS Lab
Sorbonne Université, Institut du Cerveau - Paris Brain Institute (ICM), CNRS, Inria, Inserm, AP-HP Hôpital de la Pitié Salpêtrière

September, 2024

Abstract

In an era dominated by AI, its opaque decision-making –known as the "black box" problem– poses significant challenges, especially in critical areas like biomedical imaging where accuracy and trust are crucial. Our research focuses on enhancing AI interpretability in biomedical applications. We have developed a framework for analyzing biomedical images that quantifies phagocytosis in neurodegenerative diseases using time-lapse phase-contrast video microscopy. Traditional methods often struggle with rapid cellular interactions and distinguishing cells from backgrounds, critical for studying conditions like frontotemporal dementia (FTD). Our scalable, real-time framework features an explainable cell segmentation module that simplifies deep learning algorithms, enhances interpretability, and maintains high performance by incorporating visual explanations and by model simplification. We also address issues in visual generative models, such as hallucinations in computational pathology, by using a unique encoder for Hematoxylin and Eosin staining coupled with multiple decoders. This method improves the accuracy and reliability of synthetic stain generation, employing innovative loss functions and regularization techniques that enhance performance and enable precise synthetic stains crucial for pathological analysis. Our methodologies have been validated against several public benchmarks, showing top-tier performance. Notably, our framework distinguished between mutant and control microglial cells in FTD, providing new biological insights into this unproven phenomenon. Additionally, we introduced a cloud-based system that integrates complex models and provides real-time feedback, facilitating broader adoption and iterative improvements through pathologist insights. The release of novel datasets, including video microscopy on microglial cell phagocytosis and a virtual staining dataset related to pediatric Crohn's disease, along with all source codes, underscores our commitment to transparent open scientific collaboration and advancement. Our research highlights the importance of interpretability in AI, advocating for technology that integrates seamlessly with user needs and ethical standards in healthcare. Enhanced interpretability allows researchers to better understand data and improve tool performance.

Résumé

À une époque dominée par l'IA, son processus décisionnel opaque, connu sous le nom de problème de la "boîte noire", pose des défis significatifs, particulièrement dans des domaines critiques comme l'imagerie biomédicale où la précision et la confiance sont essentielles. Notre recherche se concentre sur l'amélioration de l'interprétabilité de l'IA dans les applications biomédicales. Nous avons développé un cadre pour l'analyse d'images biomédicales qui quantifie la phagocytose dans les maladies neurodégénératives à l'aide de la microscopie vidéo à contraste de phase en accéléré. Les méthodes traditionnelles ont souvent du mal avec les interactions cellulaires rapides et la distinction des cellules par rapport aux arrière-plans, essentielles pour étudier des conditions telles que la démence frontotemporale (DFT). Notre cadre évolutif et en temps réel comprend un module de segmentation cellulaire explicable qui simplifie les algorithmes d'apprentissage profond, améliore l'interprétabilité et maintient des performances élevées en incorporant des explications visuelles et par simplifications. Nous abordons également les problèmes dans les modèles génératifs visuels, tels que les hallucinations en pathologie computationnelle, en utilisant un encodeur unique pour la coloration Hématoxyline et Éosine couplé avec plusieurs décodeurs. Cette méthode améliore la précision et la fiabilité de la génération de coloration synthétique, utilisant des fonctions de perte innovantes et des techniques de régularisation qui renforcent les performances et permettent des colorations synthétiques précises cruciales pour l'analyse pathologique. Nos méthodologies ont été validées contre plusieurs benchmarks publics, montrant des performances de premier ordre. Notamment, notre cadre a distingué entre les cellules microgliales mutantes et contrôles dans la DFT, fournissant de nouveaux aperçus biologiques sur ce phénomène non prouvé. De plus, nous avons introduit un système basé sur le cloud qui intègre des modèles complexes et fournit des retours en temps réel, facilitant une adoption plus large et des améliorations itératives grâce aux insights des pathologistes. La publication de nouveaux ensembles de données, incluant la microscopie vidéo sur la phagocytose des cellules microgliales et un ensemble de données de coloration virtuelle lié à la maladie de Crohn pédiatrique, ainsi que tous les codes sources, souligne notre engagement envers la collaboration scientifique ouverte et transparente et l'avancement. Notre recherche met en évidence l'importance de l'interprétabilité dans l'IA, plaidant pour une technologie qui s'intègre de manière transparente avec les besoins des utilisateurs et les normes éthiques dans les soins de santé. Une interprétabilité améliorée permet aux chercheurs de mieux comprendre les données et d'améliorer les performances des outils.

Scientific production

First author journal papers

1. **Ounissi, M.**, Latouche, M. & Racoceanu, D. PhagoStat a scalable and interpretable end to end framework for efficient quantification of cell phagocytosis in neurodegenerative disease studies. Nature: Scientific Reports 14, 6482 (2024).
   Paper: https://doi.org/10.1038/s41598-024-56081-7
   Code : https://github.com/ounissimehdi/PhagoStat
   Code : https://github.com/ounissimehdi/Point2Cell
   Data : https://zenodo.org/records/10803492

Submitted first author journal papers

1. **Ounissi, M.**, Sarbout, I., Hugot, J. P., Martinez-Vinson, C., Berrebi, D., & Racoceanu, D. (2024). Scalable, Trustworthy Generative Model for Virtual Multi-Staining from H&E Whole Slide Images. arXiv preprint.
   https://arxiv.org/abs/2407.00098.

Submitted patents

1. **Ounissi, M.**, Berrebi, D., & Racoceanu, D. (2024). **EP 24 305 224.8**: Trustworthy and Scalable Unpaired Virtual Multi-Staining.

2. **Ounissi, M.**, Berrebi, D., & Racoceanu, D. (2024). **EP 24 305 221.4**: Trustworthy and Scalable Paired Virtual Multi-Staining.

Conference papers

1. Valabregue, R., Khemir, I., Auzias, G., Rousseau, F., & **Ounissi, M.** (2024, July). Unraveling Systematic Biases in Brain Segmentation: Insights from Synthetic Training. In Medical Imaging with Deep Learning.
   https://openreview.net/pdf?id=B3xO0c2Q3h

2. J. Arslan, **M. Ounissi**, H. Luo, M. Lacroix, P. Dupré, P. Kumar, A. Hodgkinson, S. Dandou, R. Larive, C. Pignodel, L. Le Cam, O. Radulescu, D. Racoceanu (2023, April). Efficient 3D reconstruction of whole slide images in melanoma, Proc. SPIE 12471, Medical Imaging 2023: Digital and Computational Pathology, 124711S.
https://doi.org/10.1117/12.2657473

3. Jiménez, G., Kar, A., **Ounissi, M.**, Ingrassia, L., Boluda, S., Delatour, B., Stimmer, L. & Racoceanu, D. (2022, September). Visual deep Learning-Based explanation for neuritic plaques segmentation in Alzheimer's disease using weakly annotated whole slide histopathological images. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 336-344). Cham: Springer Nature Switzerland.
https://doi.org/10.1007/978-3-031-16434-7_33

4. K. Maňoušková, V. Abadie, **M. Ounissi**, G. Jimenez, L. Stimmer, B. Delatour, S. Durrleman, D. Racoceanu (2022, April). Tau protein discrete aggregates in Alzheimer's disease: neuritic plaques and tangles detection and segmentation using computational histopathology, Proc. SPIE 12039, Medical Imaging 2022: Digital and Computational Pathology, 1203908.
https://doi.org/10.1117/12.2613154

Submitted conference papers

1. Sarbout, I., **Ounissi, M.**, Milea, D. & Racoceanu, D. (2024). Deep Learning for Navigation of the Visually Impaired using Synthetic Data from Blind Digital Twins.

Talk

1. **Ounissi, M.**, Gabriel, J. & Racoceanu, D., (2024 April). "PBI-ILBS" Computation pathology basics and applications: Special session organized by the Paris Brain Institute / Sorbonne University, Paris, France and hosted by the Institute of Liver and Biliary Sciences (ILBS), New Delhi, India.
Computational Pathology Presentations

Conference abstract

1. Garay, G. J., Kar, A., **Ounissi, M.**, Stimmer, L., Delatour, B., & Racoceanu, D. (2022). Interpretable Deep Learning in Computational Histopathology for refined identification of Alzheimer's Disease biomarkers. Alzheimer's & Dementia, 18, e065363.
https://doi.org/10.1002/alz.065363

Scientific communications

1. Racoceanu, D. ,**Ounissi, M.**, & Kergosien Y.L. (2024, February). Explicabilité en Intelligence Artificielle ; vers une IA Responsable - Instanciation dans le domaine de la santé. Réf: H5030 v1, Édition techniques de l'ingénieur.
   https://doi.org/10.51257/a-v1-h5030

2. Jimenez, G., Kar, A., **Ounissi, M.**, & Racoceanu, D. (2023). Empowering Researchers for Understanding Alzheimer's Disease using Explainable AI.
   XAI-Alzheimers-Disease

3. **Ounissi, M.**, (2022). An open and responsible AI, with Mehdi Ounissi. Sorbonne Center for Artificial Intelligence (SCAI). "2 minutes of AI"

4. Arslan, J., **Ounissi, M.**, Jiménez, G., Kar, A., & Racoceanu, D. (2022). Responsible artificial intelligence: a review of current trends. Winter School AI4Health.
   https://hal.science/hal-03834390

Contents

List of Figures

List of Tables

Introduction

In the contemporary landscape of artificial intelligence (AI), a significant evolution has been observed, especially in its application to biomedical imaging. This domain crucially relies on advanced AI methodologies to enhance diagnostic accuracy and treatment efficacy. However, the widespread integration of AI in such sensitive areas is hindered by the opaque nature of machine learning models, often termed the "black box" issue. This opacity challenges the acceptability and reliability of AI systems, as it obscures the causal pathways through which these systems derive their conclusions.

The principal concern addressed in this thesis revolves around the opaque decision-making processes of AI systems used in biomedical imaging. These systems, driven by complex algorithms and deep neural networks, exhibit high performance but lack interpretability. This opacity is problematic in clinical environments, where understanding the basis of diagnostic or therapeutic decisions is crucial for trust and ethical medical practice. Recent regulatory developments, including mandates from the General Data Protection Regulation (GDPR), underscore the urgent need for transparency in AI systems that process personal health data, highlighting the legal and ethical imperatives for explainable AI.

This research proposes to systematically "decode" these complex AI systems, enhancing their transparency and accountability through interpretable methodologies. The objective is to develop a framework that not only elucidates the operational mechanics of AI models but also embeds ethical considerations to elevate the trustworthiness and acceptance of AI systems in biomedical contexts.

To address these challenges, the thesis will explore several cutting-edge approaches in the realm of explainable artificial intelligence (XAI). It will assess the applicability and effectiveness of various interpretability techniques, such as feature activation maps visualization, model simplification, knowledge guided and self-inspection, specifically adapted for deep learning models employed in biomedical imaging. This exploration will be grounded in empirical validation against public benchmarks and will involve collaborations with medical professionals to align AI outputs with clinical insights and needs.

Moreover, the research will investigate the dynamic between model complexity and interpretability, seeking to find a balance that maintains high predictive performance without compromising the system's interpretability. It will also examine the impact of interpretability enhancements on the clinical decision-making process, measuring how well medical practitioners can understand, trust, and effectively use AI-assisted research/diagnostic tools.

The outcome of this thesis is expected to contribute to the field of AI in healthcare by providing a scientifically robust and ethically sound compliant framework that enhances the interpretability and reliability of AI systems. This will pave the way for more responsible AI applications, ensuring that these powerful tools aid rather than obscure the critical decisions

made in medical practice. By advancing our understanding of how AI models can be made transparent and accountable, this work aims to foster a new era of trust and collaboration between AI technologies and biomedical professionals.

Contributions

This thesis presents significant advancements in the field of XAI, emphasizing the integration of explainability within AI systems to fulfill both scientific and regulatory requirements. The primary contribution of this thesis is the demonstration that XAI not only enhances understanding of the decision-making processes in deep learning models but also improves model performance and control while reducing model size.

This concept is illustrated through two distinct use cases. The first case explores XAI for cell segmentation in video microscopy, focusing on the real-time quantification of phagocytosis in unstained cells, which has implications for understanding neurodegenerative diseases such as Frontotemporal Dementia (FTD). A comprehensive, end-to-end framework is introduced that is both interpretable and scalable, featuring a novel explainable cell segmentation module that augments the interpretability of deep learning methods without compromising performance (refer to Chapter 2).

The second use case addresses XAI for generative models in histopathology, particularly through the development of virtual staining techniques using generative AI models (refer to Chapter 3). These models convert Hematoxylin and Eosin (H&E) slides into multiple outputs, thereby reducing the environmental and resource costs associated with traditional staining methods. This approach not only enhances control over these models through knowledge-guided training but also improves the performance and scalability of synthetic staining. It further incorporates real-time self-inspection mechanisms as safeguards to ensure quality and trust in critical healthcare applications.

In addition to these primary contributions, the thesis explores the state-of-the-art in XAI, emphasizing the dual role of explainability in AI as both a vital area of scientific inquiry and a regulatory necessity (refer to Chapter 1). It addresses the challenges in achieving transparency, reproducibility, interpretability, and causality in AI systems, underscoring their importance for trust and acceptance in AI applications. The thesis also makes further structural and conceptual contributions by structuring the state-of-the-art in XAI and its terminology. Significant initiatives include the development of methods to mitigate systematic bias and the proposal of a conceptual framework for XAI systems, which aims to guide the development and evaluation of robust and accountable AI systems towards Responsible AI.

Overall, the thesis encapsulates a series of methodological and theoretical contributions that significantly advance the boundaries of XAI, equipping the field to address emerging ethical, legal, and societal challenges while fostering ongoing research and practical applications in both academic and industrial contexts.

Outline of the manuscript

The manuscript is organized into the following chapters:

- Chapter 1 provides a comprehensive overview of the current state-of-the-art and highlight the dual role of explainability in AI, treating it as both an essential area of scientific inquiry and a regulatory necessity. This chapter investigates the interplay between methodological considerations and compliance with various AI regulatory frameworks, including the GDPR and the EU AI Act. It addresses the challenges and necessities in achieving transparency, reproducibility, interpretability, and causality in AI systems, emphasizing their importance for trust and acceptance in AI applications.

- Chapter 2, we present a detailed methodology with a focus on XAI and open science principles. We demonstrate the application of XAI in complex instance-segmentation tasks through a concrete use case: the quantification of phagocytosis in video microscopy within the context of neurodegenerative diseases such as Frontotemporal Dementia (FTD), showing that XAI facilitates a deeper understanding of the models and achieves better performance compared to existing state-of-the-art methods.

- Chapter 3 addresses the challenges associated with generative models, such as hallucinations, and proposes new training methodologies. It features a comprehensive case study on virtual staining in digital histopathology. Through the use of XAI, improvements in model performance and control are discussed, alongside innovative ways to leverage discriminators as safeguards. This supports the development of explainable, trustworthy, and accessible AI systems through cloud-based computation.

- In the final Chapter 4, we discuss our results and provide preliminary findings on mitigating pseudo-label bias, as referenced in Section 4.2.1. We also outline perspectives for generic XAI systems and consider the ethical implications of AI deployments in real-world scenarios, emphasizing the need for robust governance frameworks in Section 4.2.2. Additionally, we propose future research directions toward fully responsible AI.

Chapter 1

Explainable Artificial Intelligence

Scientific publication and communication

Elements of the Sections 1.1, 1.2, and 1.3 are published in:

Racoceanu, D, **Ounissi, M.**, Kergosien Y. L. "Explicabilité en Intelligence Artificielle ; vers une IA Responsable - Instanciation dans le domaine de la santé." (2024) Techniques de l'ingenieur, 29 Feb. 2024. https://doi.org/10.51257/a-v1-h5030.

Jiménez, G., Kar, A., **Ounissi, M.**, Ingrassia, L., Boluda, S., Delatour, B., & Racoceanu, D. Visual deep learning-based explanation for neuritic plaques segmentation in Alzheimer's disease using weakly annotated whole slide histopathological images. (2022) MICCAI 2022 https://doi.org/10.1007/978-3-031-16434-7_33.

Summary

In this chapter, we examine the dual nature of explainability in artificial intelligence (AI), presenting it as both a scientific domain and a regulatory necessity. This chapter explores the intersection of methodological considerations and compliance requirements across various AI regulatory frameworks. We begin by analyzing the current legislative landscape, covering the General Data Protection Regulation (GDPR), proposed new EU AI Act, and extending to non-EU regulatory frameworks, including key recommendations from UNESCO and OECD. This regulatory context highlights the need to integrate explainability into AI systems to meet ethical and legal standards. Furthermore, we address the conceptual aspects of explainability, discussing paradoxes in transparency, challenges in reproducibility, and the quest for interpretability and causality –factors critical to the trustworthiness and acceptance of AI systems. We then explore various methods used to achieve explainability in AI, including global and local approaches, model-agnostic methods, and model-specific strategies. These techniques help illuminate the decision-making processes of complex algorithms.

Résumé

Dans ce chapitre, nous examinons la double nature de l'explicabilité dans l'intelligence artificielle (IA), la présentant à la fois comme un domaine de recherche scientifique et une nécessité réglementaire. Ce chapitre explore l'intersection entre les considérations méthodologiques et les exigences de conformité dans divers cadres réglementaires de l'IA. Nous commençons par analyser le paysage législatif actuel, en couvrant le Règlement Général sur la Protection des Données (RGPD), la nouvelle loi de l'UE sur l'IA, et en étendant notre analyse aux cadres réglementaires non européens, y compris les recommandations clés de l'UNESCO et de l'OCDE. Ce contexte réglementaire souligne la nécessité d'intégrer l'explicabilité dans les systèmes d'IA pour répondre aux normes éthiques et légales. De plus, nous abordons les aspects conceptuels de l'explicabilité, en discutant des paradoxes de la transparence, des défis de la reproductibilité, et de la quête d'interprétabilité et de causalité – des facteurs critiques pour la fiabilité et l'acceptation des systèmes d'IA. Nous explorons ensuite diverses méthodes utilisées pour atteindre l'explicabilité dans l'IA, y compris les approches globales et locales, les méthodes agnostiques aux modèles et les stratégies spécifiques aux modèles. Ces techniques aident à éclairer les processus de prise de décision des algorithmes complexes.

1.1 Explainability and Compliance in AI Regulations

The scientific community has made considerable strides in the field of artificial intelligence (AI), which has led to the emergence of significant ethical and regulatory challenges. AI's remarkable capabilities and the implementation of its practical applications have raised justified societal concerns, necessitating a regulated approach to its deployment. In response, notable governmental initiatives in the United States (E. O. o. t. President, M. Holden, and Smith, 2016) and France (Villani et al., 2018) have emphasized ethical considerations, shaping new legal norms. This direction has been further supported by the formation of a recognized research domain in AI ethics, which encompasses ethical AI (AI operating under ethical constraints). This area has developed its own journals, learned societies, and educational programs. Pioneering efforts such as the IEEE initiative and research from centers in Oxford, Munich, and France's ALLISTENE Ethics Committee focus on aligning AI with human rights, particularly concerning equity, safety, privacy, and dignity.

One example of the ethical challenges in AI is the automated bank loan allocation, which, as detailed in a report by the Bank of England (Philippe Bracke and Sen, 2019), risks discriminating against certain groups if profiling based on race, ethnicity, or religion is used. Although such personal attributes are prohibited from being recorded in computerized files in France, other countries may not have similar restrictions, and indirect profiling using allowed data remains a concern. AI algorithms, especially those derived from learning procedures without legal constraints, can inadvertently engage in such discrimination, raising substantial responsibility issues for multiple stakeholders: (1) the algorithm designer, (2) the end-user, (3) the affected individuals, (4) potential certifying authorities, and (5) the data providers, especially if the data is not sufficiently aggregated to ensure only statistical treatment. The roles of (6) data annotators, who often rely on (7) annotation tools and (7bis) chosen annotation lexicons that could introduce bias, are also critical. Additionally, (8) transfer learning can perpetuate pre-existing biases, and the complexity increases with federated learning.

Moreover, in France, the law provides individuals harmed by automated decisions the right to a human review, which should include explanations of the decision-making process. This legal framework has also introduced a relatively new requirement for "explainability" of algorithms, aiming to ensure that algorithms adhere to legal and safety constraints. Explainability not only aids in verifying compliance but also enhances user trust and informs responsible parties. Beyond merely fulfilling regulatory obligations, explainability also holds interest for AI developers, particularly in applications involving human interaction, such as humanoid robots for assisting dependent persons. Here, ethical considerations and the respect for human dignity remain paramount, both from a legal and a computational perspective.

Currently, in France, legal obligations regarding AI primarily derive from three legislative texts, illustrating the nation's commitment to overseeing this transformative technology responsibly.

1.1.1 Computing, files, and freedoms regulation

Law No. 78-17 of January 6, 1978[1] was initiated due to advancements in applied computer science, particularly in data processing. It focuses primarily on managing personal data files, which include data about individuals who can be identified either directly or indirectly when stored in computerized databases. Since the law's enactment, the collection of individual data has significantly increased with the proliferation of payment terminals, web platforms, online stores, mobile phones, and social networks. Furthermore, the ability to re-identify individuals from their data often surpasses traditional anonymization techniques.

The law establishes the National Commission on Informatics and Liberty (Commission Nationale de l'Informatique et des Libertés, CNIL) as the regulatory authority for personal data. It sets forth the obligations for entities that manage computerized files, including the requirement to declare these files and the prohibition of recording sensitive information related to an individual's racial, ethnic, religious, political, or union affiliations. Additionally, it clarifies the rights of individuals whose data are stored, providing them the right to be informed, access, correct, and erase their data.

Amended by Law No. 2004-801 on August 6, 2004[2], regarding the protection of natural persons concerning the processing of personal data, the legislation introduces stringent restrictions on the automated processing of individual data, particularly concerning artificial intelligence. Article 10 of this law mandates that no judicial decision assessing a person's behavior may be based exclusively on the automated processing of personal data intended to evaluate specific personality traits. It also stipulates that no decision that has legal effects on a person can be based solely on automated data processing used to profile the individual or assess certain personality aspects.

1.1.2 General Data Protection Regulation (GDPR)

The General Data Protection Regulation (GDPR[3]) is a European regulation that differs from a directive, which requires transposition into the laws of each member state to be enforceable. Instead, the GDPR applies directly as it is published in the Official Journal of the European Union (EU) and becomes effective simultaneously across all member states from its date of implementation. This regulation was partly prompted by advancements in AI and mandates that any processing of this nature should include comprehensive safeguards. These safeguards must provide specific information to the individual concerned and grant them rights to human intervention, to express their views, to receive explanations about decisions made after such evaluations, and to challenge these decisions.

Additionally, to ensure processing is fair and transparent, given the specific circumstances and context in which personal data is handled, the data controller is required to employ suitable mathematical or statistical procedures for profiling. They must also implement adequate technical and organizational measures to correct inaccuracies in personal data and minimize the risk of errors. It is imperative to secure personal data in a way that

[1] Law No. 78-17 of January 6, 1978 (France)
[2] Law No. 2004-801 on August 6, 2004 (France)
[3] The General Data Protection Regulation "GDPR" (EU)

considers the potential risks to the interests and rights of individuals, preventing discriminatory effects based on racial or ethnic origin, political opinions, religion or beliefs, trade union membership, genetic status, health status, or sexual orientation. Furthermore, automated decision-making and profiling that involve special categories of personal data should only be allowed under strict conditions.

1.1.3 Expanding access to available technologies ethically

The French bioethics law[4], which is part of the Health Code, addresses significant medical applications of AI. It mandates explainability for algorithms or intelligent systems involved in decision-making that affects individuals. The law uses the term "explainability" without providing a detailed definition, leaving the methods for achieving this quality somewhat abstract. Article 17 specifically introduces the following key provisions:

1. Healthcare professionals using a medical device that processes data algorithmically, trained on large datasets for preventive, diagnostic or therapeutic purposes, must ensure that the data subject has been informed and, where appropriate, advised of the implications of interpreting the data.

2. Health professionals must be notified about the use of such algorithmic data processing. They should have access to the patient data used in this process and the results it produces.

3. Designers of the algorithmic treatment specified above are required to ensure that its functionality is explainable to its users.

In this context, the users in 3 are healthcare professionals who receive detailed explanations about how the algorithmic processing works. They are also tasked with managing patient information and securing informed consent.

This scenario highlights that even historical legal frameworks can be relevant in modern contexts involving AI. Imagine a situation where an intelligent device manages patient interactions directly, much like intelligent systems handle customer interactions on e-commerce platforms. However, such an arrangement would violate the Health Code as it would constitute unauthorized practice of medicine. Currently, laws explicitly prohibit the delegation of medical duties to machines. In such instances, the liability concerns for the designers and providers of these devices would be significant, paralleling, if not exceeding, those in the realm of autonomous vehicles.

1.1.4 EU regulation proposal for AI act

In addition to potential enhancements to the GDPR within EU member states, legislation pertaining to AI is also under development in the EU, the United States, and the United Kingdom, all of which carry significant economic and strategic implications. The European Commission has highlighted[5] the critical need for the EU to lead in establishing ambitious

[4] Law No. 2021-1017 of August 2, 2021, on bioethics (France)
[5] Artificial intelligence act and amending certain union legislative acts (EU)

new global standards. Following the Commission's white paper titled "Artificial Intelligence – A European approach focused on excellence and trust"[6] which proposed initial regulatory directions, a draft European regulation on AI has been released. This draft has received detailed feedback from the European Economic and Social Committee (Directorate-General for Communications Networks and Technology (European Commission), 2020). Furthermore, the European Parliament's resolution from October 20, 2020, on ethical aspects of AI, robotics, and related technologies advocates for a future regulatory framework built upon the Union's laws and values, emphasizing transparency, explainability, fairness, and accountability, though the term "explainability" is mentioned only once.

1.1.5 Non-EU regulatory movement: the United States roadmap

The global movement recognizing the societal and economic importance of AI, along with its ethical and regulatory implications, is significantly influenced by an initiative from the United States executive branch in 2016. This initiative is marked by the publication of the White House's white paper "Preparing for the Future of Artificial Intelligence" (U. S. (E. O. o. t. President, Holdren, and M. Smith, 2016), which serves as a roadmap for AI development in the U.S. This roadmap includes recommendations for government ministries and agencies responsible for funding, such as creating open databases and computational platforms to broaden access to AI, expanding AI education, and integrating ethics training into AI curricula.

Instead of initially legislating based on an abstract concept like explainability, the U.S., through the Defense Advanced Research Projects Agency (DARPA), simultaneously launched a call for research proposals on "Explainable AI" in 2016 (Agency, 2016; Gunning and Aha, 2019). This call encouraged various professional societies to develop new guidelines and standards for best practices. Concurrently, the Institute of Electrical and Electronics Engineers (IEEE) introduced the "Global Initiative for Ethical Considerations in AI and Autonomous Systems" (How, 2018) and issued a call for feedback in the form of the "Ethically Aligned Design" report (K. Shahriari and M. Shahriari, 2017). Compared to its French and European counterparts, the U.S. approach was highly effective, using a bottom-up strategy that deeply engaged professionals and researchers from the start. This collaboration promptly led to the creation of applicable standards, providing a robust foundation and valuable experience for potential future legislative texts.

1.1.6 UNESCO recommendations and reports, OECD directives

Numerous international organizations are actively addressing the ethical challenges presented by AI and are preparing for the advent of global regulatory frameworks. UNESCO's World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) has notably produced an influential report (Ethics of Scientific Knowledge and Technology, 2019), followed by a Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021). This recommendation borrows aspects of the French approach to AI ethics, with a particular focus on explainability. It further explores this concept by linking it closely with

[6] White Paper on Artificial Intelligence: a European approach to excellence and trust (EU)

transparency. It clarifies that explainability in AI systems involves understanding the inputs, outputs, and operations of various algorithmic components, and how these contribute to the system's results. Thus, explainability requires that results and the processes leading to them be transparent, making all elements clear and traceable within the given context. Additionally, the OECD issued a directive in 2019 concerning AI (OECD, 2019), which also advocates for transparency and explainability.

1.2 Conceptual Landscape of Explainability

The conceptual landscape of explainability, particularly within the realm of AI, represents a comprehensive domain at the intersection of computer science, legal regulations, and linguistic precision, which collectively compound the challenges of elucidation.

1.2.1 Transparency paradox

AI systems are often characterized as "black boxes", a term that reflects their inherent opacity. Although transparency is frequently advocated as a remedy (Barredo Arrieta et al., 2020), it embodies dual meanings that can contribute to confusion. On one hand, transparency might imply invisibility or non-interference, as in the usage, "the software update is transparent to the user," implying that the update occurs without any noticeable effect or requirement for intervention from the user's side. On the other hand, it can denote complete visibility, such as granting access to the software's source code. This ambiguity extends beyond semantics and poses substantial practical challenges within the scientific community. For example, deep neural networks, despite their perceived opacity, allow full access to all components, including structure and model weights, thus enabling replication of results. Nevertheless, fully understanding their operations is not guaranteed and, at a minimum, necessitates a detailed post-hoc analysis akin to reverse engineering a compiled binary without access to the source code.

The relationship between transparency and explainability in AI systems is complex and merits careful consideration. This analysis highlights that transparency does not necessarily lead to explainability. A system can be transparent in its operations without providing sufficient insights to explain its decision-making processes. For instance, while an XAI system may be protected under industrial property rights and comply with regulatory frameworks like the GDPR, it may still lack transparency (Jobin, Ienca, and Vayena, 2019).

It is critical to address the common misconception equating transparency with explainability. Given its ambiguity, the term "transparency" will be excluded from discussions pertaining to the general context of XAI in our manuscript. This will prevent confusion and reserve the use of "transparency" solely for discussions related to academic open science. Recognizing the fundamental distinction between these concepts is essential, as it has significantly influenced legislative decisions opposing the mandatory disclosure of AI algorithms. Such an approach facilitates the protection of confidentiality, consistent with industrial property policies (Wachter, B. Mittelstadt, and Floridi, 2017). Recognizing this distinction is crucial as it allows legislators to navigate the complexities of maintaining confidentiality under industrial property law while enforcing explainability requirements—that is, elucidating how a model processes inputs to arrive at specific decisions "the legal right to explanation" (Selbst and Powles, 2018). Therefore, it is clear that explainability must be recognized as an autonomous concept, crucial not only from a computer science perspective in dealing with AI models but also as a legal imperative ("the right to explanation"). This underscores the need for regulations that mandate clarity on how decisions are made within AI systems, independent of their transparency status.

1.2.2 Reproducibility

Reproducibility (Gundersen and Kjensmo, 2018; Gibney, 2022; Tsima, 2023), while desirable in algorithmic processes, does not inherently guarantee explainability. This is analogous to software engineering, where replicating a bug does not necessarily elucidate its underlying cause. In deterministic algorithms, consistent outcomes are expected when initiated from identical states. However, various factors contribute to non-determinism in machine learning. These include the use of stochastic optimization algorithms—although setting the seeds of pseudo-random number generators can achieve reproducibility—the uncontrolled sequencing of massively parallel processors, variations in training datasets and the sequence in which examples are presented, and potential dependencies arising from transfer learning involving other entities.

It is crucial to distinguish between the reproducibility of the final decision algorithm, which is typically deterministic (when the model's weights are frozen), and the learning process, which is inherently more complex and less predictable. Regulatory attention primarily focuses on the final decision-making algorithm, emphasizing the need for the scientific community to ensure its functional integrity. However, the rise of algorithms that undergo continuous updates through perpetual learning cycles (continual/incremental learning) presents significant challenges. These systems, which adapt based on ongoing inputs of new data, complicate efforts to maintain reproducibility and pose even greater challenges for explainability. As models evolve without explicit retraining phases, tracking changes and understanding the influence of new data on decision processes become increasingly difficult. This evolution highlights the need for advanced methods to ensure the reliability and transparency of algorithms under continuous learning paradigms.

1.2.3 Interpretability

Considerable research in the field of XAI (Barredo Arrieta et al., 2020) has underscored the importance of the term "interpretability," establishing it as a cornerstone in scholarly discussions. Interpretability plays different roles across various contexts: in statistics, it serves as a crucial, though often informally applied, heuristic; in machine learning literature, as highlighted by (Doshi-Velez and B. Kim, 2017), it is defined specifically as "the ability to explain or to present in terms comprehensible to a human." This definition not only associates interpretability closely with explainability and comprehension but also emphasizes its role in making intelligent systems more accessible and understandable to users.

In recent years, there has been a notable shift in the field toward favoring the term "Explainable AI" over "Interpretable AI." This change reflects a move toward more precise terminology, emphasizing the need for systems that can clearly articulate their processes and decisions. Additionally, the concept of 'interpretation' has long been established in mathematical logic, with foundational works by Tarski (Szczerba, 1977; Friedman, 2007) in the mid-20th century. This concept extends to various domains such as ontologies, the semantic web, and bioinformatics, where principles of interpretation are crucial for understanding complex data structures and relationships.

In our analysis of comprehension phenomena within intelligent systems, we draw upon this broad and formal framework. We acknowledge, as is commonly accepted in legal studies, that interpretations can vary widely and are not inherently singular. This recognition is crucial for understanding the diverse ways in which AI systems can be interpreted and the implications of these interpretations for both users and developers.

1.2.4 Causality

The explanation of decisions involving multiple factors or criteria can be seen as an elucidation of the "causes" behind those decisions. The concept of causality has been a subject of scholarly investigation since antiquity, with Aristotle notably identifying four types of causes: formal, material, efficient, and final. Today, this concept spans both the humanities—including philosophy, history, sociology, political science, and economics—and the natural sciences, such as physics, biology, statistics, and logic (Beebee, Hitchcock, and Menzies, 2009). In legal studies, causality is crucial for addressing issues of responsibility (legally distinct from the more recent term "responsible AI"), aligning with the motivations for XAI.

Each discipline applies its own set of rules, making it essential to specify the context in which each author operates to avoid misinterpretation. A notable example is the critique of research on perceptrons—the precursors to neural networks—by proponents of symbolic AI, such as Minsky and Papert (Olazaran, 1996; Minsky and Papert, 2017), who inappropriately extrapolated the results. In the realm of statistical theories of causality, which aim to discern causal relationships between variables—rather than mere associations—from observational data, particularly through considerations of conditional independence, it is important to temper the claims of Pearl and Mackenzie, 2018 regarding Bayesian Networks with critiques from statisticians like (Dawid, 2010). Recent comprehensive works (Allen, 2020) highlight the challenges in employing these methods, known as Graphical Models, which, according to the Lauritzen school, are not inherently Bayesian.

Counterfactual theories of causality, which are conceptually simpler and more practical, have gained widespread acceptance in political and legal sciences. These fields are particularly relevant to our discussion as they focus on human-centered issues. (Wachter, B. D. Mittelstadt, and Russell, 2017) not only provide a counterfactual methodology for explaining automated decisions to individuals but also suggest three objectives for such explanations: (1) assisting the individual in understanding the decision, (2) guiding them through potential legal challenges to the decision, and (3) aiding them in developing strategies to adapt to the algorithm. For example, in the case of a denied bank loan, it could be useful to inform the individual about how much they need to adjust their spending, reduce their loan request, or decrease the frequency of overdraft occurrences. Importantly, the paper advocates for maintaining algorithmic confidentiality while outlining legitimate demands that can be met without exposing the "black box."

1.2.5 Trustworthiness

The pursuit of trustworthiness is often intuitively regarded as a fundamental objective in the development of XAI models. However, equating a model's explainability solely with

its ability to engender trust may not fully align with established criteria for explainability (Barredo Arrieta et al., 2020). In this context, trustworthiness can be defined as the consistent reliability with which a model performs as expected in specified scenarios. While trustworthiness is an essential quality of any XAI system, it is a misconception to assume that all trustworthy models are inherently explainable. For instance, search engines like Google exemplify a situation where trust and explainability diverge. Users generally trust these algorithms to deliver relevant and high-quality results, even though their operations remain largely opaque due to their proprietary nature (Schultheiß and Lewandowski, 2023).

Furthermore, trustworthiness is not a straightforward attribute to quantify. Unlike non-XAI systems, where trust may primarily derive from performance, trust in an XAI system is based on its ability to explain its decisions. This distinction is crucial, particularly in cases where an XAI system may underperform relative to a non-XAI system but still earn trust through its capacity to acknowledge its limitations by indicating uncertainty (e.g., stating "I do not know"). This underscores a significant relationship between explainability and trust, suggesting that while the presence of explainability in XAI systems implies trustworthiness, the converse does not necessarily hold. Moreover, while the core value of an XAI system may not rely solely on superior performance, it can still benefit from it. Therefore, an XAI system can be trustworthy even if its performance is sub-optimal, provided it maintains explicit explainability in its functioning.

1.2.6 XAI towards responsible AI

Advancements in AI necessitate rigorous definitions and guidelines for XAI. While there is currently no consensus on a definitive description of XAI, several criteria or guidelines are proposed for a system to be recognized as fully XAI-compliant. Key among these is the integration of interactive features—whether textual, visual, symbolic, or otherwise—that empower users (both human and autonomous systems) to request explanations of AI-derived outcomes and conclusions. This "right to explanation" should accommodate various forms of elucidation, such as textual, visual, and causal mechanisms, tailored to specific use cases and tasks in a comprehensible manner. Furthermore, an XAI system must acknowledge the limits of its knowledge by admitting uncertainties, such as stating "I do not know." Additionally, users should have a basic understanding of the system's construction, including its adherence to traceability, fairness, and the balance of training data used, as well as the explicit elements involved in deriving the results.

The concept of Responsible Artificial Intelligence (RAI), as delineated by (Trocin et al., 2023), encompasses XAI while expanding its scope to include additional ethical and operational considerations crucial for advancing sustainable and ethical AI technologies. This broader framework notably includes privacy considerations, which are increasingly critical in domains like cloud computing. For instance, RAI emphasizes implementing the right to be forgotten, a crucial element in maintaining user trust by allowing the deletion of personal data upon request.

Additionally, RAI addresses the dual-use potential of AI systems, referring to the risk

that technologies initially developed for beneficial purposes could be repurposed for malevolent uses. A relevant example is the creation of "deepfakes," which involve generating synthetic images, voices, or combined video and voice representations. To mitigate such risks, RAI mandates the incorporation of protective mechanisms, such as digital watermarks, to verify the origins and integrity of digital content.

Security measures are also a critical component of RAI. This includes robust encryption of sensitive data to protect against unauthorized access and breaches, thereby preserving the confidentiality and integrity of personal and organizational information.

Furthermore, the RAI framework considers the environmental and energy consumption impacts of AI systems. It promotes the optimization of these systems to minimize their computational footprint, addressing the urgent need to reduce the energy demands of large-scale AI computations and mitigate their environmental impact. This focus not only enhances the efficiency of AI applications but also aligns with broader sustainability goals.

In summary, RAI represents a holistic approach that extends beyond the technical capabilities of XAI to incorporate ethical practices, security protocols, and environmental considerations, ensuring that AI technologies are developed and deployed in a manner that is accountable, secure, and sustainable.

1.3 Families of Methods for Explainability

1.3.1 Components to explain

1.3.1.1 Statistical learning and heuristics for interpretability

The field of statistical learning encompasses various methodologies that extend traditional statistical data analysis tools. Traditionally, statistics is divided into descriptive and inferential branches. Inferential statistics often link hypotheses about the phenomenon under study to explicit probabilistic models (Ghahramani, 2015), which can generate random data similar to the observed data. By quantitatively comparing the distribution of observed data to those generated by different models—whether known through analytical means or numerical simulations—researchers can select an appropriate model and validate the corresponding hypotheses. This methodology is widely practiced within a broad community and provides a consensus framework for statistical interpretation (Daly and Bourke, 2008). The relevance of explanations provided by different models can be assessed using tools understood by practitioners across various experimental sciences such as medicine, biology, and psychology.

Numerical methodologies, including simulation, Monte Carlo methods (Metropolis and Ulam, 1949), resampling, and Bayesian approaches (Howson and Urbach, 2006; Dienes, 2014; Kruschke and Liddell, 2018), have facilitated the use of complex models and allowed for the transcendence of traditional statistical approximations, such as Gaussian assumptions. Beyond linear models like linear regression (Montgomery, Peck, and Vining, 2021) and generalized linear models (e.g., logistic regression (Kleinbaum et al., 2002; Hosmer Jr, Lemeshow, and Sturdivant, 2013), tree-based methods for regression and classification (Breiman, 2001) have been developed. Subsequently, ensemble methods such as

"bagging"(Breiman, 1996; Strobl, Malley, and Tutz, 2009), which stands for Bootstrap Aggregation of resampled trees, and random forests—which combine the decisions of multiple trees, often through voting—have emerged. Boosting methods like XGBoos (T. Chen et al., 2015) further enhance these base methods. Vapnik's statistical learning theory (V. Vapnik, 2013) is crucial for quantifying a learning system's ability to generalize from data, particularly in avoiding overfitting. From this research, support vector machines (SVMs) (Boser, Guyon, and V. N. Vapnik, 1992; Noble, 2006) have emerged, resembling generalized linear methods by reducing nonlinear classification problems to the separation of points using a hyperplane.

These statistical learning methods form a common language within the community of users and are supported by software platforms like scikit-learn (Pedregosa et al., 2011) and libraries in the R language, which also incorporate tools for explainability, such as Shapley values (Lundberg and Lee, 2017a). This progression of methods ranges from elementary techniques well understood by users to more complex methods requiring greater effort to comprehend. Researchers in statistical learning adopt explainability approaches that are useful for tackling the inherently more challenging problems posed by deep learning. For example, while the final decision of a decision tree is comprehensible due to its sequence of elementary decisions, explaining a random forest's decision involves constructing a decision tree that approximates the forest's decisions with high fidelity.

In addition to these methods, radiomics (Kumar et al., 2012; Yip and Aerts, 2016; Mayerhoefer et al., 2020) in medical imaging begins with a suite of classical image processing filters or treatments. The outputs from these elementary filters—features—are then used as inputs for statistical learning systems. The elementary filters, such as edge detectors (Canny, 1986; Spontón and Cardelino, 2015) or texture analyzers (Castellano et al., 2004; Srinivasan and Shobha, 2008), have known implications for users. Although the combinatorial final step, which involves learning, poses an explanatory challenge, SVMs similarly require a "kernel" of elementary analysis for optimization. In contrast, deep learning methods discussed in the following section involve learning across all components, necessitating explanations for each component in addition to the overall decision-making process.

1.3.1.2 Deep learning in artificial neural networks

Deep learning represents an advancement in the methodology of artificial neural networks, building upon the foundational work of multi-layer perceptrons as proposed by (Rosenblatt, 1961) and influenced by neuro-biological modeling efforts by (Rashevsky, 1948; T. H. Abraham, 2004; McCulloch and Pitts, 1943; Hebb, 2005). Initially overshadowed by symbolic AI—refer to the previously mentioned critiques by J. Pearl (Pearl and Mackenzie, 2018; Olazaran, 1996; Minsky and Papert, 2017)—deep learning resurfaced in the 1980s and has demonstrated impressive results since the 2000s. For a more detailed exposition, readers are directed to foundational texts such as (Goodfellow, Bengio, and Courville, 2016).

Each neuron within the network is defined by a function that maps a set of numerical inputs (akin to signals received by dendrites) to a single numerical output (similar to signals transmitted by an axon). This involves applying a non-linear function, known as an activation function, to a linear combination of inputs. The weights of these inputs and the type

of activation function, which in contemporary practice often includes the Rectified Linear Unit (ReLU) function defined as $x \mapsto \max(0, x)$, dictate the neuron's output.

Neurons are systematically organized into layers, mimicking the anatomical layers of the visual system which has significantly influenced image processing research. Each layer takes inputs from its preceding layer and feeds outputs to the subsequent one. Among various possible configurations of inter-layer connections, two prominent types are fully connected layers, where every possible connection to preceding outputs is formed with independent weights, and convolutional layers, where each neuron is connected only to a local cluster of outputs from the preceding layer using shared weights across a defined kernel. This configuration significantly reduces the parameter space as the same weights are used by all neurons calculating the convolution.

Convolutional layers (Lecun et al., 1998; Krizhevsky, Sutskever, and G. E. Hinton, 2012; Ciresan, Meier, and Schmidhuber, 2012; Z. Li et al., 2021) are often designed with multiple output channels, analogous to the three color channels in an image. This setup allows different kernels to detect various local features, such as edges in different orientations. As the network progresses from the input layer to deeper layers, the number of neurons typically decreases while the number of channels increases, though this trend may reverse in certain architectures such as in encoder-decoder models (Baldi, 2012; Ronneberger, Fischer, and Brox, 2015). The resolution reduction across layers is achieved via max pooling, which groups outputs into tiles, each feeding into a single neuron in the subsequent layer that represents the maximum output value of its respective tile.

In practice, the behavior of a network can be simulated numerically to compute the output associated with a given input, effectively calculating the network's input-output function. Deep learning adjusts the parameters of neurons (including their weights) to approximate a desired network output function. This is commonly achieved using gradient-based optimization methods (Kingma and Ba, 2014), which adjusts parameters in a manner that minimizes the cost function—a measure of the difference between actual and desired outputs.

Deep learning is characterized by the use of multiple layers, which necessitates specific techniques to counteract issues like vanishing gradients. Techniques such as skip connections (K. He, X. Zhang, et al., 2015) have been developed to maintain gradient flow across many layers.

Explaining the operation of these networks can sometimes be relatively straightforward for certain components. A network designed for image analysis, consisting of several convolutional layers followed by a number of fully connected layers, mimics the computations performed by radiomic solutions for the initial layers. In these layers, where convolution has been imposed to enforce translational invariance, the operations often resemble those of convolution-based image filters used in radiomic systems. The operation of subsequent layers can be elucidated through visualizations that compare the outputs of one layer with the input image or the previous layer, since all these layers have comparable formats, though generally of different resolutions.

As one moves away from the input layer, explaining the network's functionality becomes more challenging, particularly for the fully connected layers (B. Zhou et al., 2016; Barredo

Arrieta et al., 2020). However, seeking explanations is an integral part of research in deep learning. For instance, akin to neurophysiology of the visual system, one can reconstruct the receptive field of a deep neuron by identifying sub-images that, when presented as inputs, maximize the response of the studied neuron. Recently developed methods for neural networks, such as attention mechanisms, serve both as means to enhance performance and as tools to facilitate the understanding of network operations. These methods allow for a more intuitive interpretation of how neural networks process information, especially in complex architectures.

It is important to recognize that the mathematical tools used to elucidate the functioning of neural mimetics extend beyond merely statistical concepts, often involving approximation theory, nonlinear dynamics, and differential geometry in high-dimensional spaces (Mhaskar, Liao, and Poggio, 2016; Sprott, 2003; Bronstein et al., 2017). This includes, at least heuristically, the use of concepts such as differentiable manifolds and their embeddings. Consequently, the interpretability in the traditional statistical sense may encounter limitations within these advanced domains. There exists a real risk of misrepresenting the inherently inexplicable aspects of deep learning as inherently obscure by specialists from other areas of artificial intelligence. For instance, the sensitivity of a neural mimetic's response to minute variations in input has been cited as a reason to prefer deductive logical systems or probabilistic systems, which do not suffer from the same issues because they maintain clear causality between inputs and outputs[7]. A more detailed discussion could revisit tools demonstrating the importance of topological stability for signs used in radiology. Such an exploration would not only enhance our understanding of the robustness required in medical imaging but could also provide deeper insights into the structural and functional complexities of deep learning models.

1.3.1.3 Federated learning explainability challenges

Federated learning represents a proposed solution designed to enable multiple stakeholders, each possessing unique datasets, to collaboratively develop a model while maintaining the confidentiality of their respective data. This approach is particularly utilized in medical imaging, facilitating the collaborative training of diagnostic models across different healthcare entities within a multi-centric study. Such an arrangement preserves data confidentiality at each entity level, as the sharing of patient data is considered unacceptably risky and fraught with liability issues (Sheller et al., 2020).

In this model, comprehensive patient images and clinical records are not centralized at a single location. Instead, only the essential differential data required for collaborative learning are transmitted from each participating site, coupled with appropriate security measures. This methodology introduces an additional layer of complexity to the challenges of explainability and necessitates targeted research on potentially unique and anticipatory methods. It is noteworthy that the effort to specify, often through modeling the stakeholders before implementing a federated system, positively contributes to elucidating numerous factors that must be considered in explainability studies.

[7] "La Méthode scientifique" France Culture (March 30, 2022), IA: par-delà le bien et le mal? 'accessed:08 July 2024'

1.3.2 Global and local methods

Global methods are designed to illuminate the operational mechanisms of an intelligent system comprehensively, encompassing all conceivable inputs. These approaches often include an examination of the learning algorithms that underpin the system's functionality. Such global insights are particularly pertinent to the dynamics between the providers of intelligent systems and their clients, as well as between the designers and their stakeholders. This wide-ranging perspective facilitates a more holistic understanding of the system's behavior and its foundational principles.

In contrast, *Local* methods focus on explicating the decision-making processes of an intelligent system for individual instances, specifically tailored to particular inputs. Although these methods are primarily localized, their scope typically extends to include a proximal neighborhood within the input space, consistent with the conventional topological definition of "local." This granularity is crucial for meeting legal mandates that require explanations of specific decisions, implemented by the entities that operate these systems. Hence, local methods serve a dual purpose: they provide clarity on the immediate decision-making context and ensure compliance with regulatory frameworks that govern the use of intelligent systems.

1.3.2.1 Model-agnostic (post-hoc) methods

Post-hoc explainability approaches are applied after the model has been trained, focusing on elucidating the behavior of complex and inherently opaque models, such as deep neural networks. These techniques serve to provide insights into the decision-making process of models by detailing the contribution of input features, visualizing influential components, or through model simplifications.

Feature importance methods, such as SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017a) and LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro, Singh, and Guestrin, 2016), assign a quantifiable value to each feature's contribution to the model's output. SHAP decomposes predictions into individual feature contributions, offering comprehensive insights across the dataset, while LIME provides local linear approximations to explain individual predictions, thus elucidating model behavior near specific instances.

In the realm of visualization, techniques like saliency maps and gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017; Chattopadhyay et al., 2018; Fu et al., 2020)highlight areas of input—such as specific pixels in images or words in text—that significantly influence model decisions. This form of visual feedback is particularly useful in domains like computer vision and natural language processing, where it helps stakeholders visualize what the model perceives as crucial for its decisions.

Additionally, decision tree approximations involve creating simpler, interpretable models such as decision trees that mimic the behavior of the complex model, either on a global or local scale. This method, although it reduces the model's fidelity for the sake of clarity, facilitates a more intuitive grasp of the decision-making process.

Counterfactual explanations provide another dimension of post-hoc explainability by suggesting minimal changes to the input that would result in a different output. This "what-if" analysis helps users understand the model's decision boundaries and provides actionable insights on how to achieve desired outcomes (Wachter, B. Mittelstadt, and Russell, 2017).

Despite the utility of post-hoc explainability methods, they are not without challenges. The fidelity of these explanations to the original model's behavior can vary, and simplifications or approximations may not always accurately reflect the model's complex dynamics across all inputs. Moreover, different explanation methods might yield diverging interpretations of the same model behavior, raising concerns about their reliability.

1.3.2.2 Model-specific (ex-ante) methods

Ex-ante explainability, also known as model-specific explainability, refers to the integration of interpretability directly into the architecture of intelligent systems during the design phase, rather than applying explanations after a model is trained.

In the ex-ante approach, the focus is on utilizing inherently interpretable models or components that allow stakeholders to grasp the logic of decisions intuitively. Common choices include linear models, decision trees, or rule-based systems, which are selected for their straightforwardness. The design of such systems involves a deliberate limitation on complexity to facilitate a more straightforward elucidation of how inputs are transformed into outputs (not losing the causal link). This prioritization of explainability can sometimes necessitate compromises in terms of model performance, leading to a fundamental trade-off between model accuracy and transparency.

This trade-off is not just a theoretical concern but a practical challenge observed across various implementations. Empirical studies suggest that simpler, more interpretable models often do not achieve the same level of predictive performance as their more complex counterparts, such as deep neural networks, which excel in tasks requiring the modeling of high-dimensional, nonlinear relationships. However, the imperative for ex ante explainability is driven by the belief that in many applications, the ability to audit, verify, and trust AI decisions outweighs the need for maximal performance (Rudin, 2019).

Recent advances in machine learning have started to challenge the notion that there must always be a compromise between performance and explainability. Techniques such as attention mechanisms, which provide insights into which parts of the data the model focuses on when making decisions, and disentangled representations, which aim to separate the underlying factors of data into distinct components that are individually interpretable, are examples of how high-performing models can also be made accessible (J. Chen, Song, Wainwright, et al., 2018).

Furthermore, the development of hybrid models that combine both interpretable and complex components offers a middle ground. For instance, an ensemble approach where simpler models handle parts of the data space while more complex models are reserved for intricate cases can balance explainability with performance.

Chapter 2

Interpretable Deep Learning for Cell Segmentation in Video Microscopy

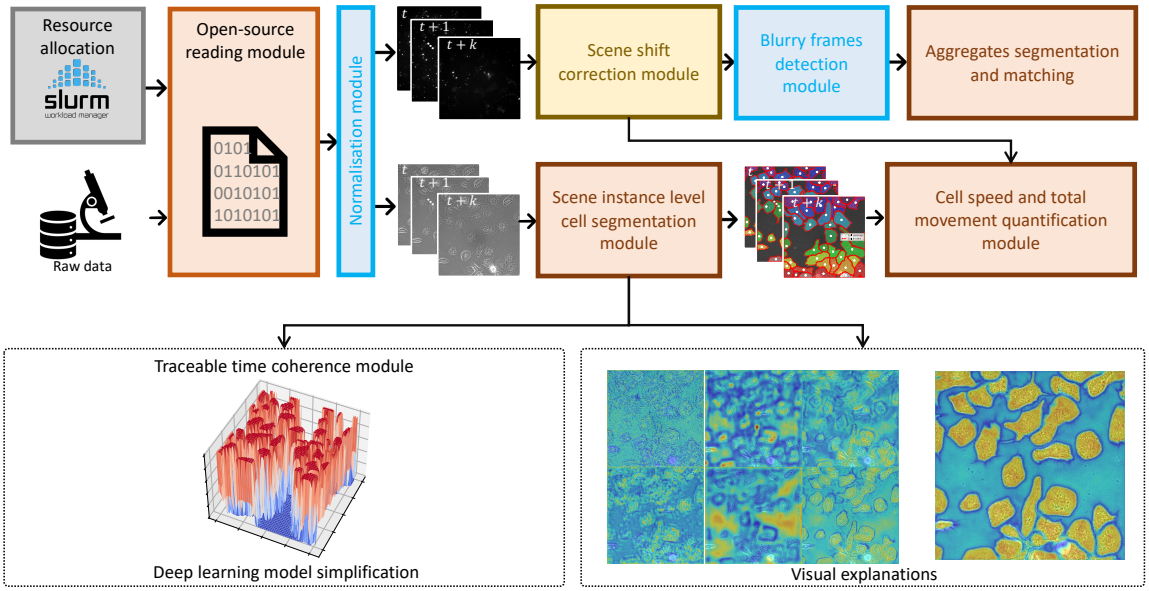


FIGURE 2.1: **PhagoStat: A comprehensive end-to-end pipeline for quantifying microglial cell phagocytosis in the context of frontotemporal dementia (FTD).** The PhagoStat pipeline is a fully operational system comprised of the following stages: (i) efficient loading of raw data (Fig.2.11.b), (ii) applying data quality checks and quantifying aggregates over time (Fig 2.12.c), and (iii) performing cell instance segmentation using an interpretable deep learning (IDL) approach (Fig.2.13, which incorporates Fig.2.3). This comprehensive pipeline streamlines the analysis process and facilitates accurate and reliable results for researchers working with microglial cell phagocytosis data.

Scientific publication

Ounissi, M., Latouche, M. and Racoceanu, D. PhagoStat a scalable and interpretable end to end framework for efficient quantification of cell phagocytosis in neurodegenerative disease studies. Sci Rep 14, 6482 (2024). https://doi.org/10.1038/s41598-024-56081-7

Github: https://github.com/ounissimehdi/PhagoStat

Dataset: https://zenodo.org/records/10803492

Summary

This chapter presents a study on quantifying phagocytosis of dynamic, unstained cells, a critical process in evaluating neurodegenerative diseases such as FTD. Due to the inherent challenges of measuring rapid cell interactions and distinguishing cells from their background in phase-contrast video microscopy, we have developed an end-to-end, interpretable, scalable, and versatile framework capable of real-time analysis of phagocytic activity. Our proposed pipeline processes large datasets efficiently and includes a data quality verification module to address common issues like microscope movements and frame blurring. Furthermore, we introduce an explainable cell segmentation module that enhances the interpretability of deep learning methods, providing a clear advantage over traditional black-box algorithms. This module integrates two key features: visual explanation and model simplification, demonstrating that interpretability does not compromise performance. We apply this innovative pipeline to the study of microglial cell phagocytosis in FTD, revealing that mutant cells exhibit larger sizes and increased aggressiveness compared to controls. Our findings are supported by statistical analyses and have been validated across several public benchmarks, achieving state-of-the-art performance. To foster further research and facilitate translational approaches, we provide an open-source version of our pipeline and a unique dataset of microglial cells phagocytosis.

Résumé

Ce chapitre présente une étude sur la quantification de la phagocytose de cellules dynamiques non colorées, un processus crucial dans l'évaluation des maladies neurodégénératives telles que DFT. En raison des défis inhérents à la mesure des interactions cellulaires rapides et à la distinction des cellules de leur arrière-plan en microscopie vidéo à contraste de phase en accéléré, nous avons développé un cadre complet, interprétable, évolutif et polyvalent capable d'analyser en temps réel l'activité phagocytaire. Notre pipeline proposé traite de grands ensembles de données de manière efficace et inclut un module de vérification de la qualité des données pour résoudre les problèmes courants tels que les mouvements de microscope et le flou des images. De plus, nous introduisons un module de segmentation cellulaire explicable qui améliore l'interprétabilité des méthodes d'apprentissage profond, offrant un avantage clair sur les algorithmes traditionnels de type boîte noire. Ce module intègre deux caractéristiques clés : l'explication visuelle et la simplification du modèle, démontrant que l'interprétabilité ne compromet pas la performance. Nous appliquons ce pipeline innovant à l'étude de la phagocytose des cellules microgliales dans la DFT, révélant que les cellules mutantes présentent des tailles plus grandes et une agressivité accrue par rapport aux témoins. Nos résultats sont étayés par des analyses statistiques et ont été validés à travers plusieurs benchmarks publics, atteignant une performance de pointe. Pour favoriser la recherche ultérieure et faciliter les approches translationnelles, nous fournissons une version open-source de notre pipeline ainsi qu'un ensemble de données unique de phagocytose des cellules microgliales.

2.1 Importance of Interpretability in Segmentation

RECENT advances in high-throughput microscopy and computer-assisted analysis are catalyzing transformative progress in fundamental cellular biology. This evolution, particularly notable in automated tasks such as cell identification, counting, movement tracking, and characteristic profiling, marks a significant departure from previously manual, labor-intensive methodologies (J., Cooper, and Heigwer, 2017; Meijering, Dzyubachyk, and Smal, 2012; Christoph Sommer, 2013).

The implications of these technological advancements extend profoundly across various biological investigations, especially in exploring the dynamics and behaviors of immune cells. The process of phagocytosis, where microglial cells engulf and degrade protein deposits or aggregates, is of particular interest within the context of neurodegenerative diseases (Scheiblich et al., 2021; Janda, Boi, and Carta, 2018; Gentleman, 2013; Q. Li and Haney, 2020; Q. Li and Barres, 2018; Boorboor et al., 2023). A comprehensive understanding of this phenomenon is pivotal for unraveling the intricate mechanisms that underlie such disorders and their progression. Consequently, the demand for precise, quantitative methodologies to advance this field is increasing, as these methodologies provide critical insights into the interactions between microglial cells and protein aggregates, thereby contributing to the development of innovative therapeutic strategies for neurodegenerative conditions.

Traditional imaging processing techniques in microscopy, such as those used to detect cells, often face challenges in accurately detecting unstained cells, measuring rapid cellular interactions, and differentiating cells from complex backgrounds (Buggenthin F., 2013). To overcome these challenges, cutting-edge approaches leveraging advancements in computer vision and deep learning (DL) are necessary (Z. Liu, Jin, and al, 2021; F. Xing et al., 2018).

DL, in particular, has facilitated significant improvements in cell segmentation. Advanced models such as U-Net, Mask R-CNN, DeepLabv3+, Stardist, and Cellpose have been widely implemented across diverse segmentation tasks (Ronneberger, Fischer, and Brox, 2015; K. He, Gkioxari, et al., 2017; L.-C. Chen et al., 2018; Schmidt et al., 2018; Stringer et al., 2021; Arbelle, Cohen, and Raviv, 2022). However, the "black-box" nature of these algorithms poses substantial barriers to their clinical adoption, as transparency in these models is crucial for building trust in their application (van der Velden et al., 2022; Barredo Arrieta et al., 2020).

In response to these challenges and to bridge the gap in the availability of tools, we introduce "PhagoStat", a scalable and interpretable DL-based pipeline designed for analyzing phagocytosis processes. PhagoStat (illustrated in Fig. 2.1) combines the precision of DL with the clarity of XAI, emphasizing interpretability to foster trust and facilitate wider adoption in cell biology research. This integration enhances cellular feature extraction, providing an accessible, comprehensive tool that propels forward our understanding of dynamic cellular processes, especially phagocytosis.

2.2 Enhancing interpretability in deep learning

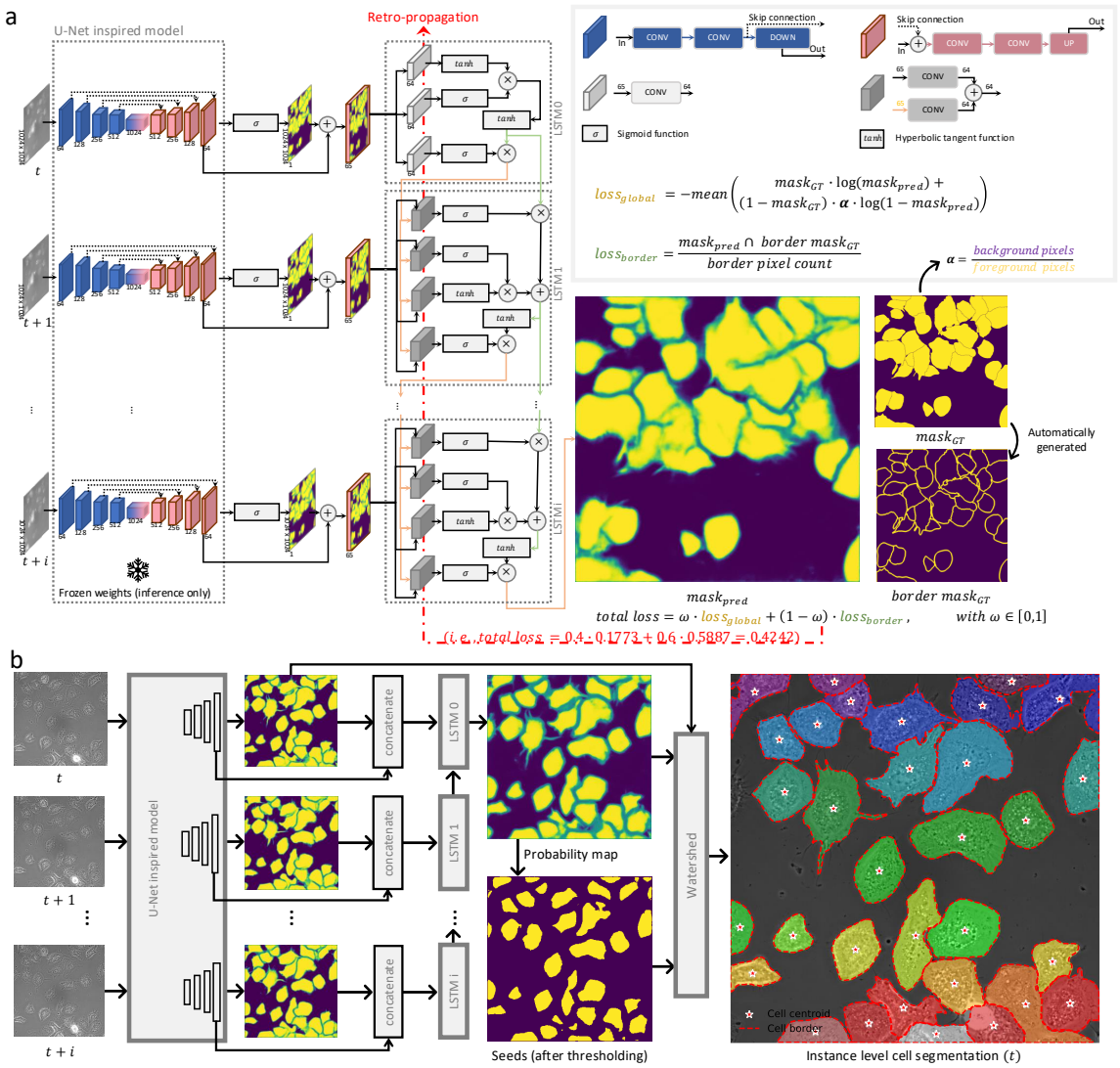


FIGURE 2.2: **Detailed Architectures of Deep Learning (DL) for Cell Instance Segmentation.** This figure provides a comprehensive view of the architectures utilized in DL for precise cell instance segmentation. **(a)** It displays the segmentation module's architecture during the training phase, featuring the application of custom loss functions, both global and local, during backpropagation in LSTM modules to refine learning outcomes. **(b)** It outlines the detailed inference phase that incorporates U-Net-like architectures with LSTM modules, along with a watershed algorithm, to achieve detailed instance-level cell segmentation.

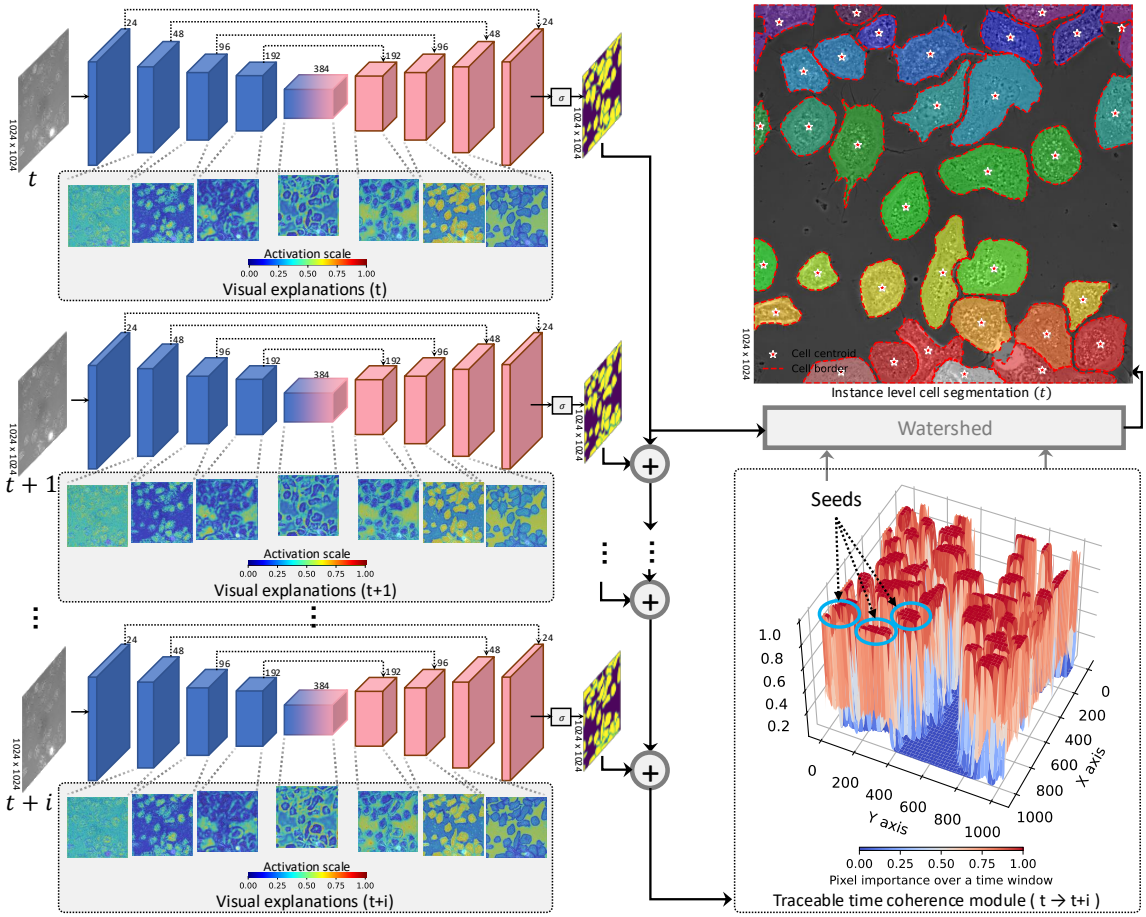


FIGURE 2.3: **Detailed Architectures of Interpretable Deep Learning (IDL) for Cell Instance Segmentation.** This figure provides a comprehensive view of the architectures utilized in IDL for precise cell instance segmentation.It explains the segmentation module, which consists of three major components: (i) streamlined U-Net-like models linked to a visualization module for real-time analysis at each time point, (ii) a time coherence module (TTCM) that efficiently extracts cell seeds, and (iii) a watershed module that integrates all signals for comprehensive cell separation, enhancing the interpretability and accuracy of the segmentation process.

2.2.1 Black-box cellular quantification models

The segmentation of cellular instances in biological images represents an important challenge due to the complex morphologies and dynamic behaviors of cells. This task is critical for a deeper understanding of cellular functions and interactions in various biological contexts. Numerous DL-based methodologies have been developed to address this challenge, employing advanced neural network architectures to discern intricate cellular shapes and structures (Turaga et al., 2010; Ronneberger, Fischer, and Brox, 2015; M. Bai and Urtasun, 2017; Arbelle and Riklin Raviv, 2019; Moen et al., 2019; Schmidt et al., 2018; Stringer et al., 2021). However, most existing techniques do not harness temporal information, which is essential for accurately segmenting cells that exhibit variable morphologies and high motility, such as microglia.

To enhance the precision of instance segmentation, incorporating spatiotemporal information has emerged as a promising strategy. By analyzing cell movement across sequential image frames, this approach can significantly refine the segmentation accuracy (Arbelle,

Cohen, and Raviv, 2022; Liang et al., 2022a). Temporal coherence in segmentation helps in identification and tracking of cells over time, facilitating a more detailed understanding of cellular dynamics and interactions (Liang et al., 2022a).

It is pertinent to note that while contemporary high-performing methods such as TrackMate (Ershov et al., 2022), often deployed as Fiji plugins, are suitable for lighter analytical tasks, they fall short in handling large-scale automated analyses due to their computational inefficiency. This limitation underscores the necessity for more robust and scalable segmentation solutions.

In response to these challenges, we propose a comprehensive framework for robust cell instance segmentation designed to effectively integrate and process extensive datasets. Our approach consists of three integral components: (i) a cell semantic segmentation module that generates precise semantic masks, distinguishing cells from the background; (ii) a time-series coherence module that utilizes information on cellular movement to improve the accuracy of instance segmentation; (iii) a post-processing step that merges semantic and instance data to more accurately delineate cell boundaries.

To evaluate the benefits of integrating temporal data, we implemented a DL-based framework utilizing variants of UNet (including UNet, AttUNet, and BioNet) (Ronneberger, Fischer, and Brox, 2015; Oktay et al., 2018; Xiang et al., 2020) to carry out semantic segmentation, combined with long short-term memory (LSTM) networks (Lindemann et al., 2021) to introduce temporal coherence (as depicted in Fig. 2.2.a). Subsequent post-processing techniques were employed to ensure distinct separation between individual cells, as shown in Fig. 2.3.b.

Our results, illustrated in Fig. 2.4.c, indicate that the integration of DL and temporal data (via AttUNet-LSTM and UNet-LSTM configurations) enhances performance, surpassing that of leading methods like Cellpose and Stardist. However, the BioNet-LSTM configuration exhibited variability in performance due to its recursive architecture which tends to over-fit the feature maps, diminishing its efficacy when paired with temporal modules.

Despite the enhanced capabilities of our DL-based approaches, they inherently lack interpretability—a significant drawback given the increasing emphasis on understanding model decisions in scientific applications. To address this, we embarked on a process to enhance the interpretability of these models, aiming to demonstrate that XAI does not compromise their effectiveness. This initiative seeks to 'whiten the black box' of DL-only methods, showing that it is not only feasible but also competitive with traditional methods (DL-only, Cellpose, Stardist) in performance metrics.

2.2.2 Interpretable cellular quantification models

2.2.2.1 Visual XAI

In this work, we conducted an extensive exploration of visual explainable methodologies in DL, as detailed in Section 1 and by the references such as (Barredo Arrieta et al., 2020; Huff, Weisman, and Jeraj, 2021; Oktay et al., 2018). A focal point of our study was the employment of post-hoc feature visualization techniques, notably through the utilization of heat maps. These maps, which transition in color from red to indicate essential features, to

blue for less critical elements, provided a robust framework for gaining insights into the DL strategies, specifically the DL-only approach. The insights garnered from these visualizations informed potential simplifications of the process, aimed at retaining the effectiveness of cell segmentation methods without compromising their efficiency.

Our empirical investigations included the generation of heat maps for previously unseen images, focusing on dissecting the feature components inherent in the DL-only approach, which integrates U-Nets with LSTM. The analysis revealed significant redundancy in the feature maps produced by U-Nets, suggesting an over-parameterization in the default model configuration. Conversely, the LSTM components were adept at capturing the less mobile internal structures of cells, leveraging temporal dynamics to refine the cell boundaries, thus facilitating enhanced segmentation.

Our method diverged from traditional approaches by optimizing the model quantitatively rather than relying on annotated test sets, as is common with established methods such as nnU-Net (Isensee et al., 2021) and NAS (Y. Zhu and Meijering, 2021). This quantitative evaluation of model efficiency in the feature space is particularly advantageous in scenarios where unbiased annotations are challenging to procure. Through meticulous monitoring of performance metrics, we were able to fine-tune the trainable parameters of the U-Nets, achieving a substantial reduction in model complexity as evidenced by a seven-fold decrease in model size (documented in Fig. 2.4.e).

2.2.2.2 XAI by model simplification

In response to the insights gained from analyzing LSTM strategies, we developed an innovative image processing algorithm termed the Traceable Time Coherence Module (TTCM). This module evaluates a temporal window of cell mask predictions, assigning probabilities to cell parts based on their motion dynamics, thereby enhancing cell segmentation through temporal analysis similar to that of the LSTM (as detailed in Fig. 2.2.b and Fig. 2.3).

This integrative approach led to the development of more compact versions of U-Nets, designated as U-Nets(XAI). These versions were optimized through guided visual interpretation, incorporating a visual explanation module designed to enhance the trustworthiness of the segmentation process, particularly for end users such as biologists and clinicians, by clearly delineating the intermediate steps in mask generation (see Fig. 2.3).

Furthermore, the TTCM demonstrated several advantages over traditional LSTM-based methods, including reduced hardware requirements and the absence of a training phase, alongside offering adjustable parameters for the temporal window, which can be tailored to the specific dynamics of the cellular activity being studied.

2.2.3 Feature-relevance-based automated optimization of DL models

In the quest for efficient deep learning architectures, the choice of model configuration significantly impacts both performance and computational overhead. In this work, we employed the U-Net architecture with its default configuration as a primary reference framework. This choice was guided by U-Net's proven efficacy in various image segmentation tasks, where it employs approximately 30 million trainable parameters in its canonical form.

Our methodology hinged on employing objective metrics such as Mean Squared Error (MSE) to evaluate the performance of the model. We initiated our experiments by generating average feature maps from a dataset devoid of annotations. Each image in this set was processed by both the reference U-Net model and its compact variants. The latter were designed with reduced convolutional layers yet retained the essential architectural stages of max-pooling and up-sampling, similar to the original model. These compact models displayed a wide range of trainable parameters, from as few as 49K to 12.5 million.

To ensure the consistency and reproducibility of our results, we maintained uniform experimental conditions across all models. This included the use of a consistent random seed, identical splits for training and validation datasets, and a uniform number of training epochs. This methodical approach allowed for a controlled comparison of model performances under standardized conditions.

During the evaluation phase, we conducted a comparative analysis using the test dataset. This involved comparing the output feature maps from the full-sized 30 million parameter reference model against those generated by its scaled-down counterparts. For quantitative assessment, we employed MSE to compute an average score that reflected the fidelity of feature representation relative to the reference model. Additionally, we monitored computational metrics such as inference times and GPU memory usage to evaluate the operational efficiency of each model variant.

For a comprehensive visualization and assessment of these metrics, we plotted the derived data, thereby facilitating an intuitive understanding of the trade-offs involved (refer to Fig.2.9). The composite score, a crucial part of our evaluation, was calculated using the following equation:

$$\text{Composite Score} = \alpha \cdot T_{\text{exec}} + \beta \cdot M_{\text{use}} + \gamma \cdot Q_{\text{FM}} \tag{2.1}$$

where α, β, and γ are coefficients summing to 1, with each ranging from 0 to 1.

These coefficients were used to weight the normalized values of execution time (T_{exec}), memory usage (M_{use}), and feature map signal quality (Q_{FM}) respectively. Normalization was performed on a min-max scale to ensure that all metrics ranged between 0 and 1, with adaptations made to reflect that lower values indicate superior performance.

Our illustrative plots not only demonstrate the non-normalized values of MSE, execution time, and GPU memory but also the composite scores using a weighting scheme where $\alpha = 0.5$, $\beta = 0$, and $\gamma = 0.5$. The highlighted red point in these plots represents the optimal trade-off between execution time and feature map signal quality.

In conclusion, the systematic approach adopted in this study not only identifies the optimal model size but also achieves a balanced optimization of feature map signal quality against execution speed. This automated framework illuminates our optimization strategy and enables users toget model parameters to best meet their specific requirements, thereby harmonizing model complexity, feature quality, memory efficiency, and processing time without the need for annotated test data (optimization performed in the features space).

2.2.4 Black-box versus XAI segmentation models

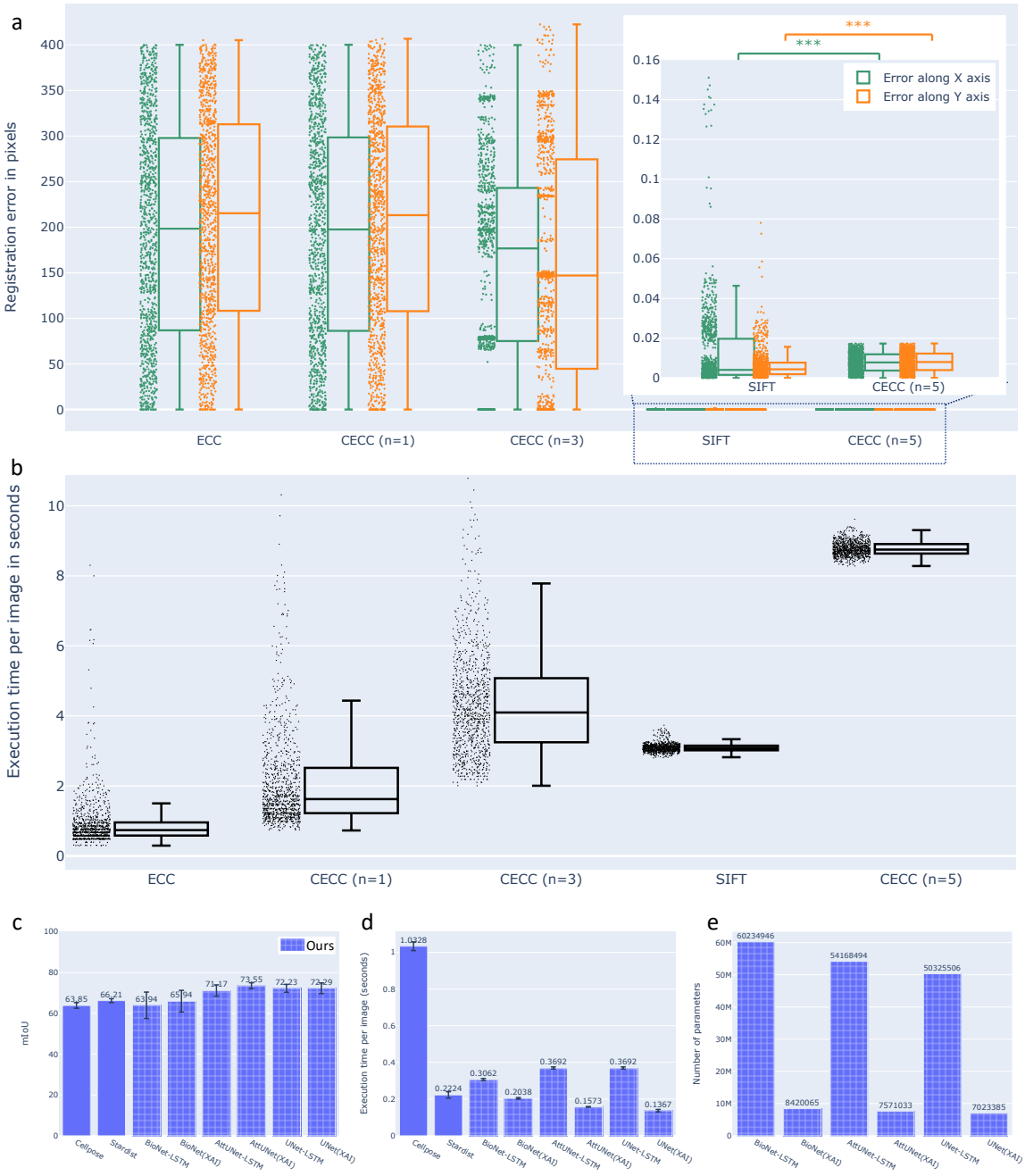


FIGURE 2.4: **Quantitative performance evaluation of the CECC module, DL/IDL cell instance segmentation module. (a)** The performance and **(b)** execution time cost of registration methods ECC, CECC (n=1, 3, 5), and SIFT were evaluated on 1000 randomly shifted frames ($x/y \pm 400px$ shift for $2048^2 px$ frame). CECC (n=5) achieved the best results with an x/y mean error of 0.008 ± 0.004, outperforming SIFT. Our cell detection approach was evaluated against Cellpose and Stardist on a 165-image test set, using a 5-fold cross-validation/testing approach to compute **(c)** mean Intersection over Union (mIoU): sum of IoU of the predicted cell masks divided by the ground-truth cell count; **(d)** the mean execution time cost per image; **(e)** number of parameters for DL and IDL approaches.

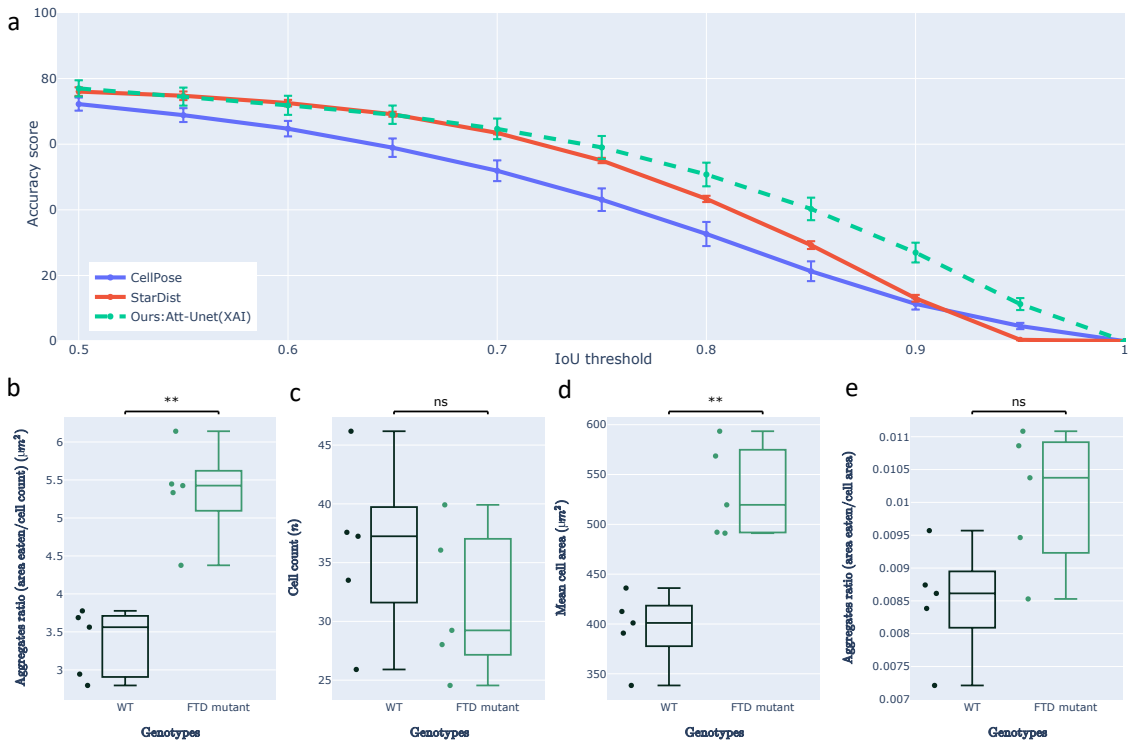


FIGURE 2.5: **Quantitative performance evaluation of the DL/IDL cell instance segmentation module and the phagocytic activity of microglial cells in FTD context. (a)** the accuracy ($0.5 \geq IoU \geq 1$) of our best performing approach 'Att-Unet(XAI)' were computed. Additionally, **(b)** the amount of TDP-43 aggregates internalized per cell; **(c)** the number of cells in the assay: cell count; **(d)** the size of the cells: mean cell area and **(e)** the amount of TDP-43 internalized per cell surface unit. Statistical tests were conducted using the Mann-Whitney-Wilcoxon test with ns (p-value ≥ 0.05), ** (p-value under 0.01), and *** (p-value under 0.001).

In this study, the central focus of our performance evaluation was the mean Intersection over Union (mIoU), a metric obtained by calculating the IoU scores for the predicted versus actual cell masks and subsequently averaging these scores across our dataset. This technique was systematically applied to a test set comprising 165 images, utilizing a five-fold cross-validation. We conducted a comparative analysis with state-of-the-art segmentation techniques, which underscored the effectiveness of both conventional deep learning-only approaches and our enhanced U-Net configurations incorporating XAI "U-Net(XAI)".

It is pertinent to mention, as depicted in Figure 2.4.d, that our evaluation extended to the inference speeds of all models considered. For instance, the Cellpose method involves extensive post-processing, converting vector gradient representations into labeled cell masks, consequently exhibiting reduced performance speed relative to alternative methodologies.

Moreover, the mIoU metric provides a comprehensive means to evaluate cell detection and segmentation quality independently of any threshold requirement. Nevertheless, it is important to distinguish that the literature often describes average precision (AP) differently, where $AP = \frac{TP+TN}{TP+TN+FP+FN}$, with $TN = 0$ (Schmidt et al., 2018; Stringer et al., 2021). We argue that this formulation essentially reflects an accuracy metric, diverging from the conventional AP metric in object detection, which is typically defined as the area under the precision-recall curve. Previous studies have also employed this accuracy definition to evaluate detection quality (Schmidt et al., 2018; Stringer et al., 2021).

Continuing with the evaluative metrics previously employed by methods such as Cellpose and Stardist (Stringer et al., 2021; Schmidt et al., 2018), Figure 2.5.a presents our application of the same accuracy metric to benchmark our top-performing mIoU method –'AttUNet(XAI)'– against state-of-the-art counterparts. By varying the IoU thresholds and computing both mean and standard deviation across the five-fold cross-validation, our approach demonstrated superior performance, particularly for $IoU > 0.8$. This result illustrates that our methodology not only excels in rapid and efficient instance-level cell segmentation but also ensures high-quality outcomes. For a detailed quantitative and qualitative comparison, please refer to the Table 2.1 and Figure 2.6.

We adhered to the default configurations of established methodologies (Cellpose and Stardist) as delineated in their respective publications, employing their publicly available source code. Notably, while Cellpose and Stardist required extensive training durations of 500 and 400 epochs respectively, our IDL approach required merely 20 epochs, and our DL-only approach necessitated 40 epochs. This substantial reduction in training time underscores the efficiency of our proposed methods.

Key insights from our findings are encapsulated as follows:

- As demonstrated in Figure 2.4.c, the integration of temporal information significantly enhances the performance of segmentation methodologies without compromising interpretability.

- Figures 2.3, 2.4.d and 2.4.e validate that enhanced interpretability (visual explanation and my model simplification) can be leveraged to optimize processing speed and diminish hardware requirements (segmentation performance and execution time).

	mIoU(%)	F1(%)	Accuracy(%)	Precision(%)	Recall(%)	Dice(%)	Epochs	Inference(sec)
Cellpose	63.85 ± 1.39	83.35 ± 1.41	72.22 ± 1.99	91.30 ± 1.75	77.09 ± 1.24	84.51 ± 1.16	500	1.032 ± 0.0238
Stardist	66.21 ± 0.96	85.82 ± 0.93	76.04 ± 1.31	**94.80 ± 0.15**	78.95 ± 1.47	87.86 ± 0.52	400	0.222 ± 0.0173
Att-UNet+LSTM	71.17 ± 2.77	86.12 ± 2.52	76.73 ± 3.54	92.20 ± 3.20	81.33 ± 2.72	91.87 ± 2.6	40	0.369 ± 0.0062
Att-UNet (XAI)	**73.55 ± 1.41**	86.53 ± 1.64	77.00 ± 2.47	89.00 ± 2.63	**84.53 ± 1.55**	93.77 ± 0.34	**20**	0.157 ± 0.0021
UNet+LSTM	72.23 ± 1.95	**86.90 ± 1.80**	**77.72 ± 2.53**	93.47 ± 1.61	81.66 ± 2.20	**94.04 ± 0.28**	40	0.369 ± 0.0062
UNet (XAI)	72.29 ± 2.6	85.44 ± 2.12	75.72 ± 2.92	87.95 ± 2.87	83.56 ± 2.46	93.18 ± 1.18	**20**	**0.136 ± 0.0069**
BiONet+LSTM	63.94 ± 6.48	80.25 ± 5.41	68.44 ± 7.01	90.83 ± 2.88	72.73 ± 7.44	92.2 ± 2.78	40	0.306 ± 0.0062
BiONet (XAI)	65.94 ± 5.31	81.43 ± 4.13	70.09 ± 5.33	88.67 ± 3.67	75.99 ± 5.95	91.24 ± 2.66	**20**	0.203 ± 0.0042

TABLE 2.1: **Five-fold Testing: Quantitative Performance Evaluation of the Cell Segmentation Module (DL/IDL) Against State-of-the-Art Methods.** The evaluation results presented are based on five-fold testing and are expressed as (mean ± standard deviation). We highlight the best metrics per column in bold, with the second-best metrics underlined. For performance assessment, instance-level segmentation (detection) evaluations were used, applying various metrics specific to each cell mask. The Mean Intersection over Union (mIoU) is calculated by dividing the sum of IoU for each predicted cell mask by the total number of ground-truth cell counts. To determine these metrics, an IoU threshold of $\geq 50\%$ between the ground truth and predicted masks was used to compute True Positives (TP), False Positives (FP), and False Negatives (FN). The F1 score is defined as $F1 = \frac{2TP}{2TP+FP+FN}$, while the Accuracy is Accuracy $= \frac{TP}{TP+FP+FN}$, Precision as Precision $= \frac{TP}{TP+FP}$, and Recall as Recall $= \frac{TP}{TP+FN}$. Additionally, we used the Dice coefficient to quantify pixel-wise separation between foreground and background. This semantic segmentation metric is defined as Dice $= \frac{2|gt \cap pred|}{|gt|+|pred|}$, where gt represents the ground truth mask, and $pred$ the predicted mask (with background as 0 and foreground as 1). Training epochs were noted as the number of cycles required to complete the training phase. The inference time per image on the test set was measured using an 8-core i7 9700K CPU, 16GB RAM, and an NVIDIA MSI 2080 GPU.

.

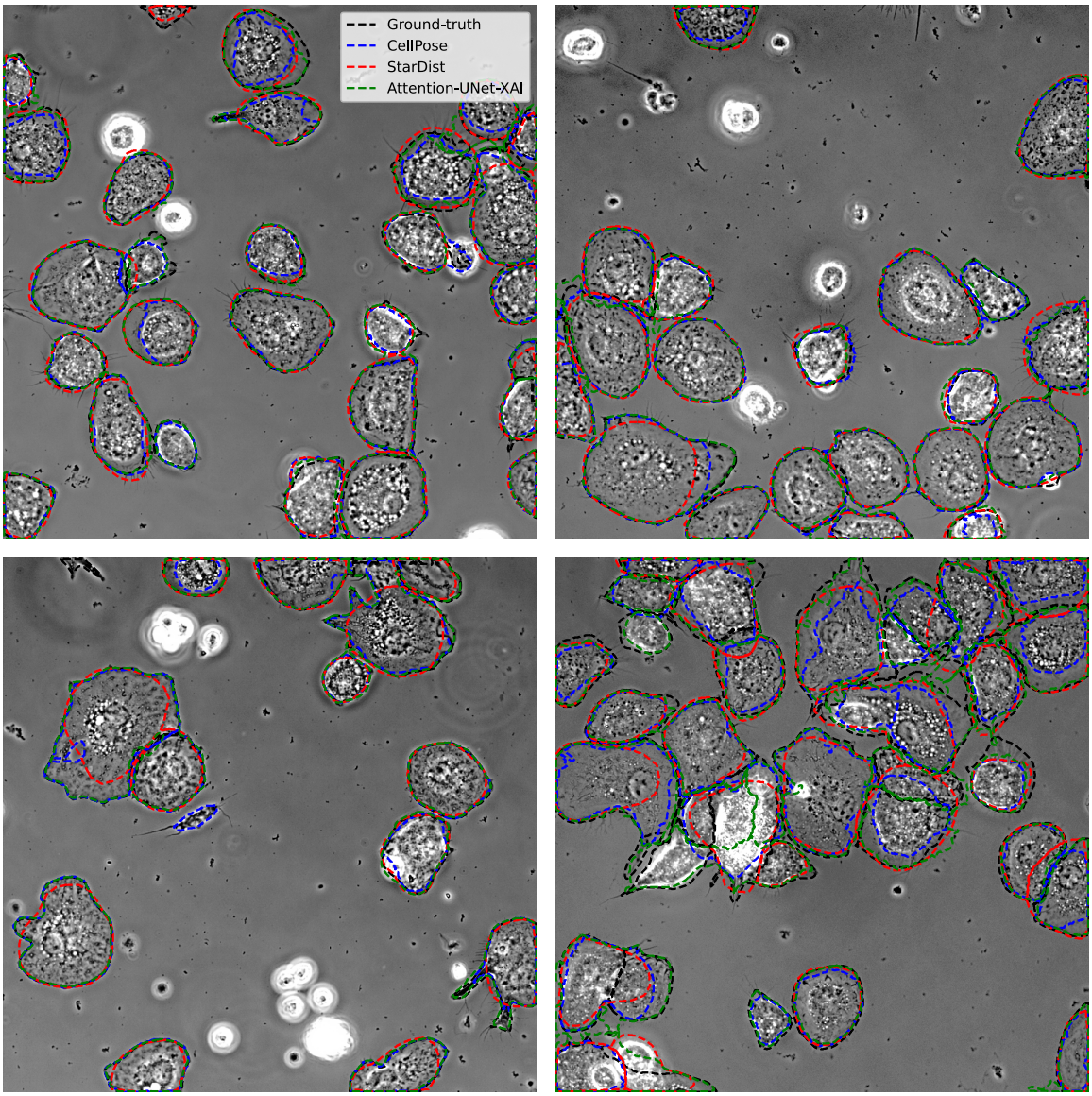


FIGURE 2.6: **Instance-level Cell Segmentation Evaluation:** Through qualitative analysis, the Attention-UNet(XAI) model demonstrates superior performance in comparison to Cellpose and Stardist, especially in addressing the complex shapes of cells. This underscores our model's robust adaptability to the varied morphologies of cells, positioning it as a viable contender against current leading methods. However, it is important to note challenges persist in scenarios where cells form dense clusters or remain in suspension, such as the depicted white cell cluster at the bottom right. In these cases, our model, along with others, faces difficulties in precise segmentation, indicating the necessity for ongoing enhancements to tackle such intricate conditions effectively.

2.2.5 Assessing XAI segmentation generalization capability:

To effectively benchmark our approach against both contemporary and forthcoming methodologies in the field of computational biology, we trained and subsequently evaluated the AttUnet(XAI) model using the comprehensive Cell Tracking Challenge (CTC) datasets. These include Fluo-N2DL-HeLa, Fluo-N2DH-GOWT1, DIC-C2DH-HeLa, Fluo-N2DH-SIM+, and PhC-C2DH-U373, focusing specifically on segmentation tasks as outlined in Table 2.2. The training regimen utilized both available sequences and their respective ground truth masks,

following an 80-20 split for training and validation, conducted over up to 200 epochs. Evaluation metrics, implemented by the CTC organizers (Maška et al., 2023), included the Normalized Acyclic Oriented Graph Matching (AOGM-D) measure (Matula et al., 2015) for detection (DET), the Jaccard similarity index for segmentation (SEG), and an overall performance metric $OP_{CSB} = 0.5(DET + SEG)$.

The quantitative assessment by CTC organizers on hidden test datasets highlighted that the AttUnet(XAI) model achieved a peak performance of 95% in the OP_{CSB} metric for PhC-C2DH-U373, with performance across the datasets ranging between 84.3% and 95%. Segmentation accuracy, denoted by the SEG metric, varied from 72.2% to 91.7%, while detection accuracy (DET) ranged from 88.5% to 98.3%. These outcomes, as provided by the challenge organizers (Maška et al., 2023), underscore the model's robust performance in critical segmentation and detection tasks, reaffirming its utility in automated cellular imagery analysis (as detailed in Table 2.2).

Despite the impressive benchmarks set by leading methods in the CTC, such as CALT-US (Guerrero Peña et al., 2020), ND-US (Liang et al., 2022b), and UNSW-AU (Y. Zhu and Meijering, 2021), their restricted availability for non-commercial, academic research highlights the practical significance of our model. Although it does not surpass these leading methods, our model remains accessible for research and amenable to method-specific refinements. Potential enhancements include the introduction of a third class for cell borders, dataset-specific hyperparameter adjustments (as suggested by nnU-Net "Isensee et al., 2021"), innovative data augmentation techniques (such as random sequence reversals, random affine, and elastic transformations detailed by NAS "Y. Zhu and Meijering, 2021"), and the integration of manual annotations (as performed by BGU-IL "Maška et al., 2023").

Moreover, our model distinguishes itself through its efficiency, utilizing U-Net architectures but at a notably reduced computational cost. The file size of our model, a mere 28.2MB, is significantly smaller—11 times less than that of nnU-Net and at least five times smaller than that utilized by BGU-IL for 2D datasets. This compactness enables the Phagostat method to serve as a baseline in automated cellular imagery analysis, particularly when considering performance relative to model size. By leveraging the AttUNet(XAI) model's compact computational footprint, Phagostat facilitates rapid and precise cell segmentation analyses, ideal for workloads in hardware-constrained environments and underscoring its broad applicability across various research scenarios.

2.2.6 Methodology limitations

In our methodology, we employ a coherence score to measure the continuity of a cell's visibility and tracking across successive frames in a time-lapse sequence. This score is particularly pivotal when a cell first enters the frame at n, prompting an assessment of its trajectory over subsequent frames within a predefined time window (until frame $n +$ time window). This evaluation is critical for determining the cell instance mask at frame n. Nevertheless, the coherence score may not always accurately reflect cell dynamics due to exceptional boundary cases:

- Cells that appear within the field of view late in the sequence—specifically after the final frame minus the time window (last frame − time window)—might receive an

	Cell tracking challenge testing datasets				
CTC metrics	PhC-C2DH-U373	DIC-C2DH-HeLa	Fluo-N2DL-HeLa	Fluo-N2DH-GOWT1	Fluo-N2DH-SIM+
$OP_{CSB}(\%)$	95 (CALT-US:96.1)	84.3 (CALT-US:92.6)	86.4 (BFR-GE:95.7)	88.1 (KTH-SE:95.2)	84.3 (KIT-GE:90.5)
$SEG(\%)$	91.7 (CALT-US:93.1)	80.1 (CALT-US:88.7)	78.8 (MU-US:92.3)	85.2 (CSU-CN:93.8)	72.2 (DKFZ-GE:83.2)
$DET(\%)$	98.3 (CALT-US:99.0)	88.5 (CALT-US:97.5)	94.1 (KIT-GE:99.4)	91.1 (TUG-AT:98.0)	96.5 (FR-GE:98.1)

TABLE 2.2: **Quantitative Performance Evaluation of the Cell Segmentation IDL Module on Cell Tracking Challenge Test Datasets (Maška et al., 2023).** This table documents the performance evaluation of the AttUnet(XAI) in segmenting cells on test datasets from the Cell Tracking Challenge, as detailed by (Maška et al., 2023). Performance metrics, calculated by the organizers following the submission of our results (where no ground truth data was available), include the OP_{CSB} score, defined as $0.5 \times (DET + SEG)$. The detection metric (DET) is derived from the normalized Acyclic Oriented Graph Matching (AOGM-D) metric, as detailed in (Matula et al., 2015). The segmentation accuracy (SEG) uses the Jaccard index, calculated as $J(S,R) = \frac{|R \cap S|}{|R \cup S|}$, where R represents the reference object pixels and S represents the segmented object pixels. A match between R and S is confirmed if the intersection is greater than half of R. This caption also mentions the highest-performing approaches by dataset and metric as reported on the CTC website (Maška et al., 2023).

erroneously low coherence score. To mitigate this, we extend our recording duration by an additional 30 minutes beyond the essential 7 hours, ensuring a total of 7 hours and 30 minutes of data to maintain a margin of safety.

- Furthermore, cells that frequently enter and exit the field of view, a phenomenon we describe as 'field of view border kill', may also adversely affect the coherence score.

Additionally, it's important to recognize that dead cells, which remain stationary, might display high coherence scores, not due to active participation in the observed processes but due to their unchanging position. Our current methodology does not adjust the coherence score for such immobility, a decision supported by the lack of qualitative detection of dead cells by neurologists in our studies. However, this methodology could be adapted to include parameters that penalize the coherence score for lack of movement. Although not essential for our research, this refinement could enhance the accuracy of cell viability assessments in other studies where distinguishing between live and non-viable cells is crucial.

2.3 Exploring Practical Use Cases of Explainability

2.3.1 Global explanation

2.3.1.1 Heat-map visualizations for trust-enhancement

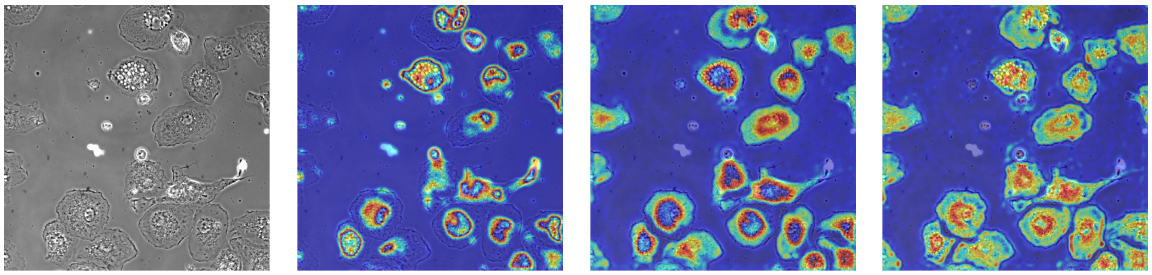


FIGURE 2.7: **Progressive Learning Visualization in AttUNet Deep Learning Model Training.** This figure qualitatively illustrates the key stages in the learning process of our UNet-based deep learning model, as depicted through mean feature map heatmaps. These heatmaps are crucial in demonstrating the model's evolving focus throughout its training. Initially, at the 10 iteration mark, the model begins to recognize cell textures, effectively distinguishing cells from the background. By 300 iterations, it further refines its capabilities, honing in on intracellular components and delineating cell boundaries and nuclei. At 800 iterations, the model displays advanced recognition abilities, identifying cells with partial visibility and precisely differentiating between individual cells. These visualizations play a vital role in building trust with neuroscientists by providing transparent insights (refer to Section 2.3.1) into the model's dynamic learning process.

In the realm of neuroscience, heatmaps serve as a pivotal visualization tool, offering an intuitive depiction of feature intensity and spatial distribution. This aligns closely with the underlying principles of neural activity and the significance of specific features. Heatmaps excel in presenting stark visual contrasts, which significantly enhance the ease of detecting underlying patterns and facilitate rapid pattern recognition. Such capabilities are crucial

in the context of neuroscience where quick and accurate interpretation of complex data is essential.

The widespread adoption of heatmaps in bioinformatics and neuroimaging underscores their established status within the scientific community. This familiarity likely contributes to a reduced cognitive load for domain experts, enabling more efficient interpretation of these visual aids. The computational efficiency of heatmaps also plays a vital role during the iterative processes of model development and inference in research, as outlined in this work.

Recent developments in deep learning visualization techniques, such as Class Activation Mapping (CAM) methods (B. Zhou et al., 2016), gradient-based approaches (Selvaraju et al., 2016; Chattopadhyay et al., 2017), and dimensionality reduction via t-SNE (Maaten and G. Hinton, 2008), alongside feature relevance assessments using SHAP values (Lundberg and Lee, 2017b; Lundberg, Erion, et al., 2020; Mitchell, Frank, and Holmes, 2020; Lundberg, Nair, et al., 2018), have proven effective in classification tasks. However, these techniques often compromise spatial information during transitions through fully connected layers. Our research focuses on dense binary segmentation tasks where maintaining pixel-level spatial detail is imperative. By ensuring spatial fidelity through network operations –including max pooling and up-sampling– our methodology not only preserves spatial integrity but also emerges as a cost-effective solution for pixel-wise precision tasks.

The application of heatmaps during the training phase of our DL model provides a vivid illustration of the model's learning trajectory, thus building trust among biologists and neuroscientists. We have meticulously documented the model's progression from its initial stages of identifying cell textures to its advanced capabilities in recognizing intracellular elements and individual cells, even under conditions of partial visibility (refer to Fig.2.7). This methodical enhancement, captured through heatmaps, furnishes domain experts with a palpable insight into the model's feature learning process, aligning with their empirical and theoretical frameworks.

Heatmap visualizations also serve as a crucial intermediary, simplifying the complexities of deep learning and presenting clear, visual steps of progression that domain experts can readily understand. These visual tools are indispensable not only for elucidating the underlying logic of the model's decisions but also for guiding subsequent improvements in the model's architecture, as detailed in Section 2.2.3.

Engagement with neuroscientists, facilitated by these visual tools, has refined our approach, highlighting the practical significance of heatmaps in bridging various disciplinary perspectives. Although a formal user study is not yet performed, initial feedback indicates that this visualization technique markedly enhances the acceptance of DL models among biologists and neuroscientists. Future research will aim to quantify the impact of these methodologies and integrate them into a framework that supports real-time expert feedback. This interactive approach will be designed to refine the training protocols of the model and augment its decision-making capabilities in real-time applications, ensuring it aligns with expert knowledge in specialized fields.

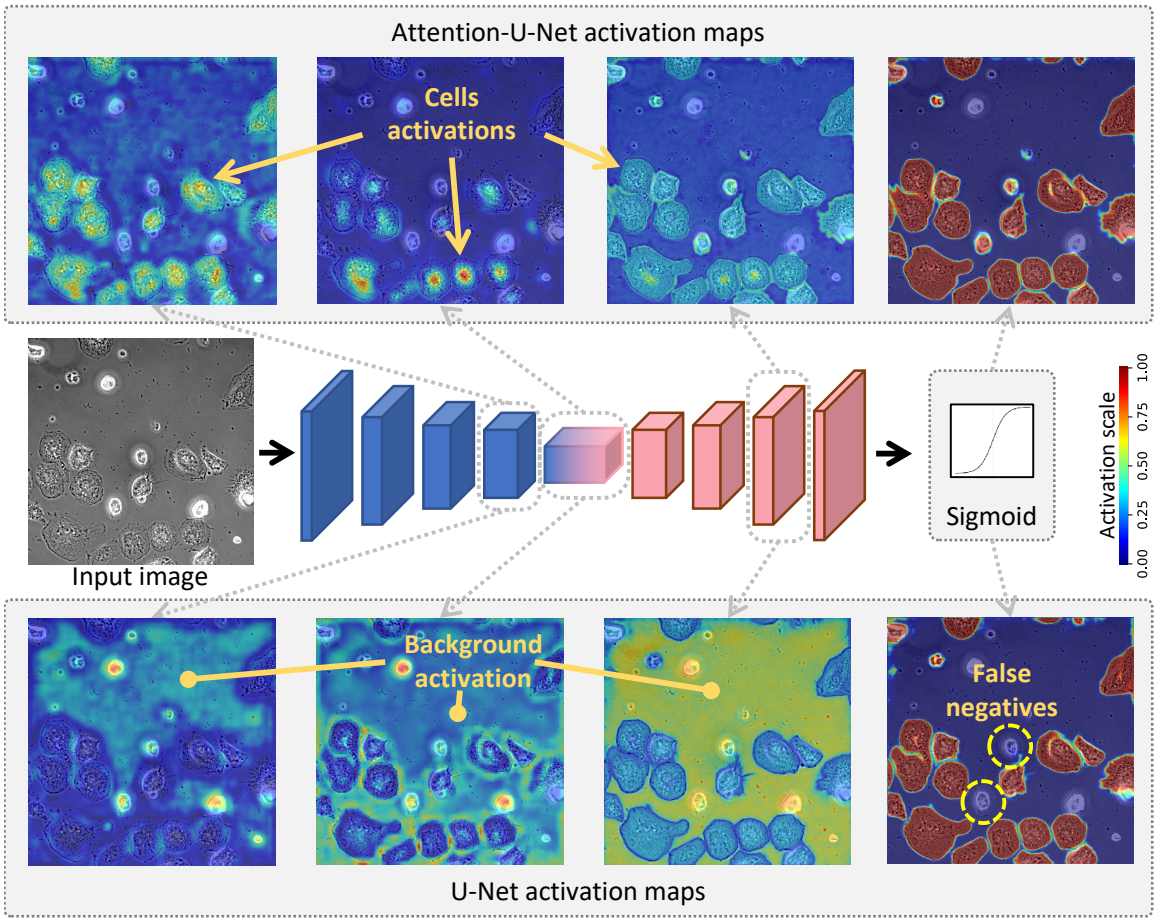


FIGURE 2.8: **Comparative Visualization of Features Learned by U-Net and Attention-U-Net.** This figure illustrates the distinct feature recognition capabilities of U-Net versus Attention-U-Net models. The U-Net model predominantly focuses on background features, as these textures are simpler to model compared to cellular textures. This focus, however, results in a higher incidence of false negatives due to inadequate cellular detail capture. In contrast, the Attention-U-Net employs an attention mechanism that prioritizes the texture of cells, leading to significantly fewer false negatives. This visualization highlights the differences in how each model processes and prioritizes image features, demonstrating the enhanced specificity of Attention-U-Net in identifying critical biological structures.

2.3.1.2 Heat-map visualizations for model comparisons

We noticed intriguing insights during the feature visualization analysis of the UNet and AttUNet models, within the domain of cell segmentation. The standard UNet model primarily harnesses its encoder to extract features from the background, notably at the 'bottleneck' mid-section. This area acts as a critical juncture where the model's architecture begins to focus less on the background, utilizing skip connections within the decoder to enhance foreground feature extraction for mask generation. This approach reflects a strategic simplification, favoring the modeling of foreground texture variability over a more uniform background, as the complexities of cellular textures demand more nuanced processing strategies.

In stark contrast, the AttUNet model introduces a significant modification to this learning dynamic by integrating an attention mechanism. This mechanism refines the model's focus, sharpening its capability to discern and prioritize cellular features more distinctly, as evidenced by the heatmaps presented in Fig. 2.8). This nuanced focus is pivotal, as it

highlights the AttUNet model's enhanced ability to adapt its feature extraction processes to emphasize biologically relevant features within the same training dataset used by the UNet model.

This differential learning strategy between the two models not only delineates their architectural distinctions but also underscores a critical aspect of deep learning in medical imaging: the ability of advanced models to selectively enhance feature recognition and extraction based on specific clinical or biological needs. Consequently, despite both models being trained on identical datasets, the introduction of an attention mechanism within the AttUNet allows it to achieve segmentation results that are not only comparable but potentially more refined in terms of reducing false positives and enhancing the accuracy of feature delineation in complex biological images.

2.3.2 Local explanation

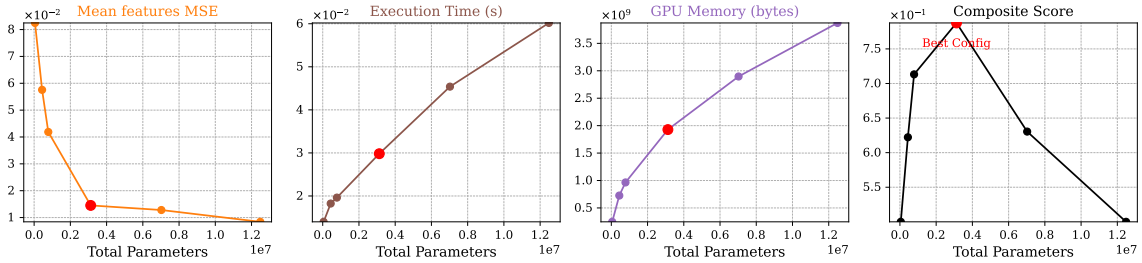


FIGURE 2.9: **Evaluation of Automated Deep Learning Model Optimization Using Feature Maps: Balancing Feature Map Signal Quality and Execution Efficiency in Unet Models.** This figure delineates the comparative analysis of several quantitative metrics across Unet models, including the Mean Squared Error (MSE) of feature map signal quality relative to a 30M-parameter reference Unet model, execution time (in seconds), and GPU memory utilization (in bytes). A composite score is derived using the formula: $\alpha \times \text{time} + \beta \times \text{memory} + \gamma \times \text{MSE}$, where all metrics are min-max normalized to range between 0 and 1. For metrics where a lower value signifies superior performance, the normalized value is adjusted to $1 - \text{metric value}$. The coefficients used are $\alpha = 0.5$, $\beta = 0$, and $\gamma = 0.5$. The red point on the graph identifies the optimal trade-off between execution time and feature map quality, indicating the most efficient parameter settings for the Unet model.

2.3.2.1 Model sensitivity analysis methodology for smart annotation

The findings of this research underscore the critical role of developing visualization tools specifically tailored for direct model ablation studies, focusing on the dynamics of input alterations and their influence on model parameters during training phases, as well as their subsequent effects on output during testing phases. This investigative process is inherently time-consuming, involving extended periods of training, validation, and visualization, thereby posing substantial challenges in real-world applications. However, as detailed in the Section 2.2.3 and depicted in Fig. 2.9, our proposed methodology effectively mitigates these challenges by reducing the model size. This reduction not only enhances training efficiency but also significantly diminishes the risk of overfitting –a pivotal concern when training models with a limited dataset– to ensure the robustness of our findings.

In specialized tasks such as segmentation in biological studies, the selection of appropriate images for annotation is a critical yet often arbitrary process, constrained by the resources at hand. Our methodology facilitates the identification of the most advantageous image types for model training. We validated this approach by evaluating the influence of varying quantities of annotated cell masks on training efficacy, with a particular focus on cell count variation. Our experimental framework employed three images per cell count from our microglial dataset, meticulously chosen from distinct acquisitions to prevent data leakage. We examined 13 different conditions, each with cell counts ranging from 14 to 40 per image, by training separate models for each scenario using the Attunet(XAI) architecture. Remarkably, this entire training process for 13 models on a standard GPU required less than 30 minutes, thus demonstrating the efficiency of our approach in swiftly assessing model performance under diverse conditions.

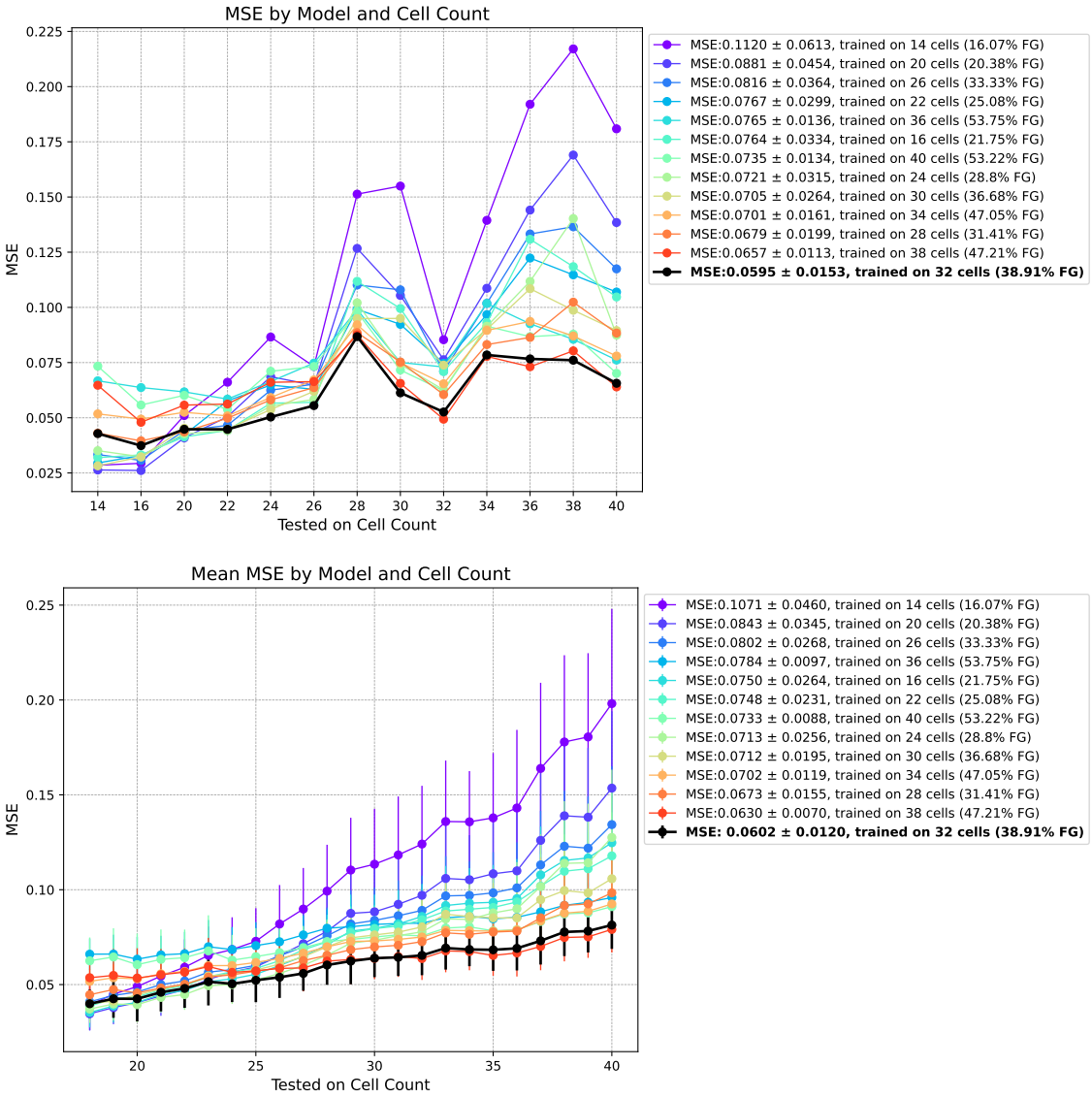


FIGURE 2.10: **Comparative Sensitivity Assessment of AttUnet(XAI) Across Varying Cell Quantities per Condition.** This figure illustrates the outcomes of two distinct test setups aimed at evaluating the performance of the AttUnet(XAI) model. On the top, results from our sensitivity analysis framework are presented, where only three images per cell count were utilized for training, validation, and testing, significantly minimizing data requirements. On the bottom, the graph displays the model's performance using 100 test images per condition across 23 conditions, involving a total of 2300 images with varying cell counts. The Mean Squared Error (MSE), where lower values indicate better performance, was calculated between the model-generated probability maps and the corresponding ground truth binary masks. The top-5 performing models provide practical guidelines, such as the prioritization of annotating images with cell counts between 28 and 38 and a foreground-to-background ratio of 31% to 47%. These results underscore the effectiveness of our sensitivity assessment framework in pinpointing key image characteristics that influence model performance, thereby guiding annotators towards more strategic and efficient processes. This approach facilitates a detailed investigation of the model's behavior under controlled conditions without significant time or computational burdens. Training the 13 distinct cell count models required less than 30 minutes in total on a single 8GB GPU (NVIDIA RTX 2080).

Our results offer valuable insights for annotators, as illustrated in Fig. 2.10.top, by identifying and prioritizing specific images that significantly enhance model performance.

Notably, the top-5 performing models revealed that images featuring cell counts between 28 and 38 with a foreground-to-background ratio of 31% to 47% yielded optimal training results. To validate these findings, we extended our testing to encompass 2300 images, as shown in Fig. 2.10.bottom, which substantiated the initial recommendations derived from a considerably smaller dataset. This extensive testing emphasizes the model's sensitivity to particular parameters, such as cell count, and underscores the effectiveness of our method in evaluating model behavior with minimal annotation and training inputs.

This streamlined and efficient approach for model evaluation and annotation strategy optimization in biological segmentation tasks. By requiring minimal resources and time, our methodology not only assists biologists in intelligent annotation but also establishes a practical framework for exploring model behavior in controlled experimental settings. To facilitate the application of our results, we have made the source code available in publicly accessible notebooks on Github, complete with visualization tools for immediate analysis, thereby simplifying the integration process across various experimental conditions.

2.4 PhagoStat pipeline components

The pipeline presented in this study is composed of several interconnected modules, each designed to carry out specific functions (refer to Appendix A for the implementation details).

Initially, the data-efficient loading and normalization module, detailed in Section 2.4.1, optimizes data handling and preprocessing. This module plays a role in minimizing the computational demands, thus facilitating more efficient downstream analysis.

Subsequently, the spatiotemporal frame registration process, covered in Section 2.4.2, includes a rigorous data quality check. Key features of this process involve correcting scene shifts and detecting blurry frames, thereby ensuring the precision and reliability of the data while maintaining consistent spatial and temporal alignment throughout the frames.

Furthermore, detailed in Section 2.2.2, the pipeline features modules for cellular and aggregate quantification that enable detailed analyses of cellular properties and interactions. The first of these employs instance-level interpretable segmentation techniques that efficiently extract and analyze features from unstained cell images, allowing for detailed observations of individual cell behavior and morphology. In parallel, the second module applies image processing techniques to segment and match aggregates from fluorescent images, thus elucidating complex interactions between cells and aggregates and uncovering critical biological and morphological insights.

2.4.1 Data efficient loading and normalization

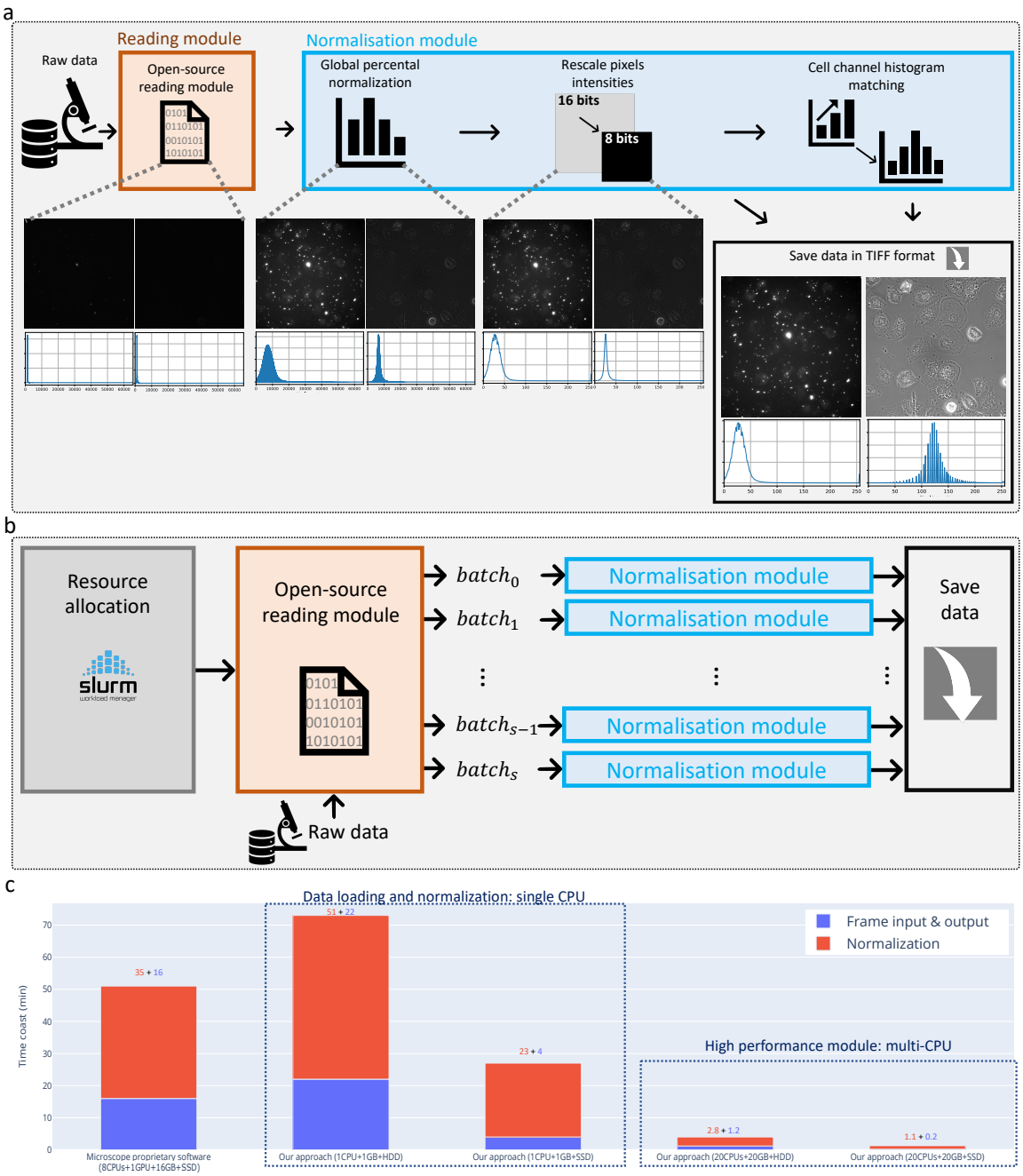


FIGURE 2.11: **Efficient data loading and normalization pipeline**. This pipeline includes: **(a)** A detailed data loading and normalization module which extracts two channels (aggregates and cells) directly from the microscope's raw data and applies both local and global normalization to standardize the data; **(b)** A High Performance Computing (HPC) cluster compatible scheme that efficiently scales to accommodate big datasets; **(c)** A quantitative comparison of our single-CPU/multi-CPU method against the GPU-accelerated Carl Zeiss ZEN software for processing a 76GB CZI file. To ensure a direct comparison, the 'Frame input & output' times encompass both reading and writing operations across all systems. An analysis of time allocation shows that our method assigns 25% for reading and 75% for saving on SSDs, while on HDDs, it allocates 76.6% for reading and 23.3% for saving.

We emphasize that our module is designed with flexibility in mind, allowing for seamless integration with various microscopy systems rather than being tailored to any specific microscope or software. This universality is demonstrated through the implementation of open-source packages. Comparative performance assessments reveal that our module functions at least twice as efficiently as conventional proprietary software and requires only one-eighth the hardware resources, as illustrated in Fig. 2.11.c.

Our comprehensive evaluation shows that the innovative combination of percentile normalization—adapted globally for each sequence—and cumulative histogram distribution matching—tailored locally to each image—significantly reduces data variability. This dual normalization approach enhances the performance of our deep learning segmentation model, leading to an improvement of up to 10% in the Dice coefficient. Additionally, we have incorporated this normalization strategy into the raw data readout module, exploiting its parallel processing capabilities. This integration significantly reduces the time required for multiple processes, namely reading, normalizing, and saving data, thereby offering a more efficient workflow compared to traditional methods where normalization follows data readout, involving steps such as reading raw data, storing it in an open format, reloading it for normalization, and subsequently saving the processed data.

2.4.2 Data quality check and correction

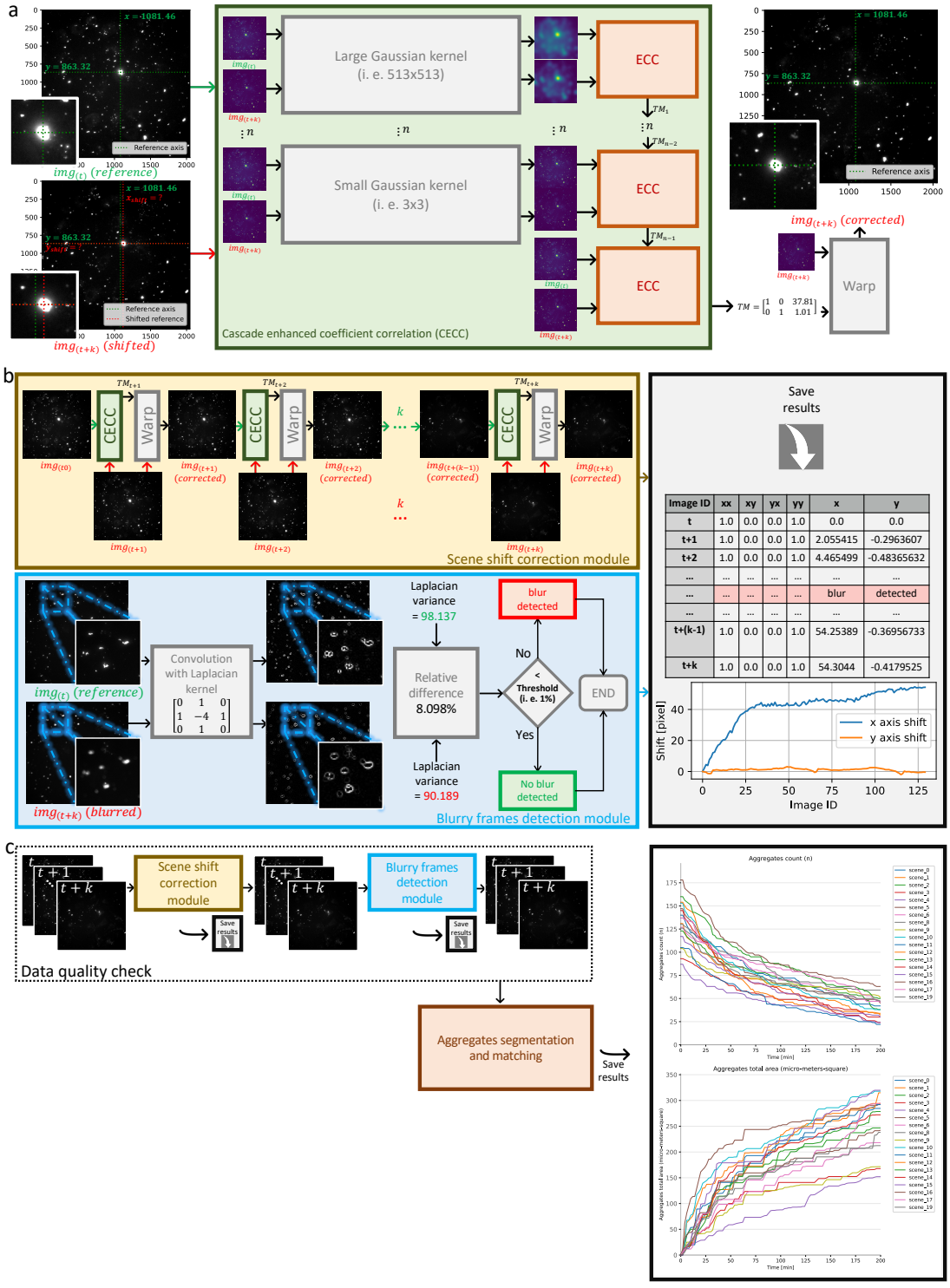


FIGURE 2.12: **Detailed Data Quality Workflow:** (a) CECC Registration Approach: Detailed description of the registration approach based on CECC. (b) Data Quality Check Modules: This includes (i) a CECC-based scene shift correction module for adjusting scene shifts using CECC, (ii) a blurry frames detection module for identifying and tagging blurry frames, and (iii) functionality for saving registration information and the rejected blurry frames. (c) Overview of the Aggregates Quantification Workflow: Combines data quality checks with segmentation and matching procedures to ensure accuracy and completeness.

During our analysis of high-content video-microscopy recordings captured over extended periods, we observed the presence of several unavoidable hardware-related acquisition faults or artifacts. One such artifact was the unintended shaking of the microscope along the x and y axes. These imperfections, although localized to specific frames, had a detrimental impact on the overall sequence quality. Consequently, these artifacts can potentially compromise the accuracy and reliability of subsequent data analysis and interpretation, underscoring the importance of addressing such issues in a systematic manner during the data processing stage.

In our study, we aimed to investigate how external disturbances affect the performance of the microscope camera sensor. To this end, we recorded 20 simultaneous scenes, each containing two channels (cells:non-fluorescent and aggregates:fluorescent), with a frequency of 1 frame every 2 minutes, over a period of 7 hours without any interruption. During the recording process, we estimated that the microscope camera sensor had only 6 seconds (2 min divided by 20 scenes) to cycle and stabilize along the x and y axes from frame n to frame $n+1$ to capture pixel intensities and write them to a local disk.

However, external disturbances, such as mechanical vibrations, lens getting out of focus, can cause the microscope camera sensor to deviate from its normal performance, further reducing the time-response gap. Likely, such external disturbances occur at least once during the 7h non-stop sessions. When this happened, it affected the quality of 1 to 10 consecutive frames.

In order to counteract the potential influence of external disturbances on our analysis, we developed a registration-based module specifically designed to align and stabilize the frames. This module effectively mitigates any potential deviations caused by external factors, ensuring the accuracy and reliability of the subsequent data analysis.

It is important to note that in the aggregate channel, the majority of pixels belong to the background. This predominance of low-intensity pixels (i.e., the background) in comparison to the high-intensity pixels (i.e., the aggregate) presents a challenge when it comes to registration (offset correction).

Given the explainable nature of our pipeline, we opted against using black-box registration approaches based on DL (Sengupta, Gupta, and Biswas, 2022). Instead, we chose to employ the Scale-Invariant Feature Transform (SIFT) algorithm (Lindeberg, 2012) as our registration method. SIFT is not only fully explainable and mature, but it has also become publicly available since the expiration of its patent in March 2020. This combination of explainability and accessibility makes SIFT an ideal choice for our pipeline.

Although the application of SIFT proved sufficient in correcting the shift, as demonstrated in Fig. 2.4.a and in Table.2.3, we observed a statistically significant directional bias in the shift correction. As a result, we concluded that an unbiased approach was necessary to address the registration problem effectively. To this end, we proposed a generalized version of the Enhanced Correlation Coefficient maximization approach (ECC) (Evangelidis and Psarakis, 2008), which we have termed Cascade ECC (CECC), as illustrated in Fig. 2.12.a.

CECC offers an unbiased solution, ensuring consistent registration performance irrespective of the shift direction. Our approach achieved sub-pixel precision for shift margins up to $\pm 20\%$ of the image size (i.e., $\pm 400px$ for $2048 \times 2048px$ frames). To provide context for

		Absolute error		
		along x-axis	along y-axis	Execution time (sec)
	ECC	196.02±119.34	206.93±116.50	**0.92±0.72**
	SIFT	0.0153±0.0609	0.0228±0.1221	3.09 ± 0.12
Ours	**CECC (n=5)**	**0.0079 ± 0.0046**	**0.0081±0.0047**	8.77±0.20

TABLE 2.3: **Performance evaluation of our CECC registration method compared to the state-of-the-art:** We report the results as the mean ± standard deviation, calculated over 1,000 registration tests. Independent random shifts along the x and y axes were generated within a range of ±400 pixels for 2048 × 2048 pixel images. The best metrics per column are bolted, and the second-best metrics are underlined. Absolute error is calculated based on the difference between the estimated registration coordinates and the ground truth, which are the generated shifts along the x and y axes. Registration time cost is determined by the time taken to register a pair of images (reference and shifted). We demonstrate that ECC is ineffective for the specified registration task, and that the SIFT exhibits a directional bias. In contrast, our proposed CECC (n=5) is unbiased and performs significantly better than both approaches. We conducted the evaluation using the following hardware: a 4-core Xeon Gold 6126 CPU and 1GB RAM. For the SIFT method, we used 2GB RAM, as 1GB was insufficient.

these results, the largest unwanted shift observed in our study was approximately 5% of the image size. This demonstrates the robustness of CECC, which offers a 15% margin in the context of the worst-case scenario observed. It is worth noting that there is a significant difference in registration speed between SIFT, with an average of approximately 3 seconds per frame, and CECC (n=5), with an average of approximately 8.7 seconds per frame (refer to Fig. 2.4.b). This discrepancy can be attributed to the maturity and optimization of SIFT, as compared to the proposed CECC. We anticipate that this gap will eventually be reduced through the contributions of the open-source community.

An other issue worth mentioning when we explored the data is that the microscope lens can get getting out of focus because of physical vibration. This introduce some blurry frames into the scene, which unnaturally amplify the size of aggregates, thus, biasing the phagocytic quantification.

We addressed this issue, by including a blur detection module, that uses image processing to detect the loss of details in images. Then discard them from the stack.

To streamline the data quality check process we combining the CECC-based scene shift correction and the blurry frames detection module (see Fig. 2.12.b).

This gives the user full traceability over the data quality, making the definition of objective criteria possible. For example, the maximum tolerated shift along either axis x and y can not be more than 50 pixels at any frame, and the maximum tolerated blurry frames in a given scene can not be more than 5%. The data quality check module plays a crucial role in detecting (i.e., blurry frames), correcting (i.e., scene shift) and objectively quantifying the severity of hardware/human errors in real-world conditions.

2.4.3 Aggregates quantification

To accurately detect and quantify aggregates in time-lapse videos, we developed a specialized module that processes data via a sequence of steps. These steps encompass: (i) employing

a threshold-based segmentation specifically designed for fluorescent aggregates; (ii) generating a binary mask to distinguish individual aggregates; (iii) calculating and recording the count, surface area, and coordinates of aggregates; and (iv) quantifying phagocytized aggregates through changes in surface area and coordinates. Our tests revealed that tracking the reduction in surface area captures two crucial phenomena: the decrement observed as a cell internalizes an aggregate piecemeal, and the coordinate shifts occurring when the aggregate becomes sufficiently small to be engulfed by the cell. This dual measurement approach enables accurate quantification of phagocytosis, defining complete internalization as phagocytosis.

In this study, the presence of fluorescent aggregates was indicated by high pixel intensities, while their absence corresponded to low intensities, allowing for effective segmentation via a simple threshold. Our experimental protocol ensured that aggregates remained stationary in the presence of cells initially, indicating that any subsequent movement was attributable to cellular activity. Further, our imaging highlighted the dynamic interplay between microglia and fluorescent aggregates, which appeared as brightly fluorescent against a dimmer background. During internalization, aggregates were intermittently masked by the cell plasma membrane, creating transient, moving white shadows due to cellular motion. As microglia degraded the aggregates, fluorescence decreased, culminating in complete darkness when fluorescence ceased. Our analytical method effectively differentiated true fluorescence of both static and mobile aggregates from background noise, using particle movement detection to ensure accurate and consistent identification and segmentation of non-internalized aggregates.

The efficacy of this technique is predicated on two primary assumptions: the fluorescence and initial immobility of aggregates in the presence of cells. Thus, observed movements of aggregates are ascribed to cellular actions. This monitoring relies on detecting shifts in the aggregate centroid from its original position, a strategy that may prove unreliable if aggregates were not stationary or if the assumption of fluorescence was invalid.

2.4.4 Spatiotemporal analysis of aggregates and cells

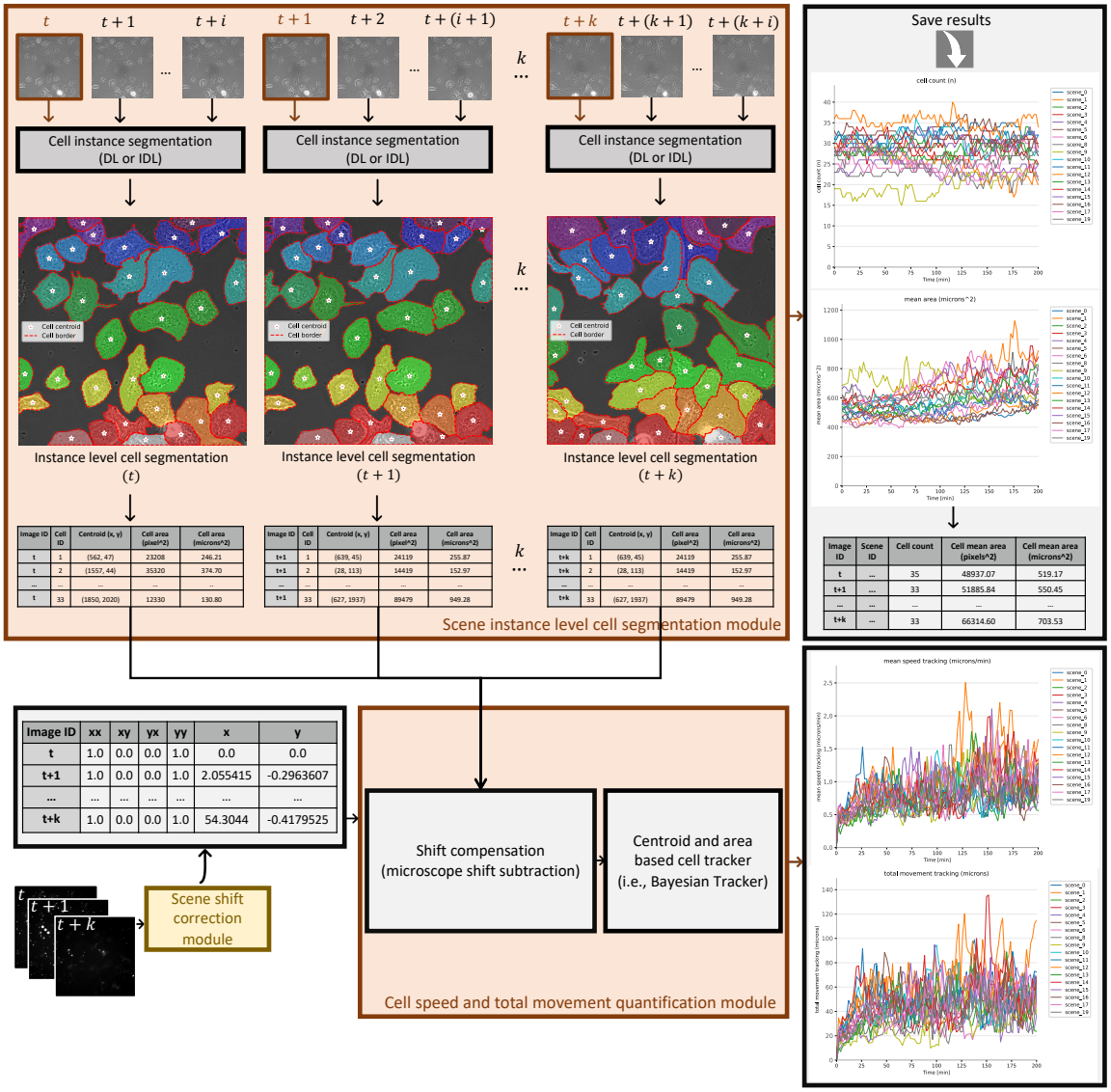


FIGURE 2.13: **Scene cell instance segmentation and tracking.** The scene instance-level segmentation module leverages either the DL module (Fig.2.2.b) or the IDL module (Fig.2.3) to perform scene cell instance segmentation, quantifying cell count, area, and coordinates for each frame. This is further supported by the scene shift correction module (Fig.2.12.b) that adjusts cell centroids, essential for accurate tracking. A tracking algorithm, such as the Bayesian Tracker, is then applied to these corrected features to calculate cell speed and total movement. The integration of these modules allows for the results to be compiled and saved in an open-source CSV format, facilitating data sharing and analysis.

We delineate the components of our phagocytosis quantification pipeline, 'PhagoStat', illustrated in Fig.2.1. The pipeline begins by automatically allocating computational resources, utilizing a high-performance computing cluster for extensive computations as depicted in Fig.2.11.b, or a local machine for less demanding tasks (Fig.2.11.a). Upon loading, the raw data undergoes normalization and division into two distinct channels to separately analyze cells and aggregates. This approach aids in correcting alignment shifts and removing any blurred frames, thereby enhancing the accuracy of subsequent analyses (Fig.2.12.b). The

aggregates are then meticulously segmented to quantify morphological attributes such as area and count (Fig.2.12.c).

Parallelly, the cell data is processed through a scene instance-level segmentation module (Fig.2.13), facilitating the extraction of cellular metrics like area and coordinates. These metrics are crucial for tracking cellular movements and calculating dynamic parameters such as cell speed and total displacement, as described by recent studies in cellular tracking algorithms (Ulicna et al., 2021; Bove et al., 2017). Moreover, the pipeline generates heat map visualizations and time-coherent data traces for enhanced interpretability (Fig.2.3).

The results from various experimental conditions are aggregated to produce a detailed statistical report. It is pertinent to mention that the current version of our statistical module supports analyses under two conditions. For experiments involving more conditions, the reporting framework would require modifications to handle the additional conditions data.

Regarding the pipeline's efficiency, it is designed to process data from 20 scenes –each spanning 7 hours with frame pairs captured each 2 minutes–within approximately 20-30 minutes. Initially evaluated on a CPU-only cluster, the pipeline's performance suggests a potential reduction in processing time when utilizing GPUs, underscoring its capacity for swift and effective data analysis.

2.5 Microglial cells phagocytosis use case

The dual role of microglial phagocytosis, which includes the clearance of protein aggregates as well as the problematic phagocytosis of live neurons and synapses, has been extensively studied in the context of neurodegenerative diseases such as Alzheimer's and Parkinson's (Scheiblich et al., 2021; Janda, Boi, and Carta, 2018; Gentleman, 2013; Q. Li and Haney, 2020; Q. Li and Barres, 2018). In developing our assay, we specifically addressed FTD, highlighting mutations in the genes *C9ORF72* and *GRN*, which are prevalent in familial forms of FTD and are known to modulate microglial functions like phagocytosis (Lui H, 2016; Lall D, 2021; Haukedal H, 2019). The accumulation of TDP-43 protein aggregates in neurons, (Neumann M, 2006; Arai T, 2006) suggests that enhancing aggregate clearance could be beneficial therapeutically. However, the risk of exacerbating neurodegeneration through excessive synaptic pruning or the phagocytosis of live neurons has been shown in cases of *C9ORF72* mutations (Lall D, 2021).

Furthermore, while existing phagocytosis assays employ simple targets like latex beads or pH-sensitive fluorescent particles, there is a significant interest in developing more sophisticated assays. These assays aim to analyze microglial phagocytic activity using physiological targets such as protein aggregates and intact neuronal networks, which is essential for identifying and developing therapeutic compounds targeting abnormal phagocytic activities.

To assess the specific phagocytic activity of both WT and FTD-mutant microglial cells, we quantitatively evaluated the uptake of TDP-43 aggregates per cell, as depicted in Fig. 2.5.b, which presents the ratio of aggregate area internalized to cell count. Remarkably, FTD-mutant cells exhibited a 70% higher rate of phagocytosis compared to their WT counterparts, suggesting an enhanced phagocytic aggression. Despite ensuring an equal cell count in the assays (Fig. 2.5.c), we noted a significant increase in the size of the FTD-mutant

microglial cells (Fig. 2.5.d: mean cell area), with these cells being approximately 30% larger than the WT cells, a novel observation in our studies. Consequently, we further analyzed the amount of TDP-43 internalized relative to the cell surface area (Fig. 2.5.e: aggregates ratio = area eaten/cell area), revealing that the increased cell spread might account for the heightened phagocytic activity in FTD mutants. Importantly, there were no significant changes in the overall mobility or movement speed of the cells, as indicated in Fig. 2.15 and Fig. 2.14.

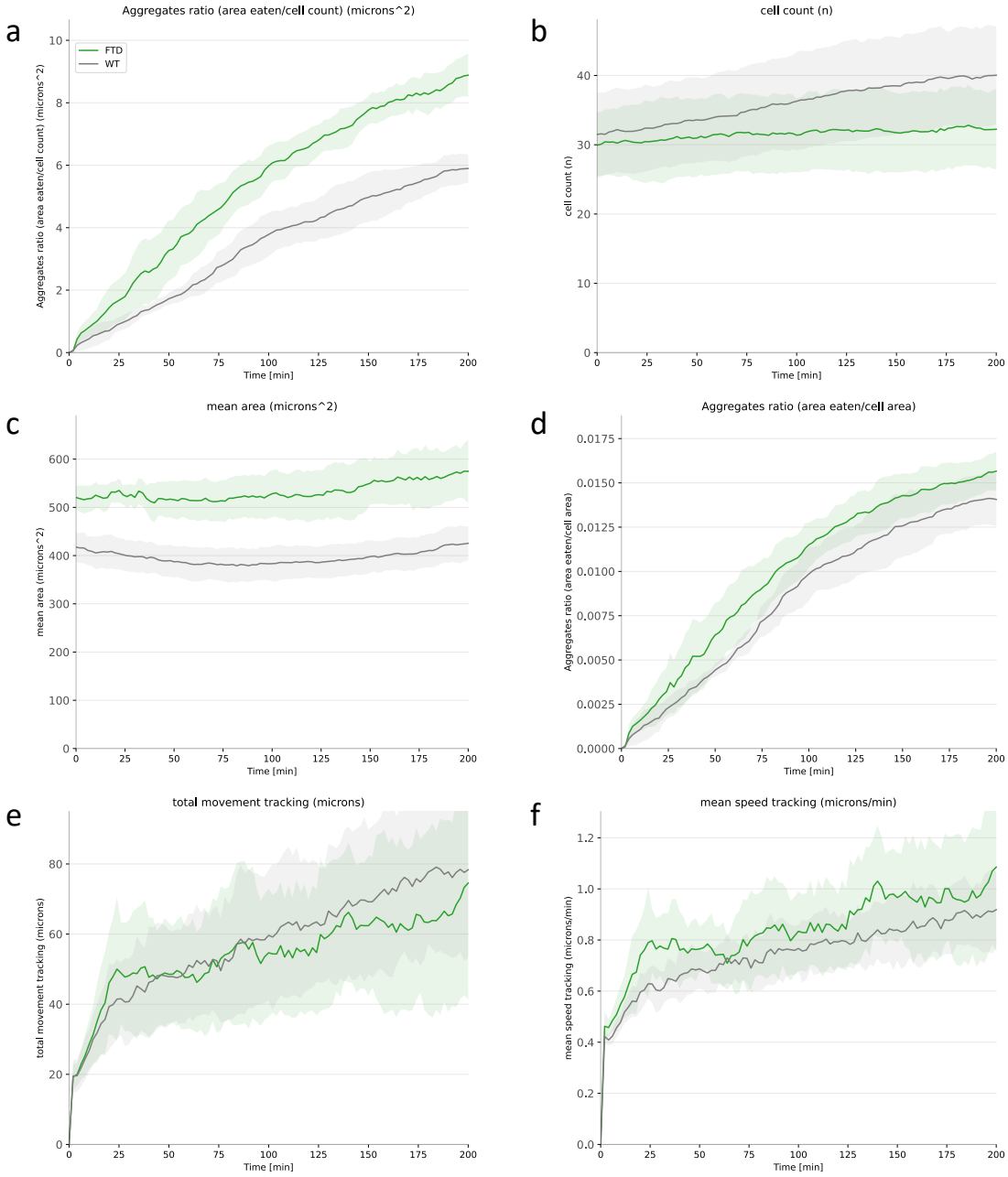


FIGURE 2.14: **Comparative Analysis of Phagocytosis Metrics Over Time for WT and FTD Groups:** This figure offers a detailed comparison of phagocytosis-related metrics between WT and FTD groups, capturing their dynamic differences over time. It includes a series of panels illustrating various parameters: (a) the aggregate area consumed by cells, (b) cell count, (c) mean cell area, (d) cell surface area consumption, (e) total cell movement, and (f) cell speed. Through this comparative analysis, the figure facilitates a comprehensive understanding of the distinct phagocytic behaviors characterizing each group.

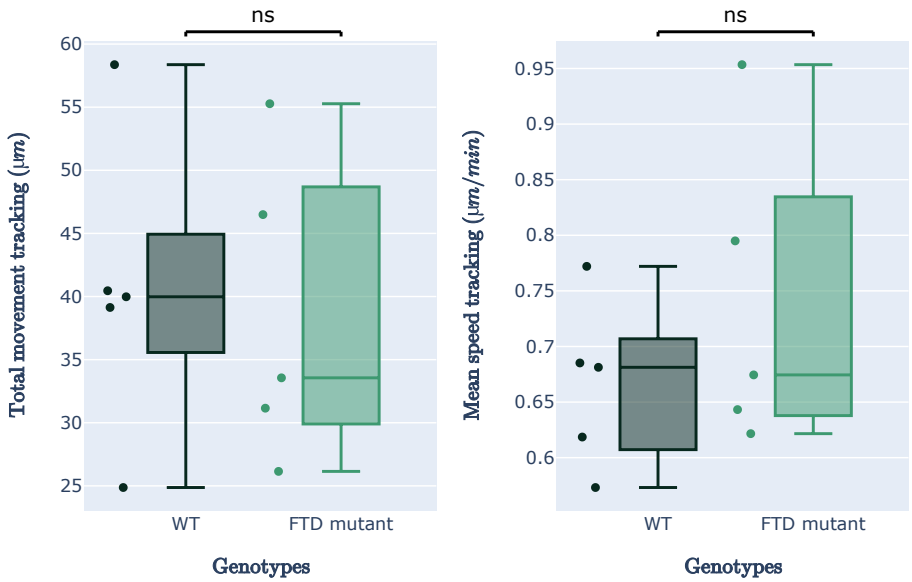


FIGURE 2.15: **Additional quantitative results of FTD-mutant versus WT microglial cells:** On the left, the quantification of the cells' mean speed and on the right, the quantification of total cells movement are presented. Statistical analysis was conducted using the Mann–Whitney–Wilcoxon test, where a non-significant result is indicated by a p-value ≥ 0.05 (ns).

Phagocytosis dataset for microglial cell

In this study, we have undertaken a detailed analysis of the phagocytosis of protein aggregates by microglia within the context of FTD. FTD is a neurodegenerative disease characterized by mutations in genes that regulate microglial functions, such as *C9ORF72* and *GRN*, which are associated with specific types of aggregates composed of the TDP-43 protein (Bright F, 2021; Neumann M, 2006; Arai T, 2006). For the purpose of this study, we will collectively refer to these mutations as FTD mutants, as distinguishing between *C9ORF72* and *GRN* is not pertinent to our objectives.

The data for this study comprises acquisitions from wild-type (WT, n=5) and frontotemporal dementia (FTD, n=5) microglial cells during the phagocytosis process. Each acquisition consists of 20 distinct scenes, recorded over seven hours of time-lapse video microscopy at one frame every two minutes, capturing both cell and aggregate images in two separate channels. Our team of biologists has meticulously generated a comprehensive dataset, which has been rigorously validated by the laboratory's ethical committee. This dataset includes 36496 normalized cell images and 36496 aggregate images, along with 1306131 individual instance masks for cells and 1553036 for aggregates. Additionally, it contains 36496 registered aggregates and data of the intermediate steps in tabular format, all generated using the PhagoStat algorithm. To maintain high data quality, we applied various data quality correction techniques. The dataset offers an extensive array of biological features, such as area, position, and speed, presented on a frame-by-frame basis to facilitate in-depth analysis.

To further enhance the dataset's utility, we incorporated 226 manually annotated images, which collectively contain 6430 individual cell masks (seed dataset). These images, encompassing a variety of conditions including WT and FTD, were selected randomly from

a broad spectrum of scenes. Initially, a polygon-based method was employed by our team of expert biologists to annotate 61 images, a process facilitated by the use of [1]. Despite their meticulous efforts, the annotation process was slow and fraught with inconsistencies, particularly due to the irregular shapes of the cells. To address these challenges, we developed "Point2Cell," a GUI-based annotation tool[2], which markedly enhanced the efficiency of our annotation process. This new tool not only improved the Dice score for precision from 91.3% to 94.97%, but also reduced the average annotation time from 96.1 seconds to 14.6 seconds per image. Consequently, Point2Cell was utilized to annotate the remaining 165 images, accounting for 4694 individual cell masks, which were designated as the test set for objective model evaluation. The initial set of 61 images, encompassing 1736 cell masks and annotated via LabelMe, were retained for training and validation to ensure the models were well-tuned prior to testing.

The resulting dataset comprises a robust collection of 235288 files (94GB) of '2D + time' aggregate and cell images in a monolayer configuration, providing researchers with a valuable resource for investigating various cellular and aggregate properties in their studies.

Firstly, our dataset, available at [3], serves as a comprehensive benchmark, providing a reliable reference for researchers aiming to test the efficacy of their algorithms or techniques against PhagoStat. This meticulously curated dataset includes data from both WT and FTD mutant microglial cells engaged in the phagocytosis process, establishing a robust basis for comparative analysis. Secondly, by detailing parameters such as area, position, and speed on a frame-by-frame basis, our dataset exemplifies the capabilities of PhagoStat to thoroughly analyze microglial phagocytosis of protein aggregates, especially pertinent in FTD contexts. Lastly, the dataset is invaluable during the pre-training phase for those seeking to extend the application scope of PhagoStat across different data modalities. Utilizing our dataset for model pre-training can potentially hasten model convergence and enhance generalization performance. To facilitate this, we employed AttUNet(XAI) and UNet(XAI) models on subsets WT-1 to WT-3 and FTD-1 to FTD-3 for training, with subsets WT-4 and FTD-4 used for validation. Subsequent evaluations on WT-5 and FTD-5 resulted in both models achieving a Dice score of 97.88%.

2.6 Discussion

Phagocytosis, a crucial cellular process, acts as a primary defense mechanism against danger signals and pathogens, vital for the immune system's functionality. Within the brain, microglial cells are exclusively responsible for phagocytosis. The importance of their phagocytic activity has become increasingly recognized in neurodegenerative disease research, where neuroinflammation is implicated in disease pathology, potentially through mechanisms such as the clearance of aggregate formations or the aberrant phagocytosis of live neurons and synapses. Recent studies have also explored the potential of antibody-mediated clearance of aggregates by phagocytic microglia as a therapeutic avenue for Alzheimer's disease and other dementias.

[1] LabelMe: polygon-based annotation tool
[2] Point2Cell: seed based annotation tool
[3] Microglial dataset

Quantifying the phagocytosis of amorphous and highly active unstained cells presents significant challenges, crucial for advancing our understanding of neurodegenerative diseases. This process typically involves the use of phase-contrast time-lapse video microscopy to capture rapid cellular interactions, although distinguishing these cells from their background remains difficult.

Addressing these challenges, the PhagoStat framework offers a scalable, real-time analysis solution that utilizes high-performance computing clusters to efficiently process large datasets, as demonstrated by its ability to manage 750 GB across ten CZI files in merely 97 minutes with only CPU utilization. PhagoStat's adaptability is further evidenced by a tripling in data retrieval speeds upon upgrading from HDD to SSD storage, signifying ongoing performance improvements. Designed to process data concurrently with its acquisition, PhagoStat enables completing analyses of 7-hour recordings within 20 minutes, synchronizing perfectly with microscope operations to ensure immediate availability of results upon recording completion, thus optimizing workflow efficiency.

PhagoStat also emphasizes transparency and compliance with General Data Protection Regulation (GDPR) guidelines, enhancing trust and understanding among users while ensuring data safety and reproducibility. Such features not only support scientific integrity but also accelerate the translation of research into practical applications.

By providing an interpretable and transparent pipeline, PhagoStat empowers a diverse range of users—from laboratory technicians to biologists and physicians—to deepen their understanding of the processes under study. This approach also fosters opportunities for pipeline optimization tailored to specific needs, contributing to sustainable practices by reducing the carbon footprint of technological operations.

The quality of data is paramount; hence, the PhagoStat pipeline includes robust data quality checks adaptable to varying acquisition conditions. However, our current registration method, CECC, although less biased than the commonly used SIFT method, requires further optimization to enhance processing speed. Investigating the biases associated with the SIFT method, particularly in landmark identification due to background dominance or geometric similarity of aggregates, could provide valuable insights for improving our pipeline.

Our transition from a deep learning-only to an interpretable deep learning approach has significantly reduced model size while maintaining high performance, facilitated by an automated evaluation system for feature map integrity. This system not only optimizes model selection but also conserves computational resources, expanding the models' usability across lower-specification hardware.

In this study, we employed FTD-mutants and wild-type cells to showcase our pipeline's utility, grouping GRN and C9ORF72 mutations under FTD-mutants for simplicity, though acknowledging the potential biological distinctions between these mutations as an area for future research. Our findings on the increased size and activity in mutant cells contribute significantly to neurodegenerative disease research, with the public release of an extensive dataset on microglial cell phagocytosis adding a valuable resource for the community.

Looking ahead, we are excited about the potential of moving into 3D spatio-temporal analysis, which represents not just a shift in dimensionality but a leap towards a more genuine in vivo understanding of cellular behavior. This transition introduces substantial

computational and data management challenges, heralding a new era in methodological computer vision and interpretable AI, crucial for advancing models of neurodegenerative diseases like brain organoids.

Chapter 3

Trustworthy Generative Models in Computational Pathology

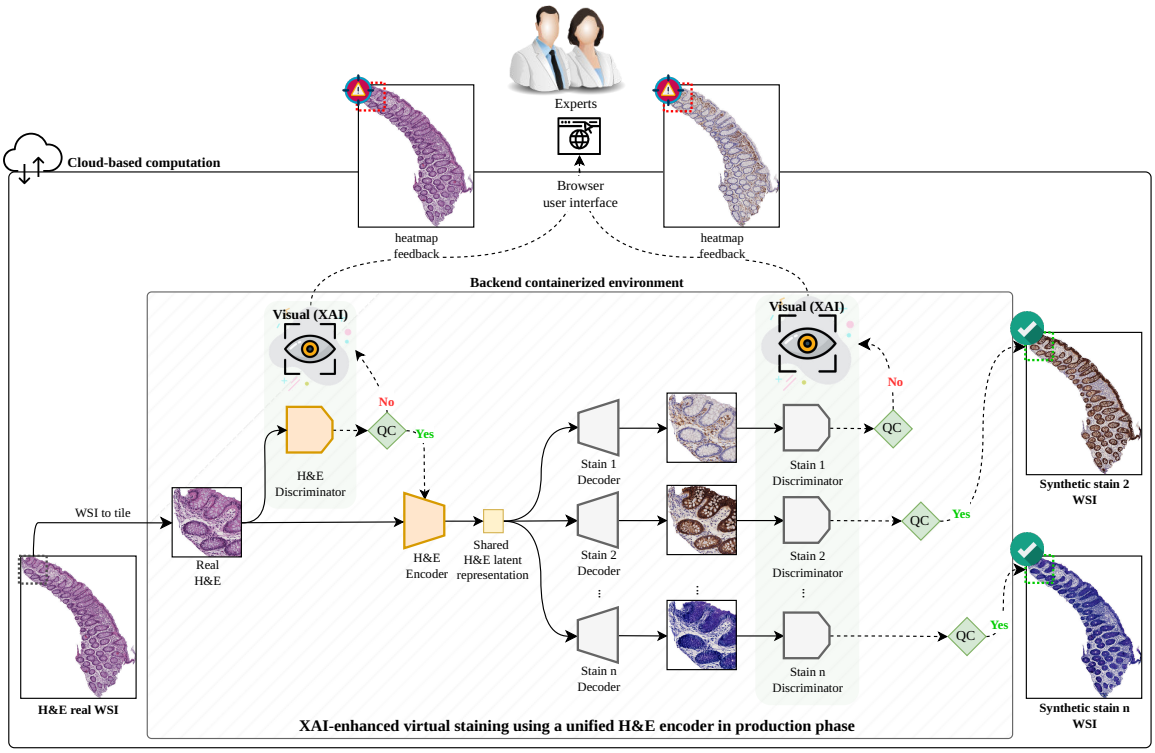


FIGURE 3.1: **Visual-XAI-enhanced trustworthy virtual staining approach.** End-to-end virtual staining approach generating synthetic IHC stains by using a single H&E encoder and multiple stain decoders. Quality check (QC) protocol based on self-inspection features uses trained discriminators to consolidate trust in the generated synthetic stains, by ensuring the alignment of the new H&E slides with the trained distribution and by validating the quality of the generated stained slides. Integration of cloud-based computing enhances accessibility and adoption by enabling pathologists to efficiently process large datasets from anywhere, while end-to-end system's algorithms are handled in a back-end containerized environment.

Submitted scientific publication and patents

Ounissi, M., Sarbout, I., Hugot, J. P., Martinez-Vinson, C., Berrebi, D., & Racoceanu, D. (2024). Scalable, Trustworthy Generative Model for Virtual Multi-Staining from H&E Whole Slide Images. arXiv preprint.
https://arxiv.org/abs/2407.00098.

Ounissi, M., Berrebi, D., & Racoceanu, D. (2024). Patent submitted: **EP 24 305 224.8**: Trustworthy and Scalable Unpaired Virtual Multi-Staining.

Ounissi, M., Berrebi, D., & Racoceanu, D. (2024). Patent submitted: **EP 24 305 221.4**: Trustworthy and Scalable Paired Virtual Multi-Staining.

Summary

Chemical staining methods, while reliable, are time-consuming and resource-intensive, raising environmental concerns. Virtual staining offers a faster, more flexible alternative without the physical and chemical costs. Generative AI technologies can address these challenges, but their opaque processes complicate adoption in high-stakes healthcare decisions, especially in computational pathology. Our work introduces an innovative approach using generative models for virtual stain transformations, enhancing performance, trustworthiness, scalability, and adaptability. A single Hematoxylin and Eosin (H&E) encoder supports multiple stain decoders, prioritizing critical regions in the latent space for precise synthetic stain generation. Our method, tested to generate eight different stains from one H&E slide, offers scalability by loading only necessary components during production. We integrate label-free knowledge during training to minimize artifacts, enhancing virtual staining accuracy in both paired and unpaired settings. To build trust, we use real-time self-inspection with trained discriminators, providing pathologists with confidence heat-maps. Automatic quality checks on new H&E slides ensure high-quality synthetic stains. Our open-source, cloud-based proof-of-concept system allows easy virtual staining through a browser, addressing common hardware and software challenges and facilitating real-time feedback. Additionally, we have curated a novel dataset of eight different paired H&E/stains related to pediatric Crohn's disease, providing 30 whole slide images for each stain set (480 WSIs total) to stimulate further research in computational pathology.

Résumé

Les méthodes de coloration chimique, bien que fiables, sont chronophages et consomment beaucoup de ressources, soulevant des préoccupations environnementales. La coloration virtuelle offre une alternative plus rapide et flexible sans les coûts physiques et chimiques. Les technologies d'intelligence artificielle générative peuvent relever ces défis, mais leurs processus opaques compliquent l'adoption dans les décisions de santé à enjeux élevés, notamment en pathologie computationnelle. Notre travail introduit une approche innovante utilisant des modèles génératifs pour les transformations de coloration virtuelle, améliorant la performance, la fiabilité, l'évolutivité et l'adaptabilité. Un seul encodeur Hematoxyline et Éosine (H&E) prend en charge plusieurs décodeurs de coloration, en donnant la priorité aux régions critiques dans l'espace latent pour une génération précise de taches synthétiques. Notre méthode, testée pour générer huit colorations différentes à partir d'une seule lame H&E, offre une évolutivité en chargeant uniquement les composants nécessaires lors de la production. Nous intégrons des connaissances sans étiquette pendant la formation pour minimiser les artefacts, améliorant ainsi la précision de la coloration virtuelle dans des contextes appariés et non appariés. Pour instaurer la confiance, nous utilisons une auto-inspection en temps réel avec des discriminateurs entraînés, fournissant aux pathologistes des cartes de confiance. Des contrôles de qualité automatiques sur les nouvelles lames H&E garantissent des colorations synthétiques de haute qualité. Notre système de preuve de concept open-source et basé sur le cloud permet une coloration virtuelle facile via un navigateur, en répondant aux défis matériels et logiciels courants et en facilitant les retours en temps réel. De plus, nous avons créé un nouvel ensemble de données comprenant huit différentes paires H&E/colorations liées à la maladie de Crohn pédiatrique, fournissant 30 images de lames entières pour chaque ensemble de colorations (480 images au total) afin de stimuler la recherche en pathologie computationnelle.

3.1 Opportunities and Challenges in Virtual Staining

Hematoxylin and Eosin (H&E) staining, a cornerstone of histopathology, is globally recognized for its cost-effectiveness and has solidified its place in routine diagnostic protocols, including cancer grading (Saha, Chakraborty, and Racoceanu, 2018; Echle et al., 2021). Despite its merits, H&E staining falls short in identifying specific proteins, a critical factor for precise disease diagnosis and severity assessment.

To address this limitation, Immunohistochemical (IHC) staining has been adopted as an effective alternative, particularly for identifying specific proteins critical for classifying various tumor types and pinpointing the origins of metastatic tumors. This technique is indispensable for detecting minute tumor cells that might elude standard staining procedures and is particularly beneficial for diagnosing diseases that elude traditional biopsy cultures and serological diagnostics (Magaki et al., 2019; Oumarou Hama et al., 2022). However, the IHC method is resource-intensive, prone to errors, and potentially delays diagnosis, which could be detrimental to patient care. Moreover, the chemicals used in IHC can hinder further tissue analysis and present environmental risks (B. Bai et al., 2023).

These challenges highlight the urgent need for an automated, digital, and reliable staining process to optimize the selection of stains and enhance diagnostic accuracy. Advances in computational pathology, especially the application of deep learning techniques, are pivotal in this regard. These methods have successfully enabled the conversion of H&E stains to IHC stains, improving diagnostic accuracy (B. Bai et al., 2023; Haan et al., 2021).

Both supervised and unsupervised deep learning strategies have shown promise in transforming H&E to IHC stains across various organs. The supervised method, often termed "paired", and the unsupervised method, known as "unpaired", do not necessarily require aligned slides, which adds flexibility to the staining process (Borhani et al., 2019; Rivenson et al., 2019; T. M. Abraham et al., 2022).

Despite the potential of these computational techniques, the adoption of deep generative models faces skepticism due to trust issues among clinicians and pathologists, particularly concerning their applicability in real-world scenarios. The requirement for specialized hardware and software further complicates their integration into routine pathology practices.

In response to these challenges, we propose a novel computational pathology pipeline that enhances the scalability, accuracy, trustworthiness, and utility of virtual staining techniques. Our approach integrates a unified encoder with multiple stain decoders, incorporates trust-building mechanisms through self-inspection, and utilizes advanced training methods without the need for additional annotations. A cloud-based deployment strategy and a unique dataset focused on pediatric Crohn's disease further enhance the utility and applicability of our pipeline in computational pathology.

Virtual staining has emerged as a transformative approach for efficient stain transformations in WSIs. In recent years, generative adversarial networks have facilitated the generation of multiple stains, marking significant advancements in the field. Despite these improvements, challenges persist in terms of scalability, accuracy, trustworthiness, and accessibility for clinicians (Ciompi et al., 2017; B. Bai et al., 2023). This section delves into the literature, focusing particularly on the H&E to IHC transformation. Our aim is to underscore the existing limitations and provide a detailed overview of the current research landscape, thereby situating our study within the broader context of computational pathology.

3.1.1 Advancements in stain synthesis through deep learning techniques

Computational pathology intensively investigates the capabilities of stain transformations and synthesis. These scholarly pursuits are oriented towards the precise digital replication of tissue slide staining utilizing paired datasets, which incorporate both H&E stains along with corresponding WSIs in alternative stains. Several pivotal studies exemplify progress in this area. For instance, (Haan et al., 2021) implemented a deep learning algorithm that processes H&E tiles and concurrently produces Jones, MT, and PAS stains. Likewise, (Burlingame et al., 2020) formulated the SHIFT method utilizing a paired pancreas dataset to transmute H&E into virtual immunofluorescence imagery, estimating the distribution of the tumor cell marker pan-cytokeratin. Building upon these developments, (Hong et al., 2021) employed a paired gastric carcinomas dataset to synthesize cytokeratin staining from H&E, aiding in the diagnosis of gastric cancer.(Xie et al., 2022) leveraged a paired prostate dataset to transform H&E to CK8 IHC stains, a fundamental step towards reconstructing 3D segmented glands for prostate cancer risk stratification. Additionally, (S. Liu, C. Zhu, et al., 2022) devised a pyramid approach to generate human epidermal growth factor receptor 2 IHC stain from H&E using a paired breast cancer dataset.

Despite these technological advancements, the methodology of paired H&E/IHC staining presents significant challenges. (Yang et al., 2022) underscores that the staining procedures are generally irreversible and pose logistical and technical obstacles in acquiring paired data. Moreover, inconsistencies within one staining type can undermine the accuracy of the other, thereby diminishing the overall diagnostic efficacy.

In response to these challenges, innovative methodologies have emerged. One such approach involves the C-DNN (Yang et al., 2022) method, which utilizes cascaded deep neural networks to transform images from auto-fluorescence to H&E, and subsequently to PAS, effectively bypassing the difficulty of acquiring paired data. Additionally, the utilization of unpaired dataset configurations has been explored, notably through the CycleGAN (Goodfellow, Pouget-Abadie, et al., 2014; J. Zhu et al., 2017), a widely adopted model. Unpaired datasets have facilitated transformations such as those by (Levy and al., 2020) from H&E to trichrome, and by (Mercan et al., 2020) from H&E and PHH3 stains. Further investigations by (Lahiani et al., 2021) incorporated a perceptual embedding consistency loss in Generative Adversarial Networks (GANs), and (S. Liu, B. Zhang, et al., 2021) generated Ki-67-stained images from H&E-stained samples. Moreover, the MVFStain framework (R. Zhang et al.,

2022) succeeded in converting H&E-stained images into multiple virtual functional stains across diverse scenarios.

In the academic state-of-the-art, two predominant methodological frameworks for domain representation are identified. The first approach engages separate pairs of encoders, decoders, and discriminators for each domain pairing, as illustrated by CycleGAN (J. Zhu et al., 2017) and its derivatives. This strategy mandates the training of $3 \times n$ models, often precipitating scalability impediments during the training phase due to the substantial computational resources required. In contrast, methodologies such as StarGAN (Y. Choi, M. Choi, et al., 2017; Y. Choi, Uh, et al., 2019) employ a consolidated model that includes a mapping network, a style encoder, a generator, and a discriminator. This architecture facilitates the generation of multiple latent domain representations, each styled as distinct domains, thereby streamlining the training process by utilizing a single model to support multiple transformations.

Nonetheless, these methodologies present certain limitations, particularly in specialized applications. For instance, a pathologist requiring a specific subset of stains—namely, s out of S potential stains derived from H&E—encounters a significant computational challenge. They must either deploy $2 \times s$ models (an encoder and a decoder for each stain required) or rely on an overarching model that incorporates all S stains. Both scenarios demand extensive computational resources, consequently impeding prompt real-time responses. Moreover, the existing virtual staining technology does not facilitate the concurrent synthesis of more than three IHC stains in a single session, representing a notable constraint in scalability and adaptability to the diverse demands of clinical settings.

Despite these hurdles in digital pathology, the broader domain of image processing has witnessed appreciable advancements in addressing similar scalability concerns. For instance, approaches such as those delineated in (Anoosheh et al., 2017), which employ a separate encoder, decoder, and discriminator for each domain, demonstrate considerable scalability potential. This success in other fields suggests that analogous methodologies could be transposed to digital pathology, potentially amplifying scalability and efficacy in a domain where these qualities are critically needed.

The progression of computational pathology has immensely benefited from diverse training strategies, loss functions, and regularization techniques. Contributions from studies such as (Q. Liu et al., 2018; Tellez et al., 2018) have led to significant enhancements in model performance. Yet, embedding knowledge in a self-supervised manner without depending on additional labels remains a complex challenge.

Additionally, the contextual framework within which computational pathology operates, particularly the selection of magnification in WSI interpretation, has attracted increasing scrutiny. Research such as (Sirinukunwattana et al., 2016; Courtney et al., 2018; Kosaraju et al., 2020à has underscored the pivotal role of context in augmenting the efficacy of deep learning models in tissue characterization and cell classification. However, within the realm of virtual staining, there exists a considerable disconnect, with methods often reliant on arbitrary magnification scale selections. This accentuates the necessity for further exploration in virtual staining techniques that leverage both paired and unpaired datasets, aiming to enhance their applicability and effectiveness.

In conclusion, the substantial impact of synthetic stains on patient outcomes necessitates that these methods be both efficient and reliable. Concerns pertaining to their interoperability and consistency remain critical areas for enhancement, which our research endeavors to address within the existing academic milieu.

3.1.2 Advancements in cloud-enabled computational pathology

In recent years, the field of collaborative image analysis systems has experienced remarkable advancements, with the emergence of several influential platforms that have reshaped the domain. QuPath (Bankhead, Loughrey, and Fernández, 2017) pioneered the introduction of web-based remote collaboration in computational pathology, enabling annotations and the incorporation of modifiable algorithms through JavaScript and Groovy. Additionally, the Open Reproducible Biomedical Image Toolkit (ORBIT) (Stritt, Stalder, and Vezzali, 2020) was launched, specializing in integrating existing analysis tools for medical imaging, with its collaborative capacities augmented by the inclusion of OMERO (Besson et al., 2019; Linkert et al., 2010).

Although some tools exhibit limited AI capabilities, Cytomine (Marée et al., 2016) sets itself apart with its innovative web-based interface. It was the first platform to facilitate the display of multiple WSIs within a web environment, obviating the need for software installation. Cytomine's platform is notably comprehensive, encompassing all essential components for server deployment—including web servers, job concurrency management, data storage, and a robust API. This integration renders it particularly suitable for histopathology applications.

Furthermore, the platform enhances inclusivity and reproducibility of results by supporting any dockerized algorithm. This feature grants authorized users access to an extensive array of tools for collaborative medical image analysis. The platform's design also facilitates job monitoring and enhances user interaction, which in turn improves collaboration and workflow management. Due to its efficacy in promoting collaboration, efficiently managing medical image data, and integrating advanced machine learning techniques, Cytomine is increasingly favored across various applications.

To the best of our knowledge, no cloud-based open-source platform has previously incorporated virtual staining in a reliable manner. In response to this gap, and in alignment with our technological advancements and research objectives, we have integrated our virtual staining method into the platform as a proof of concept. This integration provides a framework that empowers pathologists by eliminating the need for specific hardware and software requirements. It saves time and enhances their diagnostic and research capabilities in medical imaging analysis. This achievement is realized through a browser interface where all complex computations are managed in the backend, streamlining the user experience.

3.1.3 Datasets in virtual staining: challenges and opportunities

In the field of computational pathology, the transformation and synthesis of stains constitute critical areas of research that aim to enhance diagnostic precision. Numerous studies in this domain rely on the utilization of diverse datasets, particularly paired datasets, which include

both H&E stains and their corresponding IHC stains on the same tissue slide. For example, a significant investigation by (Burlingame et al., 2020) harnessed a paired dataset of pancreas tissues to develop the SHIFT method, which converts H&E images into virtual PanCK immunofluorescence images, thus estimating the distribution of the tumor cell marker pan-cytokeratin. Similarly, (Haan et al., 2021) used a dataset of paired tissue slides to transmute H&E tiles into Jones, Masson's Trichrome, and Periodic Acid–Schiff stains.

Research also extends to datasets that feature various cancer types. (Hong et al., 2021) employed a paired dataset of gastric carcinomas to generate cytokeratin staining from H&E, aiding in the diagnosis of gastric cancer. (Xie et al., 2022) utilized a paired prostate dataset to transform H&E into CK8 IHC stains, with the goal of reconstructing 3D segmented glands for prostate cancer risk stratification. Moreover, (S. Liu, C. Zhu, et al., 2022) concentrated on producing the human epidermal growth factor receptor 2 (HER2) IHC stain from H&E using a paired breast cancer dataset.

While paired datasets are fundamentally valuable, they are not devoid of challenges. Given that staining procedures are generally irreversible, acquiring such data can present technical difficulties (Yang et al., 2022). In response to these obstacles, the exploration of unpaired datasets has commenced. For instance, (Levy and al., 2020) successfully transformed H&E to trichrome using an unpaired liver dataset and modified H&E to SOX10 IHC using a skin and lymph node dataset. Innovatively, (Lahiani et al., 2021) implemented a perceptual embedding consistency loss in Generative Adversarial Networks (GANs), using an unpaired liver dataset to convert H&E into FAP-CK IHC stain. Additionally, studies like (S. Liu, B. Zhang, et al., 2021) have generated Ki-67-stained images from H&E-stained samples using unpaired and unbalanced datasets from neuroendocrine tumors and breast cancers. The MVFStain framework (R. Zhang et al., 2022) also stands out, transforming H&E-stained images into various virtual functional stains for tissues including mouse lung, breast cancer, and rabbit cardiovascular system.

Despite these advances in computational pathology, the availability of data remains a limiting factor, particularly the paucity of high-quality public paired H&E/IHC stain datasets. For example, although (Haan et al., 2021) publicly shared the source code of their approach, the dataset they employed remains proprietary. Furthermore, specific domains, such as pediatric Crohn's disease at the diagnosis stage (pre-treatment), are under-researched and offer avenues for future investigations.

3.2 Multi-Virtual Staining: Scalability and Performance

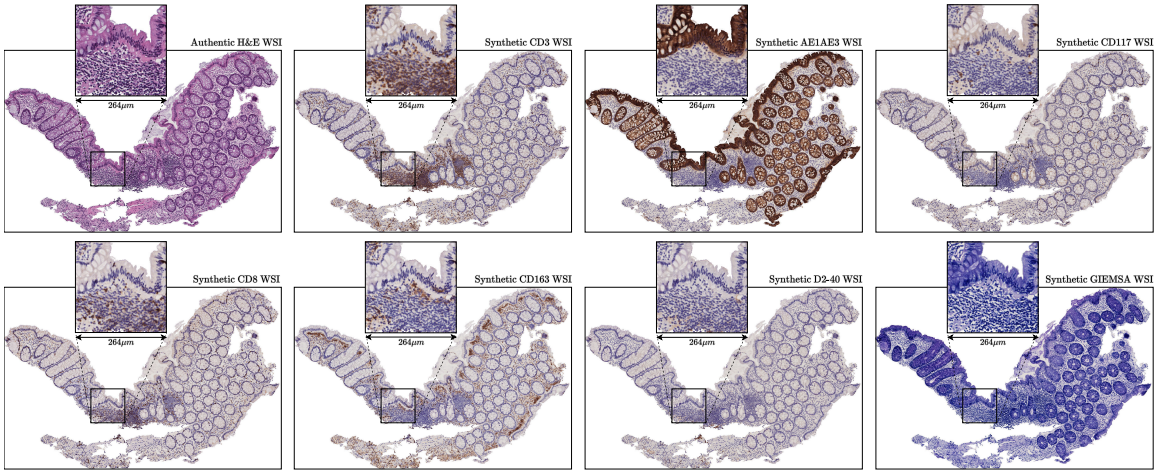


FIGURE 3.2: **Multi-Virtual Staining Outcomes Associated with Crohn's Disease.** This figure illustrates the high-resolution WSIs of diverse synthetic stains, generated through the application of $\mathcal{L}_{\text{IHC}}$ and $\mathcal{L}_{\text{H\&E}}$ loss functions within an unpaired framework.

3.2.1 Comparative analysis of unified versus individual H&E encoders

In this investigation, we endeavored to devise an advanced methodology capable of simultaneously generating multiple virtual stains. Current practices, as explicated in Section 3.1.1, restrict the simultaneous production to a maximum of three stains. Motivated by a desire to transcend these constraints, we adopted style transfer methods from frameworks such as ComboGAN (Anoosheh et al., 2017), which is proficient in managing up to 14 disparate art styles. This approach was tailored for histopathological contexts by integrating a novel framework comprising a dedicated H&E encoder, generator, and discriminator. This architecture facilitates the concurrent training from H&E to multiple stains S, as depicted in Section 3.4.1 – Figs. 3.1, 3.6, and 3.5.

Our empirical findings, delineated in Table 3.1, underscore the superiority of utilizing a unified H&E encoder for multi-virtual staining applications. Synthetic stains derived from this encoder consistently exceeded the performance of those from separate encoders tailored to individual stains. We assessed the efficacy using the Mean Square Error (MSE) metric by comparing the synthetic stains against genuine counterparts using a paired test set of H&E samples. These comparisons demonstrate that our methodology significantly surpasses the performance of the CycleGAN approach.

Both the unified and CycleGAN methodologies were evaluated under identical experimental conditions, encompassing the same dataset, equivalent training durations, and consistent architectures for the encoders, generators, and discriminators. Our approach not only enhances computational efficiency through the utilization of a single encoder, decoder, and discriminator across the staining process but also minimizes the number of trainable parameters compared to the CycleGAN method (refer to Fig. 3.10). This streamlined architecture increases computational efficiency and fosters scalability by accommodating a wider array of output stains and expediting the training phase.

In summary, the unified H&E encoder method surpasses alternative techniques by producing more accurate synthetic stains and demonstrating superior computational efficiency, thereby establishing itself as a scalable and robust solution for extensive histopathological investigations.

Method	AE1AE3	CD117	CD15	CD163	CD3	CD8	D240	GIEMSA	Overall MSE↓	Parameters↓
CycleGAN (J. Zhu et al., 2017)	0.1027	**0.0504**	0.0773	**0.0660**	0.0464	0.2659	0.0379	**0.0321**	0.0848±0.0717	409M
Ours	**0.0854**	0.0516	**0.0674**	0.0693	**0.0463**	**0.0836**	**0.03**	0.0365	**0.0588±0.0209**	**230M**

TABLE 3.1: **Enhanced Performance and Efficiency in Multi-Virtual Staining via a Unified H&E Encoder.** This table presents the Mean Squared Error (MSE) metrics (mean±standard deviation) of synthetic stain generation in an unpaired setting using our unified H&E encoder compared to the traditional distinct H&E encoders per stain (CycleGAN). The results are computed on a patch-by-patch basis, which is a common approach used in CycleGAN models. The data underscore the enhanced accuracy and computational efficiency of our method, evidenced by a significantly reduced count of trainable parameters. This attribute showcases the scalability and clinical efficacy of our approach for histopathological applications. For reproducibility details, please see Section B.2.1.

3.2.2 Context-importance

Owing to the substantial dimensions of WSIs, typically on the order of 10000×10000 pixels, contemporary GPUs are incapable of processing an entire slide in one pass during the training phase. Consequently, virtual staining methodologies frequently adopt a sliding window tiling strategy. This technique partitions the slide into smaller tiles, or patches, which align with the operational constraints of deep learning architectures and computation capacities (e.g. GPU memory). The commonly employed dimensions for these patches are 128×128 pixels and 256×256 pixels (R. Zhang et al., 2022; Lin et al., 2022). Implementing this strategy necessitates meticulous selection of the magnification level for analysis (e.g., $10\times$, $20\times$, and $40\times$), as it significantly influences the learning paradigm in both paired and unpaired scenarios. Utilizing smaller patches escalates the total number of patches per WSI, thereby raising concerns regarding the inference time, specifically the time required to reconstruct a virtually stained slide. Addressing these technical challenges is essential not only for enhancing performance metrics but also for understanding the practical implications concerning inference times, a critical consideration for pathologists.

In our empirical investigations, by adopting a modular approach that obviates the need for simultaneous loading of all model components, we successfully processed tiles of 512×512 while concurrently training on 8 stains plus H&E on a conventional 16GB GPU. This configuration provides a spatial resolution at least fourfold greater than those reported in (R. Zhang et al., 2022; Lin et al., 2022), offering enhanced flexibility and the ability to incorporate more contextual data within each patch. To ascertain the optimal magnification for virtual staining, we trained our model at various magnifications, each resized to a uniform dimension of 512×512 pixels for consistent image processing. The magnifications tested included $10\times$ (original tile size of 2048×2048 pixels $\approx 450.56 \times 450.56 \mu$m), $20\times$ (original tile size of 1024×1024 pixels $\approx 225.28 \times 225.28 \mu$m), and $40\times$ (original tile size of 512×512 pixels $\approx 112.64 \times 112.64 \mu$m), under both paired and unpaired learning settings to evaluate the impact of magnification on model efficacy. As depicted in Table 3.2, in the paired settings, all magnifications provided comparable outcomes due to the direct correspondence between the H&E-stained WSI and other stained WSIs. Our studies in unpaired settings disclosed that lower magnifications, offering wider contextual views, significantly enhance model performances, underscoring the importance of integrating extensive contextual information for effective learning where direct stain correspondences are absent, thus guiding future advancements in virtual staining technologies.

In the paired analysis, initial experiments employed a $10\times$ magnification, correlating to a resolution of 512×512 pixels (approximately 0.88μm per pixel). The original images were further resized to 1024×1024 pixels (0.44μm per pixel) to more effectively assess the impact of pixel density on high-context paired training. To fully exploit the capabilities of high-end GPUs, such as the NVIDIA A100 80GB, we additionally experimented with a maximal image size of 1400×1400 pixels (0.32μm per pixel) during the training phase.

Table 3.3 illustrates the scalability of our modular approach, capable of processing images with eight stains plus H&E up to the 1400×1400 resolution. Comparative analyses presented in Tables 3.2 and 3.3 underscore that augmenting contextual information within the images is substantially more beneficial than merely increasing pixel density.

	Trained on			Tested on			Metrics			
Setting	x10	x20	x40	x10	x20	x40	MSE($\times 10^{-2}$)↓	PSNR↑	SSIM↑	Inference time↓
Paired	✓			✓			**1.658±0.586**	22.85±2.60	86.37±6.9	**14.06 sec**
		✓			✓		1.681±0.618	22.84±2.79	85.42± 6.8	19.81 sec
			✓			✓	1.714±0.636	22.75±2.80	**87.22±5.3**	49.67 sec
		✓		✓			1.659±0.632	**22.94±2.89**	86.46±7.1	19.81 sec
			✓	✓			1.673± 0.640	22.90±2.90	86.50±7.1	49.67 sec
Unpaired	✓			✓			**2.329±0.871**	**21.34±2.39**	**83.79±7.6**	**14.06 sec**
		✓			✓		3.476±2.227	20.06±3.25	79.60±10.8	19.81 sec
			✓			✓	4.498±3.668	19.54±4.20	80.59± 9.7	49.67 sec
		✓		✓			4.389±3.639	19.72±4.30	76.45±15.7	19.81 sec
			✓	✓			3.639±2.226	20.22±3.33	79.64±12.8	49.67 sec

TABLE 3.2: **Performance Analysis of the Model Across Varied Magnifications.** This table delineates the efficacy of our modular training methodology at different magnifications (10×, 20×, and 40×), wherein tiles extracted at each magnification were uniformly resized to 512×512 pixels to maintain consistent image dimensions for analysis. The models were subjected to training utilizing the $\mathcal{L}_{\text{IHC}}$ and $\mathcal{L}_{\text{H\&E}}$ loss functions. In the paired learning context, the performance across magnifications appeared homogeneous, indicating no discernible preference for any specific magnification. In contrast, in the unpaired learning context, the lower magnifications, which encapsulate a broader contextual window, exhibited a pronounced advantage. This enhancement emphasizes the critical role of extensive contextual information in scenarios where direct stain correspondences are absent, thus facilitating more effective learning. For detailed information on experimental procedures and reproducibility, refer to Section B.2.3.

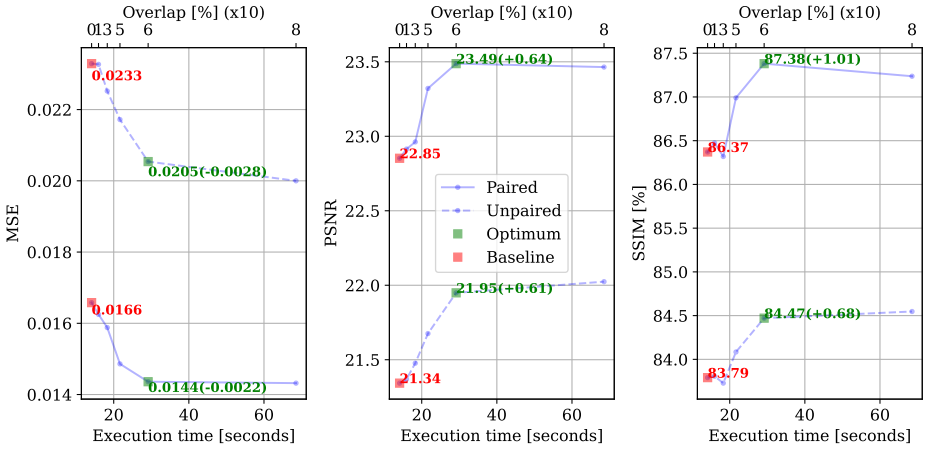
Trained on patch size			Tested on patch size			Metrics			
512^2	1024^2	1400^2	512^2	1024^2	1400^2	MSE($\times 10^{-2}$)↓	PSNR↑	SSIM↑	Inference time↓
✓			✓			**1.745±0.626**	22.63±2.61	86.24±6.8	**14.06 sec**
	✓			✓		1.766±0.654	22.61±2.74	85.27±6.8	22.04 sec
		✓			✓	1.87±0.723	22.40±2.79	85.43±6.6	37.51 sec
	✓		✓			1.748±0.667	**22.70±2.84**	**86.33±7.0**	22.04 sec
		✓	✓			1.853±0.714	22.45±2.84	86.11±7.1	37.51 sec

TABLE 3.3: **Scalability Assessment of the Multi-Virtual Staining Approach Across Various Training Resolutions in a Paired Setting.** This table delineates the outcomes of training our virtual staining model on images enhanced with eight stains plus H&E across diverse resolutions. The results reveal a consistent performance across different pixel densities, illustrating the robustness of our approach. Additionally, the data underscores the effective utilization of advanced GPU capabilities, thereby emphasizing the scalability of our method.

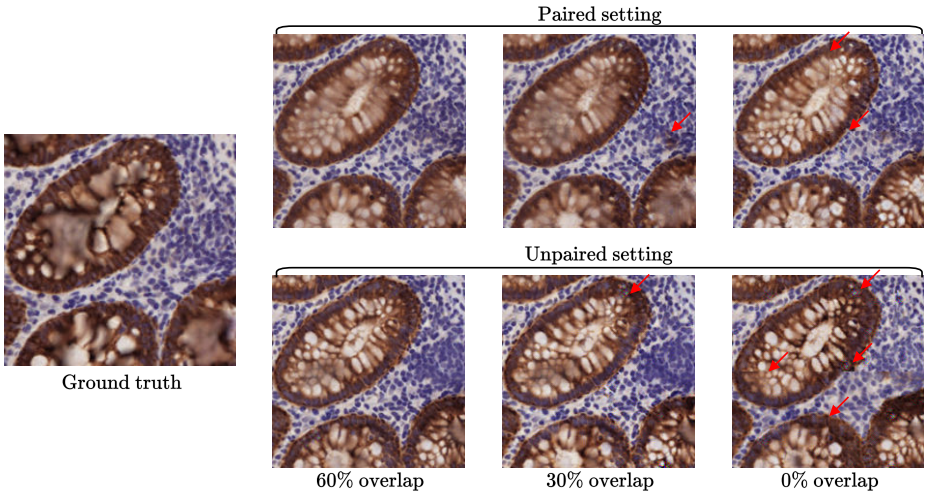
3.2.3 Eliminating stitching artifacts in synthetic slides

In our comprehensive examination of existing virtual staining methods (Mercan et al., 2020; Burlingame et al., 2020; Hong et al., 2021; S. Liu, B. Zhang, et al., 2021; Haan et al., 2021; R. Zhang et al., 2022; Lin et al., 2022), both paired and unpaired, it is apparent that a predominant number utilize a sliding window tiling approach during model training, as elucidated in Sections 3.1.1 and 3.2.2. This prevalent training technique frequently culminates in challenges associated with reconstructing WSIs from the resulting patches. Notably, this can induce visible stitching artifacts, such as abrupt color transitions at tile borders and errors adjacent to these boundaries, as demonstrated in Figure 3.3.b (0% overlap in both settings, indicated with red arrows). These artifacts not only diminish the reliability of these tools in the eyes of pathologists but also escalate cognitive strain and error rates during slide evaluations. This complication is ubiquitous among all tile-based virtual staining methods, emphasizing the imperative for a universal remediation strategy.

To address these challenges, we have devised a post-processing technique specifically tailored for WSIs within the realm of virtual staining. Our findings suggest that the models exhibit context sensitivity, with enhanced accuracy at the center of tiles and diminished precision near the edges. Capitalizing on this observation, our methodology involves stitching tiles with deliberate overlap, focusing on the central regions of the tiles using a Hamming window (Hamming, 1998; Oppenheim and Schafer, 1999) (refer to Section B.2.5), thus augmenting performance without necessitating additional training. This approach, illustrated in Figure 3.3, markedly ameliorates all evaluated metrics in both paired and unpaired configurations and leads to superior perceived image quality compared to the ground truth. It is noteworthy that while our post-processing technique slightly extends processing time, it yields a substantial improvement in the performance-to-time ratio. An optimal overlap of 60%, as depicted in the figures, achieves an optimal balance between performance enhancement and execution time (>1min per 8 stains). This post-hoc processing strategy not only effectively mitigates stitching artifacts but also enhances the overall utility of virtual staining technologies in clinical environments. By streamlining the workflow, it facilitates the routine application of these technologies in fast-paced clinical settings, potentially expanding their adoption and fostering trust among pathologists. This adjustment ensures the high quality of the generated WSI virtual stains (refer to Figures 3.2 and 3.4) while providing necessary spatial context and maintaining manageable processing times, thus aligning with the requirements and dynamics of contemporary anatomopathological practice.



(A) Overlap post-processing quantitative performance evaluation



(B) Overlap post-processing qualitative performance evaluation

FIGURE 3.3: **Post-Processing Effects on Stitching Artifacts and Objective evaluation in Virtually Stained Slides. (a)** Illustrates the enhanced outcomes achieved through various overlap strategies employing a Hamming window, highlighting the improved image quality and diminished artifacts. The optimal performance-to-time execution ratio is realized at a 60% overlap. **(b)** Demonstrates typical stitching artifacts at tile borders with overlaps of 0%, 30%, and 60%, indicated by red arrows, which exemplify the abrupt color transitions and errors near the boundaries. This figure elucidates the comparative analysis across performance metrics (MSE, PSNR, SSIM) in both paired and unpaired settings, underscoring the efficacy of the post-processing strategy in elevating the overall quality and promoting the integration of virtual staining technologies within clinical practices. For reproducibility details, refer to Section B.2.5.

3.2.4 Generalizing across diverse stain types

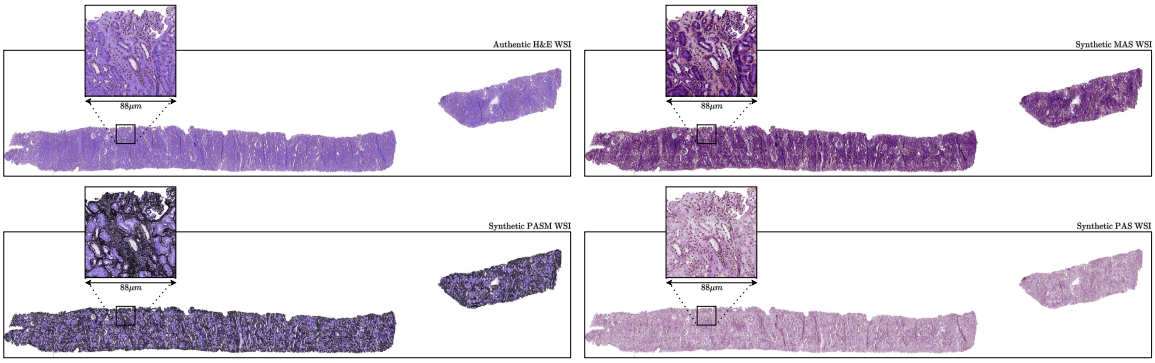


FIGURE 3.4: **Multi-Virtual Staining Results on Kidney Slide No. 5 from the AHNIR Dataset.** This figure demonstrates the high-quality synthetic stains produced by our methodology, showcasing the effectiveness of our approach.

The ANHIR dataset (Borovec et al., 2020) encompasses five collections of high-resolution human kidney tissue slides, where each collection comprises four sequential tissue slides stained with diverse histological stains (H&E, MAS, PAS, and PASM). Despite their structural similarities, these slides are not pixel-level aligned and all exhibit magnification at 40×.

Following the experimental protocol defined in UMDST (Lin et al., 2022), four collections (Kidney 1, Kidney 2, Kidney 3, Kidney 4) were designated as training datasets, with the fifth collection (Kidney 5) set aside for testing purposes. In alignment with the UMDST protocol, the H&E-stained slide from Kidney 1 was excluded due to notable color variation relative to other collections. For computational processing, the slides were segmented into 256x256 pixel tiles, with an overlap of 192 pixels.

Our methodology involved simultaneous training using the three stains, excluding H&E, through the utilization of four encoders, four decoders, and four discriminators. The training regimen consisted of 150000 iterations at a constant learning rate of 2×10^{-4}, followed by an additional 150000 iterations with a linearly decreasing learning rate, culminating in 300000 iterations in total. We employed the Adam optimizer with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and a batch size of 1, consistent with the UMDST protocol (Lin et al., 2022). Data augmentation was limited to random flipping and random rotation. The training was executed on a single NVIDIA A100 80GB GPU. Loss weights were defined with $\lambda_{\text{cyc}} = 10$ and $\lambda_{\text{adv}} = 1$, and the cycle consistency losses $\mathcal{L}_{\text{cyc},i}$ were calculated by averaging over components from the kidney dataset every three iterations, setting $\mathcal{L}_{\text{idt}} = 0$, $\mathcal{L}_{\text{lat}} = 0$, and $\mathcal{L}_{\text{fwd}} = 0$.

Evaluation on Kidney 5 was conducted as depicted in Figure 3.4, with qualitative outcomes summarized in Table 3.4. To compare with the leading-edge, the Contrast Structure Similarity (CSS) metric (Lin et al., 2022; L. Xing, Zeng, et al., 2017; L. Xing, Cai, et al., 2018) was applied, as detailed in Table 3.4.

Table 3.4 demonstrates enhanced performance and generalization capabilities relative to contemporary state-of-the-art approaches in MAS, PAS, and PASM staining techniques. The marginal underperformance in PASM staining, relative to the benchmarks established

Method	MAS	PAS	PASM	Overall	Tile-output	WSI-compliant	XAI	Scalability
MUNIT (X. Huang et al., 2018)	0.145	0.115	0.110	0.123±0.015	✓	×	×	×
FUNIT (M. Liu et al., 2019)	0.332	0.318	0.246	0.298±0.037	✓	×	×	×
StarGAN (Y. Choi, M. Choi, et al., 2017)	0.520	0.543	0.491	0.518±0.021	✓	×	×	×
UGATIT (J. Kim et al., 2019)	0.584	0.510	0.399	0.497±0.076	✓	×	×	×
UMDST (Lin et al., 2022)	0.682	0.674	**0.600**	0.652±0.036	✓	×	×	×
Ours	**0.797**	**0.784**	0.536	**0.705±0.120**	✓	✓	✓	✓

TABLE 3.4: **Comparative Analysis of Contrast Structure Similarity (CSS) Across Staining Methods and Computational Models.** This table delineates the CSS metrics for a range of computational methodologies applied to human kidney tissue slides stained with H&E, MAS, PAS, and PASM. Each model's efficacy is quantified via metrics including overall CSS, tile-based outputs, whole-slide imaging (WSI) compliant outputs and assessments, explainable AI (XAI) capabilities, and scalability. The results underscore the superior proficiency of our approach in addressing the multifaceted challenges of multi-virtual staining, where higher CSS values indicate enhanced preservation of structural fidelity across diverse stains.

by (Lin et al., 2022), can be attributed to the methodological emphasis on preserving the histomorphological features characteristic of H&E stains. This preservation is advantageous as it enhances the CSS metric, which quantifies the correlation between H&E and PASM stains. Notably, PASM staining inherently diminishes certain morphological details through its application of black coloration; consequently, our model is trained to replicate this effect, which should be regarded as an intrinsic feature rather than a flaw. Should there be a requirement to retain these histological details (yielding a less authentic PASM representation but conserving all H&E characteristics), a forward loss strategy may be employed as detailed by (Lin et al., 2022). This, however, introduces a compromise between the retention of morphological detail and the fidelity of staining. Additionally, our methodology excels in scalability during both the training and inference phases and distinctively incorporates XAI features. These capabilities, absent in alternative methodologies, are particularly vital within a clinical context.

3.3 Multi-Virtual Staining: XAI

3.3.1 Annotation-free knowledge guided training and H&E regularization

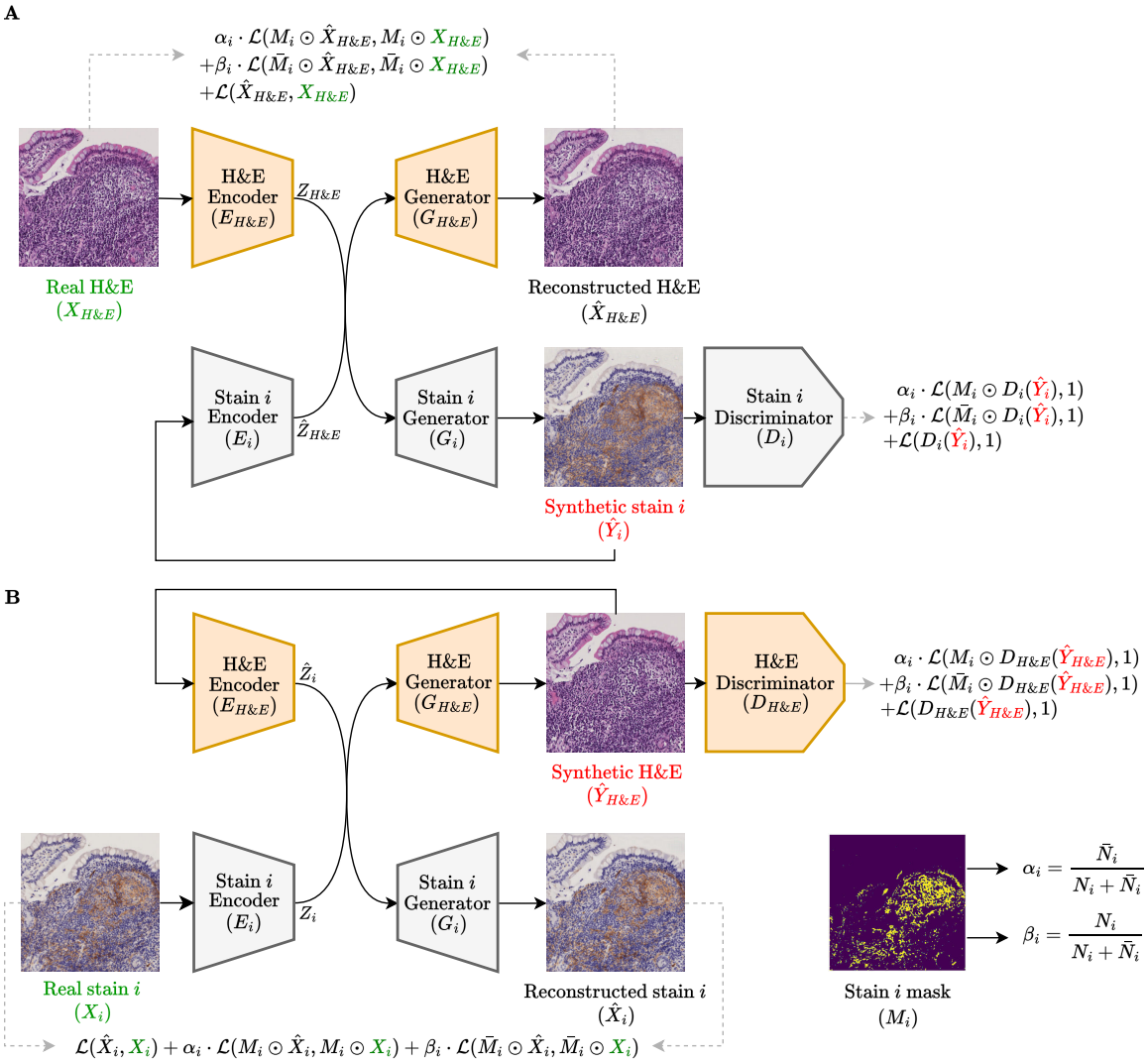


FIGURE 3.5: **An Overview of the Training Mechanism for Paired Stain Synthesis and Loss Function Computation in H&E $\leftrightarrow$ Stain i Conversion. A.** This part delineates the initial training cycle, initiating with a genuine paired H&E image $X_{H\&E}$, synthesizing a corresponding image in stain i denoted as $\hat{Y}_i$, and subsequently reconstructing the original H&E image $\hat{X}_{H\&E}$. This reconstruction serves to facilitate the computation of the loss function components, as elaborated in Section 3.4.1. **B.** This section outlines the second training cycle, commencing with a genuine stain i image X_i, generating a corresponding H&E image $\hat{Y}_{H\&E}$, and concluding with the reconstructed stain i image $\hat{X}_i$. The use of the staining mask M_i (where $\bar{M}_i$ denotes the complementary mask of M_i) is pivotal in computing various elements of the loss function, further detailed in Section 3.4.1. Each panel illustrates the model's enhancements aimed at increasing the precision and consistency of stain synthesis and discrimination within paired training scenarios.

To augment the reliability and trustworthiness of virtual staining techniques in histopathology, our approach aim to enriches the training model by integrating constraints derived from chemically stained slides. Unlike artistic style transfer applications, such as those documented in (Anoosheh et al., 2017; Y. Choi, M. Choi, et al., 2017; Y. Choi, Uh, et al., 2019), where the domain discrepancies simplify the discriminator's role, enhancing the pressure on

the generator for precise image reproduction. However, the intricacies in histopathological staining lie in the universal morphological characteristics present across various stains, with deviations primarily manifesting in activation responses to specific proteins. This scenario presents two primary challenges: (i) the model might underrepresent regions with activated stains, which appear less frequently compared to non-activated regions, thereby causing inaccuracies in stain generation where the discriminator fails to distinguish between genuine activated areas and false negatives produced by the model; (ii) with an increase in the diversity of output stains, the encoder might disproportionately favor certain stains, potentially distorting the learning process and hindering overall performance.

To mitigate these issues, our methodology integrates $\mathcal{L}_{\text{IHC}}$ loss functions that autonomously discern stain-specific attributes (see Figure 3.11), and adaptively adjusts the loss functions to accentuate underrepresented activated regions (refer to Section 3.4.1, Figure 3.6 and Figure 3.5). This adjustment aims to minimize errors and reduce false staining. Initially, stain-activated regions are identified and subsequently employed to spatially modify the loss functions, thereby focusing more on pertinent tissue sections.

Additionally, we implement an H&E regularization, $\mathcal{L}_{\text{H\&E}}$, to ensure balanced attention across diverse stains, recalibrating the model to uniformly consider all stains by backpropagating the average error across the H&E components exclusively.

This strategic integration not only consolidates the training process but also demonstrates scalability, as evidenced by enhanced results documented in Table 3.5. By combining $\mathcal{L}_{\text{IHC}}$ with $\mathcal{L}_{\text{H\&E}}$, the model's performance in both paired and unpaired configurations is significantly improved, ensuring consistent attention across various stains and markedly increasing overall effectiveness.

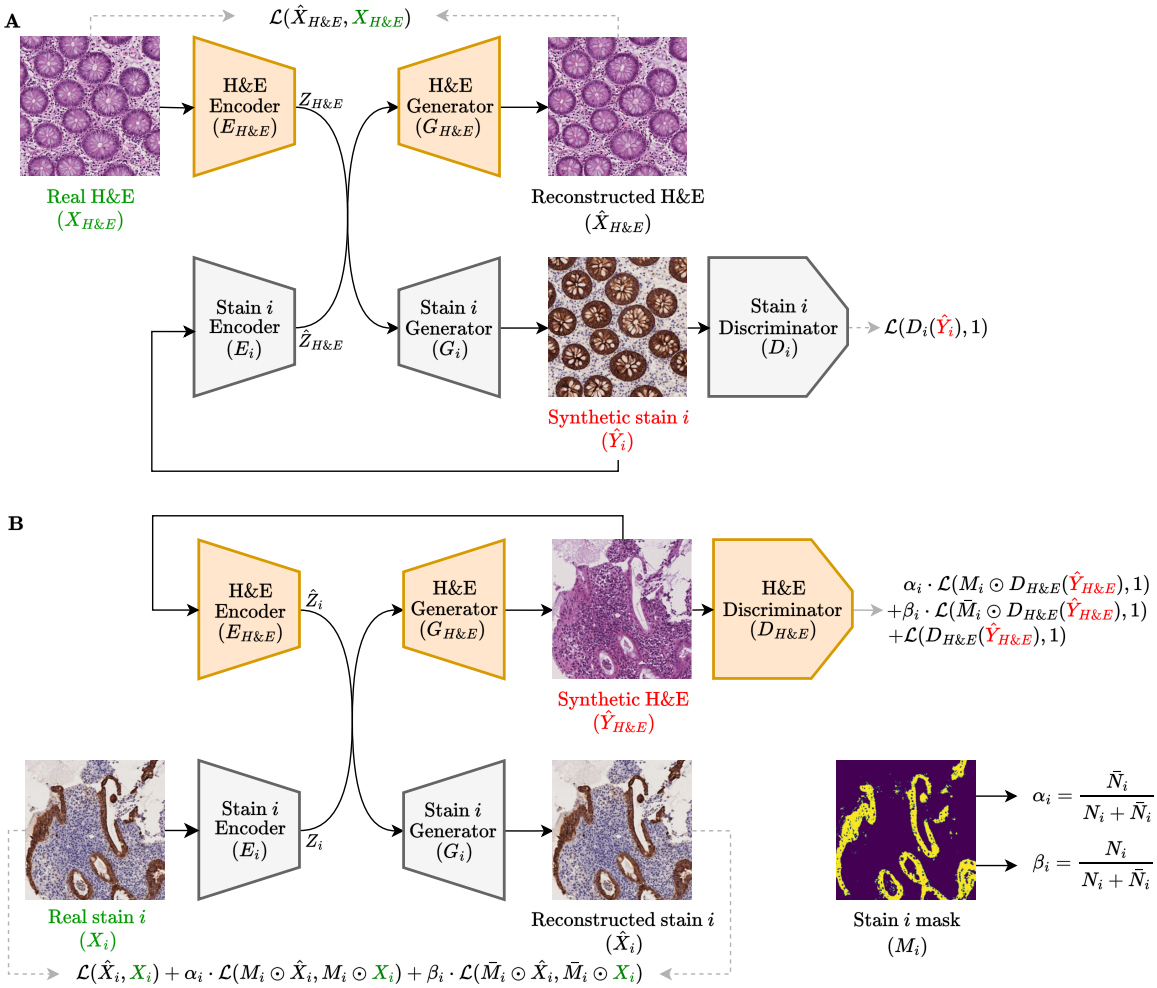


FIGURE 3.6: **An Overview of the Training Mechanism for Unpaired Stain Synthesis and Loss Function Computation in H&E $\leftrightarrow$ Stain i Conversion. A.** This part elucidates the initial training cycle, commencing with an authentic H&E image $X_{H\&E}$, proceeding to generate a synthetic stain i image $\hat{Y}_i$, and culminating with the reconstructed H&E image $\hat{X}_{H\&E}$. This progression is essential for the computation of the loss function components. **B.** This part depicts the subsequent training cycle, initiating with a genuine stain i image X_i, leading to the creation of a synthetic H&E image $\hat{Y}_{H\&E}$, and ending with the reconstructed stain i image $\hat{X}_i$, integrating the staining mask M_i (with $\bar{M}_i$ representing the complementary mask of M_i). This setup facilitates the computation of various elements of the loss function, as detailed in Section 3.4.1. Each panel underscores the model's strategic modifications and refinements, designed to target and enhance underrepresented activated regions, thereby ensuring more precise and consistent stain synthesis and discrimination.

	Synthesis loss functions		Metrics		
Setting	$\mathcal{L}_{\text{IHC}}$	$\mathcal{L}_{\text{H\&E}}$	MSE($\times 10^{-2}$)↓	PSNR↑	SSIM (%)↑
Paired			2.413±0.951	21.22±2.51	83.60±7.7
		✓	1.768±0.572	22.51±2.49	85.10±7.1
	✓		1.673±0.579	22.78±2.51	86.38±6.8
	✓	✓	**1.658±0.586**	**22.85±2.60**	**86.37±6.9**
Unpaired			2.451±0.903	21.10±2.31	83.59±7.5
		✓	2.921±0.748	20.21±1.62	79.50±7.8
	✓		2.413±0.892	21.17±2.31	83.68±7.6
	✓	✓	**2.329±0.871**	**21.34±2.39**	**83.79±7.6**

TABLE 3.5: **Evaluating the Effectiveness of $\mathcal{L}_{\text{IHC}}$ Loss Functions and $\mathcal{L}_{\text{H\&E}}$ Regularization on Stain Synthesis Quality.** This comparison delineates the outcomes for both paired and unpaired staining configurations, assessed through metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) on the Crohn-dataset. For details on methodology and replication, please see Section B.2.2.

3.3.2 Regularization impact on unpaired multi-virtual staining quality

Most style transfer methods emphasize the importance of regularization to enhance and stabilize the training process. For instance, the regularization via identity mapping loss is pivotal for retaining the color integrity in input paintings within artistic contexts, as demonstrated by the CycleGAN framework (J. Zhu et al., 2017). Similarly, the forward loss regularization plays a critical role in virtual staining applications by preserving morphological features when transitioning from H&E staining to other types, a feature central to the UMDST model (Lin et al., 2022). However, despite the diversity of these techniques, comprehensive ablation studies assessing their effectiveness in virtual staining remain understudied.

In our study, detailed in Table 3.6, we perform an extensive ablation analysis to assess both individual and combined impacts of various regularization techniques on virtual stain synthesis quality. This investigation explores multiple configurations of synthesis loss functions and regularization methods to determine the most effective setups. The evaluation metrics employed, including MSE, PSNR, and SSIM, measure error, quality, and visual similarity of the synthesized images, respectively. Our findings, presented in Table 3.6, offer a granular analysis on how different combinations of loss functions–specifically identity loss $\mathcal{L}_{\text{idt}}$, latent loss $\mathcal{L}_{\text{lat}}$, and forward loss $\mathcal{L}_{\text{fwd}}$ (refer to Section 3.4.1.2)–influence key performance indicators such as MSE, PSNR, and SSIM. Each row in the table delineates the effects of these loss functions on the assessment metrics, providing critical insights into the efficacy of each configuration.

Notably, applying the forward loss $\mathcal{L}_{\text{fwd}}$ alone yields superior outcomes compared to baseline approaches that include or exclude the combination of $\mathcal{L}_{\text{IHC}}$ and $\mathcal{L}_{\text{H\&E}}$. The $\mathcal{L}_{\text{fwd}}$ proves particularly efficacious in conserving the morphological attributes highlighted by the H&E staining, vital for precise virtual staining.

Furthermore, the optimal performance is achieved when $\mathcal{L}_{\text{fwd}}$ is combined with $\mathcal{L}_{\text{idt}}$. This synergy not only preserves the morphological integrity of the stains but also retains the original features of the input images, thereby ensuring high fidelity in the virtual staining process. This finding underscores the value of integrating both forward and identity losses as a robust approach to enhance the quality and accuracy of the synthesized stains, especially beneficial for applications demanding high precision in unpaired virtual staining.

Conversely, incorporating latent loss $\mathcal{L}_{\text{lat}}$ in the tested combinations does not positively impact staining outcomes. In fact, setups including $\mathcal{L}_{\text{lat}}$ consistently underperform across all metrics compared to those excluding it. This observation suggests that latent loss might disrupt the preservation of essential staining characteristics, specific to each stain, thus rendering it less suitable for virtual staining applications where accuracy and fidelity are paramount.

Synthesis loss functions		Regularization			Metrics		
$\mathcal{L}_{\text{IHC}}$	$\mathcal{L}_{\text{H\&E}}$	$\mathcal{L}_{\text{idt}}$	$\mathcal{L}_{\text{lat}}$	$\mathcal{L}_{\text{fwd}}$	MSE($\times 10^{-2}$)↓	PSNR↑	SSIM (%)↑
					2.451±0.903	21.10±2.31	83.59±7.5
✓	✓				2.329±0.871	21.34±2.39	83.79±7.6
✓	✓	✓			2.378±0.850	21.23±2.32	83.93±7.5
✓	✓		✓		2.477±0.940	21.08±2.38	83.32±7.7
✓	✓			✓	2.258±0.845	21.50±2.50	84.28±7.6
✓	✓	✓	✓		2.459±0.952	21.13±2.46	83.43±7.6
✓	✓	✓		✓	**2.244±0.797**	**21.51±2.45**	**84.31±7.6**
✓	✓		✓	✓	2.315±0.897	21.41±2.51	84.19±7.5
✓	✓	✓	✓	✓	2.488±0.882	21.03±2.23	83.23±7.6

TABLE 3.6: **Effects of Various Regularization Techniques on Unpaired Virtual Staining Performance.** This table showcases an ablation study exploring different combinations of synthesis loss functions—$\mathcal{L}_{\text{IHC}}$ and $\mathcal{L}_{\text{H\&E}}$ (discussed in Sections 3.4.1.1 and 3.4.1.2), and regularization methods–$\mathcal{L}_{\text{idt}}$, $\mathcal{L}_{\text{lat}}$, and $\mathcal{L}_{\text{fwd}}$ (outlined in Section 3.4.1.2). It presents the impact of these configurations on performance metrics such as MSE, PSNR, and SSIM on the Crohn-dataset. Each row in the table corresponds to a unique configuration of loss functions, highlighting their effects on the accuracy and quality of virtual staining results. For details on reproducibility, see Section B.2.4.

3.3.3 Enhancing trustworthiness: self-inspection for anomaly detection

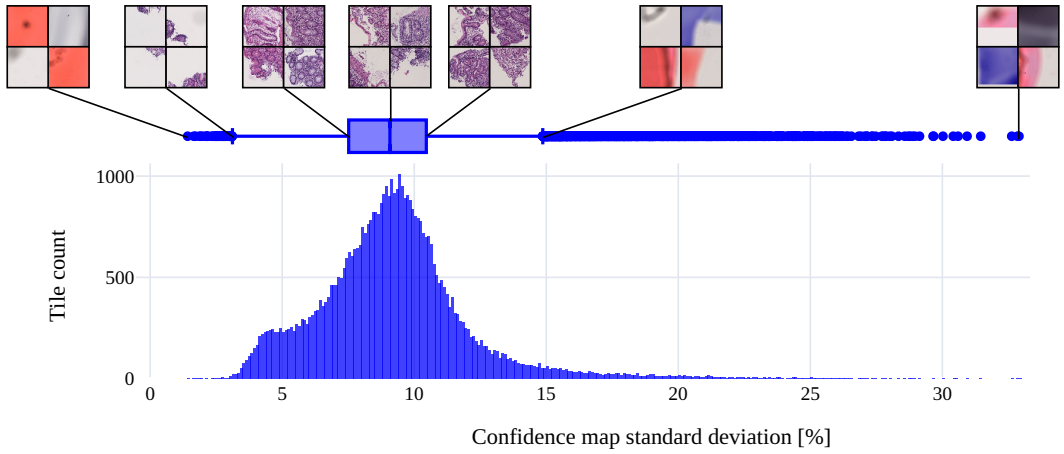


FIGURE 3.7: **Discriminator Confidence Analysis for Anomaly Detection in H&E-Stained Tiles Across Multiple Scanners.** This figure presents the evaluation of the authenticity of 47984 H&E-stained tiles derived from 2022 authentic WSIs, which were stained over a 20-year period using various scanners. Discriminator confidence maps assess the authenticity of each tile, using the standard deviation of the map values. A histogram illustrates the acceptable range for H&E staining authenticity, defined empirically between 3.11% and 14.86%. Tiles falling within this range are considered highly authentic, while those outside are flagged as outliers. Such outliers are typically either background or significantly degraded tiles, characterized by unusually high or low deviations in confidence levels. These results underscore the discriminator's ability to detect and quantify tile authenticity, providing pathologists with a crucial tool for excluding unreliable artifacts in the H&E staining and scanning processes. This method enhances the quality control within the multi-virtual staining pipeline, effectively minimizing potential errors in synthetic stains and improving the reliability and accuracy of the resulting images. For details on reproducibility, refer to Section 3.4.2.

A significant challenge in applying generative models, especially in healthcare, is their lack of a confidence score. This limitation prompts several critical questions: How can we identify issues in the input H&E data? What is the model's confidence in its virtual stains? How do we evaluate the synthetic stains' quality and spot potential errors that could affect a pathologist's reliance on virtual stains or decision to seek traditional chemical stains for confirmation?

These concerns are crucial in contexts where high-stakes decisions are made. It is imperative to develop interpretable methods to ensure the reliability of these advanced generative techniques. In this study, we employ knowledge-guided training to not only improve control during the learning process—which has shown to enhance performance as discussed in Section 3.3.1—but also to furnish pathologists with a narrative that they can interpret. Our method focuses on stain masks, directing attention to medically relevant features for better clarity than what a purely black-box model would provide.

Furthermore, we exploit the discriminator's knowledge acquired during training—a resource typically discarded after training—to assess the authenticity of images. This novel use supports the inspection of data quality and deviations from the learned distribution.

To validate our approach, we processed H&E tiles and assessed global degradation due to factors like incorrect stain concentration or scanner settings, as depicted in Figure 3.8. The discriminator successfully identifies domain shifts in these images, aligning with an anomaly detection framework by marking deviations in red (see Figure 3.8). These findings corroborate our hypothesis that the discriminator can detect newly emerging defects in H&E tiles.

An evaluation was conducted using 2022 authentic WSIs from a private dataset, which included 47,984 tiles of 512x512 H&E stained tiles. Given the possibility of both local and global anomalies, we utilized the standard deviation of the discriminator's confidence map as a diagnostic tool, as illustrated in Figure 3.7. This analysis not only verifies the discriminator's ability to identify outliers, such as artifact-laden tiles and predominantly background tiles, but also supports the empirical establishment of a confidence interval, specifically $3.11\% < \text{acceptable} < 14.86\%$. This technique introduces an effective filter to prevent the incorporation of substandard H&E images into the multi-virtual staining process, thereby minimizing potential errors in synthetic stains and increasing the reliability and credibility of the outcomes.

Furthermore, we applied the discriminator's confidence maps to the generated virtual stains to create pixel-wise confidence scores. These scores provide pathologists with crucial insights by highlighting areas where the virtual staining diverges from the expected stain appearance. This functionality serves as an additional verification layer in the output stage of our pipeline and is visually depicted through heat maps in Figure 3.9. The figure contrasts the discriminator's responses to identical tissue sections—one stained authentically and the other exhibiting a staining error in the virtual WSI. Remarkably, the discriminator identifies discrepancies in staining, highlighted in red, which corresponds to the actual differences observed between the authentic and virtual images. This approach enables the provision of dependable confidence scores to pathologists, offering further context for evaluating the importance of specific regions for particular applications. It also determines the necessity of performing a chemical stain to verify results, thereby reducing uncertainties. By clearly marking areas of doubt in the output, this tool enhances trust in virtual staining technologies, thereby ensuring greater reliability and boosting overall confidence in the results.

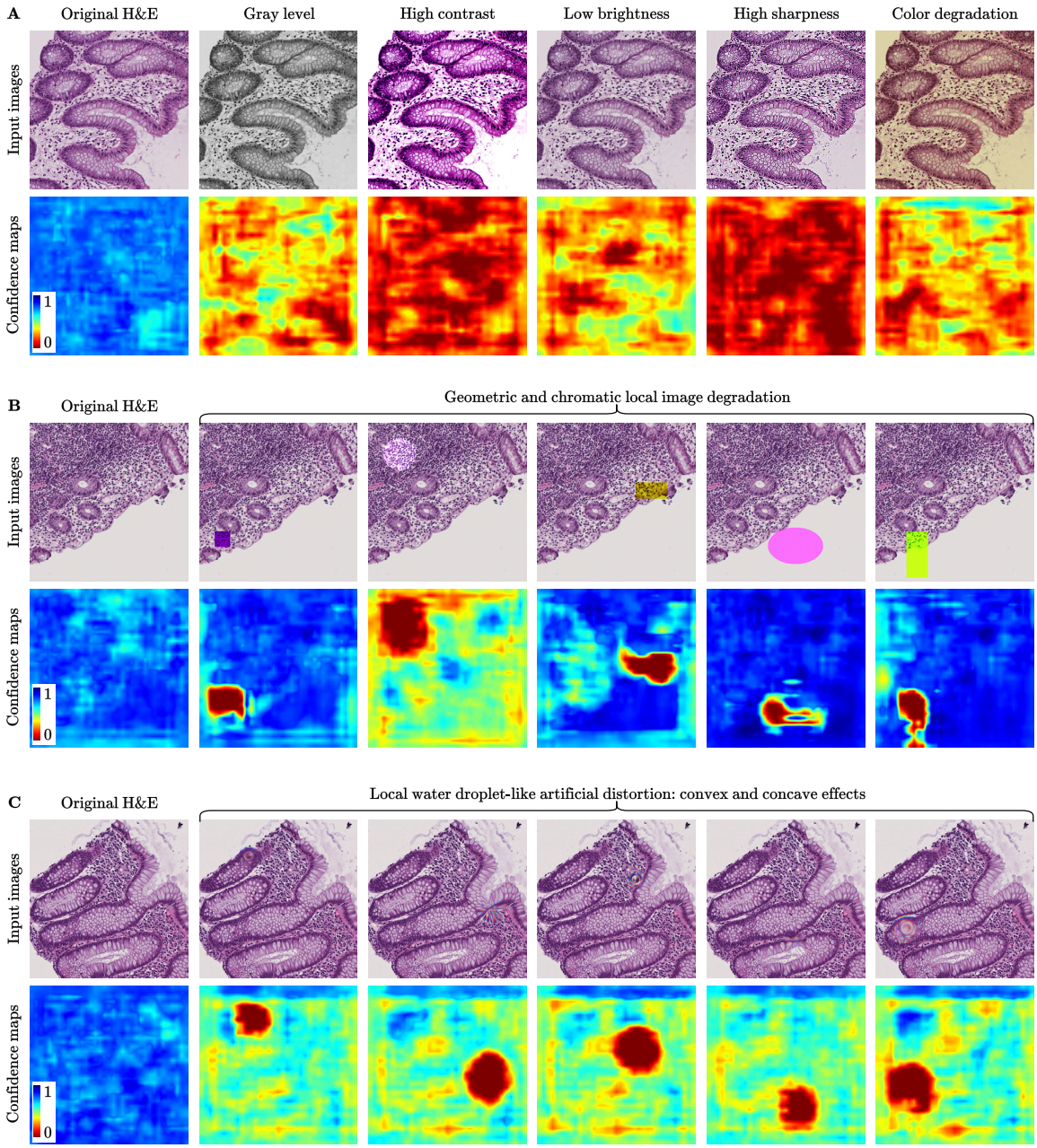


FIGURE 3.8: **Comparative Analysis of Original vs. Degraded H&E Stained Tiles with Discriminator Confidence Mapping**: Panels **A**, **B**, and **C** showcase the analysis of H&E-stained tiles. Each panel consists of two rows; the upper row presents the original H&E tile next to its five degraded variants, and the lower row displays the discriminator's confidence maps identifying areas of perceptual inconsistencies highlighted in red. Panel **A** focuses on global degradation likely stemming from chemical staining or scanning mishaps, like imprecise staining concentrations or scanner setting errors, with the model effectively detecting these widespread issues. Panel **B** illustrates local imperfections, possibly from staining faults or physical anomalies on the scanner glass, with precise identification by the model. Panel **C** reveals artifacts resembling water droplets, possibly sticking to slides during preparation and causing analytical errors, where the model marks the droplet locations, drawing attention to these critical areas. For further reproducibility information, see Section 3.4.2.

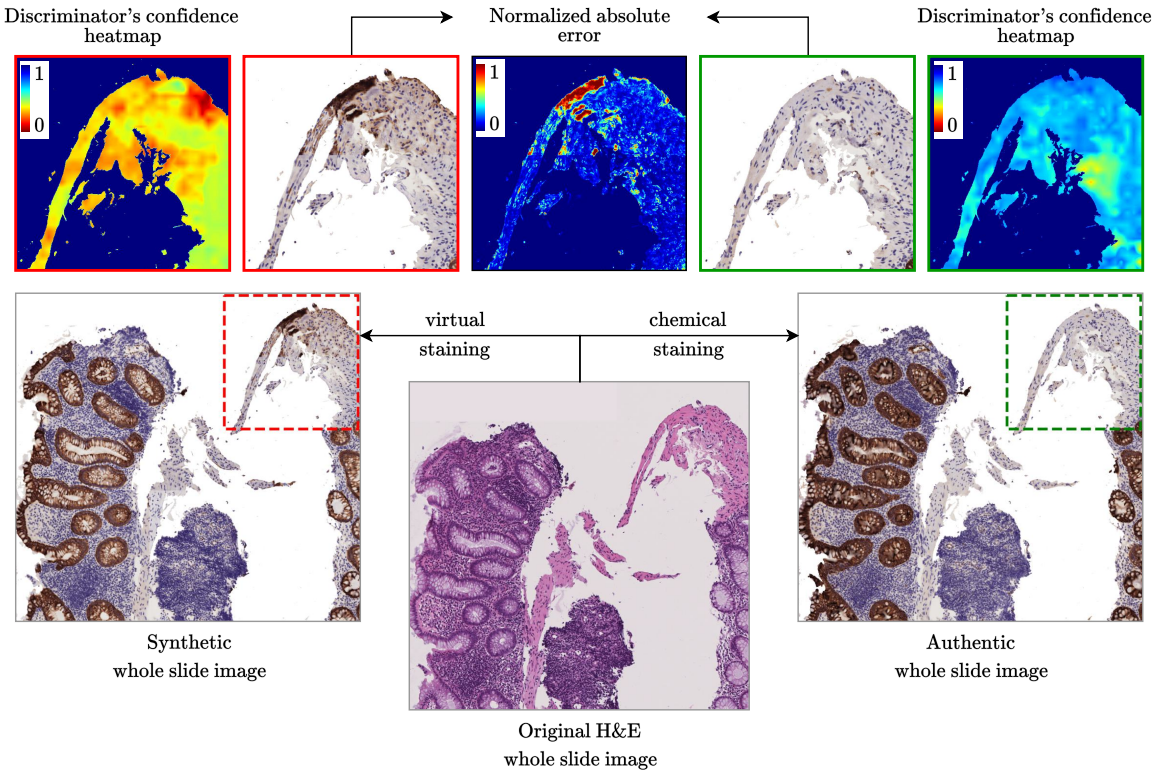


FIGURE 3.9: **Discriminator Confidence Visualization in Virtual Staining Analysis.** The efficacy of using discriminator confidence maps to assess both virtual and genuine stained WSIs is depicted in this figure. It presents two sections of tissue: one with genuine staining and another with virtual staining, wherein an error is clearly evident. The response of the discriminator is represented using heat maps, which highlight areas of inconsistency in red. These areas indicate substantial deviations from the anticipated staining pattern, offering pathologists a pixel-level confidence measure. Such visual aids are crucial for deciding whether additional chemical staining confirmation is required and for pinpointing areas needing detailed scrutiny. By accurately depicting errors, this tool enhances the trust in virtual staining technologies and assists pathologists in making informed decisions. Refer to Section 3.4.2 for details on reproducibility.

This study underscores the efficacy of incorporating discriminator confidence maps into the digital and virtual staining workflow in pathology, as illustrated in Figure 3.1. By enabling the identification of discrepancies and artifacts at both the input and output stages, our methodology ensures the utilization and generation of only high-quality, reliable data. This approach effectively addresses the pivotal concern of "garbage in, garbage out" in medical imaging.

3.4 XAI Methodology for Multi-Virtual Staining

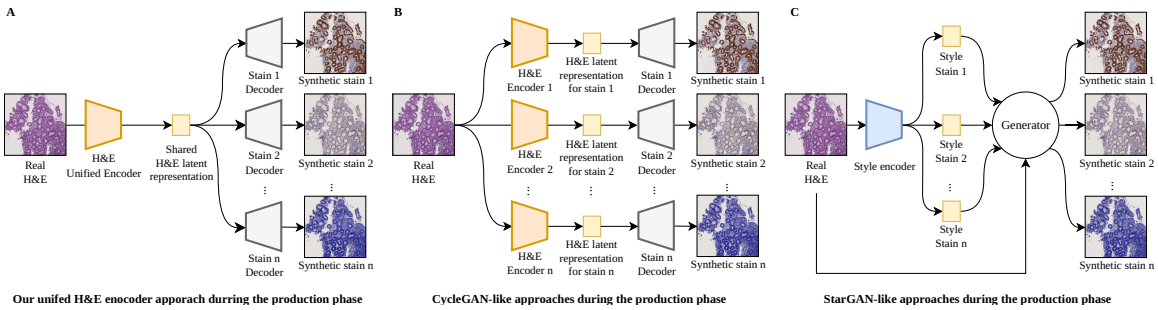


FIGURE 3.10: **H&E Staining-Based Methods for Virtual Stain Generation in Computational Histopathology during the Production Phase.** Panel **A** introduces the unified H&E encoder strategy, adapting the ComboGAN model (Anoosheh et al., 2017) for virtual staining. This method utilizes a single encoder along with multiple decoders to create various synthetic stains, enhancing computational efficiency and scalability (for detailed comparisons on XAI capabilities, refer to Figure 3.1). Panel **B** displays the conventional methodologies akin to CycleGAN (Goodfellow, Pouget-Abadie, et al., 2014; J. Zhu et al., 2017), employing multiple distinct encoders and decoders for each stain type, which increases both model complexity and computational demands. Panel **C** illustrates the methods similar to StarGAN (Y. Choi, M. Choi, et al., 2017; Y. Choi, Uh, et al., 2019; Lin et al., 2022; R. Zhang et al., 2022), incorporating a style encoder and a single generator capable of handling multiple stains. Although this architecture streamlines the model, it demands significant computational power and struggles to scale effectively with the increase in the number of stains, necessitating the maintenance of a large generator even for processing a subset of stains, which introduces inefficiencies. The approach presented in panel **A** marks a notable improvement by reducing the reliance on multiple models, thereby enabling faster and more efficient processing. This model is capable of generating only the necessary stains and loads minimal components into memory, thus minimizing hardware requirements and computational expenses in cloud-based implementations.

Our approach combines advanced deep learning techniques with cutting-edge computational pathology strategies. This integration aims to improve the scalability, accuracy, reliability, and practical application of virtual stain transformations. We will detail the specific deep learning models utilized, and the training methodologies in the subsequent sections.

As highlighted in Section 3.1.1, addressing the inherent limitations of virtual staining is crucial. Our methodology adheres to several core principles. Central to our approach is the development of a unified H&E encoder, designed to support multiple staining generators. This configuration significantly enhances the H&E encoder's ability to accurately delineate critical morphological areas within the tissue samples. As a result, this architecture not only enriches the latent feature representation but also mitigates the risk of overfitting by distributing the learning task across a denser network, thereby boosting overall performance.

Our method improves performance and trust during training by utilizing computationally activated regions via IHC. By integrating knowledge from stains autonomously and focusing on specialized loss functions and regularization tailored to stain generation, we achieve medically precise stains, enhancing the training process's credibility.

A key aspect of our approach is the use of discriminators in production to maintain quality across two primary dimensions. First, the discriminators perform a pixel-wise evaluation of the input H&E WSIs, identifying potential impurities in the data and providing visual feedback through heatmaps using XAI techniques. Second, they generate pixel-wise

confidence scores for the synthetic stains, which can be visualized as heatmaps using XAI methods to further ensure quality.

We propose a compartmentalized design that offers flexibility and practicality, enhancing trustworthiness. By making different model components separable, it becomes easy to integrate data quality check methods and visual XAI tools during production. During deployment, only the necessary parts of the model need to be loaded based on a pathologist's specific stain requirements, eliminating the need to load the entire model. Furthermore, if a new stain type is added to the dataset, our design allows for efficient training of this addition alone, avoiding the need for comprehensive model retraining.

3.4.1 Architecture and training methodologies

Our study focuses on the adaptation of two significant neural network architectures, ComboGAN (Anoosheh et al., 2017) and CycleGAN (J. Zhu et al., 2017), for the specific task of transforming H&E-stained histological slides into various other stain types. The architectural framework of our approach encompasses crucial components such as an encoder E_i, a generator G_i, and a discriminator D_i, where i denotes the index corresponding to each unique stain type within the set $\{1, \ldots, S\}$. Our methodology is the deployment of shared (unique) H&E-specific encoder, generator, and discriminator across all S stains, designated as $E_{\text{H\&E}}$, $G_{\text{H\&E}}$, and $D_{\text{H\&E}}$ respectively. This implementation is intended to enhance the specificity and scalability of the virtual staining process (refer to Appendix B for more details about the validation and implementation).

Synthesis Training Protocol: Our training methodology is predicated on a bidirectional cyclic framework designed to transform and reconstruct histopathological stain patterns in tissue slides. The primary objective of this framework is to facilitate fluid conversion between H&E stained tiles and various alternative staining modalities, while preserving essential structural features throughout the transformations.

In the inaugural cycle, termed the "H&E cycle", the procedure initiates with an H&E-stained tile $X_{H\&E}$ (refer to Fig. 3.6.A). An H&E-specific encoder, designated as $E_{\text{H\&E}}$, processes this tile to encode it into a latent representation $Z_{\text{H\&E}}$. Subsequently, this latent vector is input into a stain-specific generator G_i, corresponding to the target stain type i, yielding an image $\hat{Y}_i$ that replicates the attributes of the target stain. To complete this cycle, the resultant image is encoded by another stain-specific encoder E_i to derive a new latent representation $\hat{Z}_{\text{H\&E}}$, which is then utilized by the H&E generator $G_{\text{H\&E}}$ to regenerate an H&E-stained tile $\hat{X}_{\text{H\&E}}$. This cyclical transformation, encapsulating the transition from H&E staining to target stain i and reconversion to H&E, is mathematically represented as follows:

$$\hat{Y}_i = G_i\left(E_{\text{H\&E}}\left(X_{\text{H\&E}}\right)\right), \tag{3.1}$$

$$\hat{X}_{\text{H\&E}} = G_{\text{H\&E}}\left(E_i\left(\hat{Y}_i\right)\right) \quad \forall i \in \{1, \ldots, S\} \tag{3.2}$$

In the secondary cycle, referred to as the "stain i cycle", the methodology encompasses the reverse transformation: initiating with a tile stained by a specific type i, the objective is to convert it into an H&E-stained representation and then revert it to its original stain (refer to Fig. 3.6.B). The tile X_i is first subjected to an encoding process by E_i to produce a latent representation Z_i. This latent vector is subsequently transformed by the H&E generator $G_{\text{H\&E}}$ into $\hat{Y}_{\text{H\&E}}$, an H&E-stained tile, thereby translating X_i into the H&E domain. The resultant H&E image undergoes re-encoding by $E_{\text{H\&E}}$, yielding a new latent representation $\hat{Z}_i$. This new vector serves as the input for the generator G_i, which reconstructs the original stained image $\hat{X}_i$. This cycle facilitates the bidirectional conversion between a specific stain type and an H&E representation, encapsulated mathematically as follows:

$$\hat{Y}_{\text{H\&E}} = G_{\text{H\&E}}\left(E_i\left(X_i\right)\right), \tag{3.3}$$

$$\hat{X}_i = G_i\left(E_{\text{H\&E}}\left(\hat{Y}_{\text{H\&E}}\right)\right) \quad \forall i \in \{1, \ldots, S\} \tag{3.4}$$

To optimize our model's performance, we introduce a composite global synthesis loss, denoted $\mathcal{L}_{\text{IHC},i}$, to facilitate the accurate translation between H&E-stained images and target stains designated by i. This framework involves a comparative analysis between the reconstructed image $\hat{X}_i$ and its corresponding target image X_i for each specific stain i, as well as between the reconstructed H&E image $\hat{X}_{\text{H\&E}}$ and the original H&E image $X_{\text{H\&E}}$. These comparisons contribute to the estimation of the cycle-consistency loss $\mathcal{L}_{\text{cyc},i}$, which measures the translational fidelity in a bidirectional context between the H&E stain and the specific stain i.

Moreover, our architecture incorporates an adversarial loss, $\mathcal{L}_{\text{adv},i}$, leveraging discriminators, specifically $D_{\text{H\&E}}$ for the H&E stain and D_i for each target stain i, to evaluate the authenticity of the generated images $\hat{Y}_i$ and $\hat{Y}_{\text{H\&E}}$. These discriminators are tasked with distinguishing real images from synthesized counterparts, thereby promoting the production of images that closely mimic authentic stained tissue samples. Furthermore, the integration of a regularization component, $\mathcal{L}_{\text{reg},i}$, supports the stabilization and convergence of the training process. The comprehensive synthesis loss for each specific stain i, among a total of S stains, is expressed mathematically by the following equation:

$$\mathcal{L}_{\text{IHC},i} = \lambda_{\text{cyc}} \cdot \mathcal{L}_{\text{cyc},i} + \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{adv},i} + \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg},i} \quad \forall i \in \{1, ..., S\} \tag{3.5}$$

In this context, the parameters λ_{cyc}, λ_{adv}, and λ_{reg} serve as weighting coefficients, configuring the emphasis on the respective loss components within our training framework. This methodological design fosters a flexible and robust training regime capable of accommodating a diverse array of stains, symbolically represented by S. Such a configuration enhances the model's capacity for generalization across a broad spectrum of staining patterns.

Discriminator Training Protocol: Our methodology incorporates two distinct categories of discriminators; a specific discriminator for H&E stains, labeled $D_{\text{H\&E}}$, and individual discriminators for each IHC stain, designated D_i for the i^{th} stain. Each discriminator is trained using a dedicated loss function, tailored to enhance its discriminative efficacy. For

the discriminator associated with H&E stains, $D_{\text{H\&E}}$, the corresponding loss function is articulated as follows:

$$\mathcal{L}_{D_{\text{H\&E}}} = \lambda_{\text{D}} \cdot (\mathcal{L}_{\text{real}_{\text{H\&E}}} + \mathcal{L}_{\text{synthetic}_{\text{H\&E}}}) \tag{3.6}$$

This expression consolidates the losses $\mathcal{L}_{\text{real}_{\text{H\&E}}}$ and $\mathcal{L}_{\text{synthetic}_{\text{H\&E}}}$, associated respectively with real and synthetic H&E-stained images (see Equation (3.15)). These losses are modulated by a scaling factor λ_{D}, which is used to evaluate the discriminator's efficacy in differentiating between authentic and synthetically generated H&E images.

In a parallel manner, for each specific stain discriminator D_i, the loss function is devised to gauge its capacity to distinguish between real and synthetic images corresponding to that particular stain type, defined as follows:

$$\mathcal{L}_{D_{\text{IHC},i}} = \lambda_{\text{D}} \cdot (\mathcal{L}_{\text{real},i} + \mathcal{L}_{\text{synthetic},i}) \quad \forall i \in \{1, \ldots, S\} \tag{3.7}$$

In this formulation, $\mathcal{L}_{\text{real},i}$ and $\mathcal{L}_{\text{synthetic},i}$ denote the losses incurred from real and synthetic images of the i^{th} stain, respectively (refer to Equation (3.14)). The aggregation of these losses, modulated by a weighting factor λ_{D}, constitutes the comprehensive loss for each discriminator. This ensures that each discriminator is effectively trained to distinguish between authentic and synthetically generated samples of its designated stain. This structured approach is consistently applied across all S stains, empowering the discriminators to specialize and enhance their proficiency in discerning genuine from generated images within their respective staining domains.

3.4.1.1 Annotation-free knowledge via loss function integration

In the formulation of computational models dedicated to the synthesis of IHC slides, it has become apparent that conventional metrics such as $\mathcal{L}_1$, $\mathcal{L}_2$, and MSE introduce substantial limitations. The primary challenge associated with these metrics is their uniform application across disparate regions of the slides, leading to a lack of differentiation between tissue sections and regions highlighted by IHC staining. This issue is exacerbated by the prevalent staining disparity on IHC slides, characterized by a dominance of IHC-negative areas, thus highlighting the localized nature of IHC staining which is confined to a minor portion of the slide's total area.

To mitigate these deficiencies, our methodology employs cycle consistency loss $\mathcal{L}_{\text{cyc}}$ and adversarial loss $\mathcal{L}_{\text{adv}}$ (J. Zhu et al., 2017; Anoosheh et al., 2017). Leveraging these advanced losses, we introduce an innovative technique that integrates areas activated by IHC staining into the training regime of the model for the synthesis of S stains from H&E-stained slides. This approach enables the derivation of knowledge without reliance on annotations, while effectively assimilating the unique properties of each stain into the training model. As a result, our model is capable of generating features with better fidelity, enhancing the synthesis quality significantly. By directing the synthesis process through insights specific to each stain, our method not only improve performance but also increases the trustworthiness and reliability of the generated images.

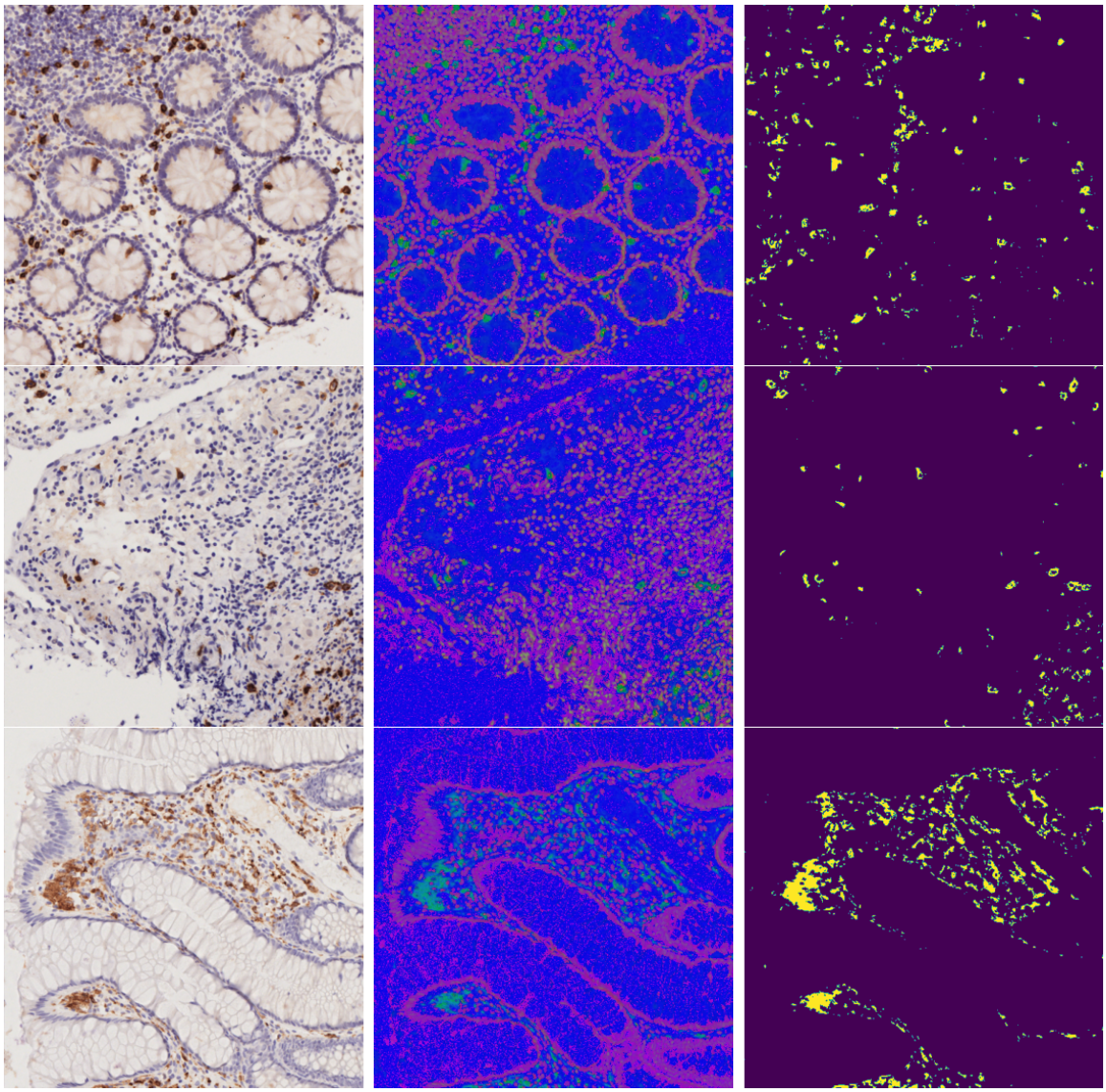


FIGURE 3.11: **Visualization of Immunohistochemical Activation and Extraction in Stained Tissue Samples:** For each biomarker, exemplified by CD8, CD117, and CD163, the extraction workflow is delineated across a tripartite columnar display. The initial column presents the original RGB stained tile (X_i), followed by the central column illustrating the transformation into the HSV color space, which isolates the distinctive chromatic signatures resultant from antigen-antibody interactions. The terminal column exhibits the derived binary mask (M_i), accentuated in yellow, depicting the areas of activation.

To facilitate the automated delineation of a mask, designated M_i, from the i^{th} IHC stained image, X_i, we initiate the process by converting X_i from its native RGB color space to the HSV color space. This transformation enhances the segregation of target regions by leveraging their chromatic and luminance properties. Following the transformation to the HSV space, we implement a thresholding procedure to delineate a clear mask, M_i, as depicted in Fig.3.11. The mask M_i is subsequently utilized in the derivation of dynamic weighting coefficients, α_i and β_i, which are incorporated into the formulation of the loss function. These coefficients are mathematically expressed as follows:

$$\alpha_i = \frac{\bar{N}_i}{N_i + \bar{N}_i}, \quad \beta_i = \frac{N_i}{N_i + \bar{N}_i} \tag{3.8}$$

In this formulation, N_i denotes the aggregate pixel count within the foreground, aligning with the IHC-activated regions, while $\bar{N}_i$ signifies the comprehensive pixel count within the background, or the IHC-non-activated regions, of the i^{th} immunohistochemically stained image. This methodology yields a refined distinction between the regions of interest and their corresponding background, thereby enhancing the precision of subsequent analyses.

Enhancing Synthesis Training with Integrated Knowledge (cycle loss): In instances where direct correspondences between H&E-stained and IHC images are absent, the synthesis is conducted within an *unpaired setting*. This approach incorporates a tailored cycle loss, denoted as $\mathcal{L}_{\text{cyc},i}$, which is instrumental in facilitating the synthesis process in the absence of directly paired images. The formulation of the cycle loss for the unpaired setting is expressed as follows:

$$\begin{aligned}\mathcal{L}_{\text{cyc},i} &= \mathcal{L}(\hat{X}_i, X_i) + \mathcal{L}(\hat{X}_{\text{H\&E}}, X_{\text{H\&E}}) \\ &\quad + \alpha_i \cdot \mathcal{L}(M_i \odot \hat{X}_i, M_i \odot X_i) + \beta_i \cdot \mathcal{L}(\bar{M}_i \odot \hat{X}_i, \bar{M}_i \odot X_i) \quad \forall i \in \{1, \ldots, S\}\end{aligned} \tag{3.9}$$

This methodology rigorously delineates between IHC-activated regions (M_i) and non-activated regions ($\bar{M}_i$), thereby preserving the fidelity of the synthesis process in the absence of direct image correspondences (refer to Fig.3.6).

In contrast, under *paired settings* where direct correspondences between H&E and IHC images are established, the cycle loss is strategically formulated. This ensures not only the global fidelity of image reconstructions but also the precise replication of designated IHC-activated regions, leveraging the available direct correspondence (refer to Fig.3.5). The detailed equation governing the paired setting is as follows:

$$\begin{aligned}\mathcal{L}_{\text{cyc},i} &= \mathcal{L}(\hat{X}_i, X_i) + \mathcal{L}(\hat{X}_{\text{H\&E}}, X_{\text{H\&E}}) \\ &\quad + \alpha_i \cdot [\mathcal{L}(M_i \odot \hat{X}_i, M_i \odot X_i) + \mathcal{L}(M_i \odot \hat{X}_{\text{H\&E}}, M_i \odot X_{\text{H\&E}})] \\ &\quad + \beta_i \cdot [\mathcal{L}(\bar{M}_i \odot \hat{X}_i, \bar{M}_i \odot X_i) + \mathcal{L}(\bar{M}_i \odot \hat{X}_{\text{H\&E}}, \bar{M}_i \odot X_{\text{H\&E}})] \quad \forall i \in \{1, \ldots, S\}\end{aligned} \tag{3.10}$$

In both scenarios, dynamic weighting factors, α_i and β_i, are employed, which are determined by the ratio of IHC-activated to non-activated regions within each image. This calibration ensures that the training of the model prioritizes not only the overall accuracy of the stain transformation but also the precise reproduction of regions essential for IHC analysis. Additionally, the methodology integrates masks M_i and $\bar{M}_i$ in the computation of the cycle loss. This enhances the model's ability to incorporate distinct stain characteristics directly into its architecture, thereby facilitating the incorporation of annotation-free knowledge. Consequently, this rigorous approach leads to the generation of high-quality, reliable synthetic images that authentically replicate the intricacies of IHC staining.

Enhancing Synthesis Training with Integrated Knowledge (adversarial loss): In the unpaired configuration, the adversarial loss, designated as $\mathcal{L}_{\text{adv},i}$, is computed by assessing the authenticity of the generated images, $\hat{Y}_i$ and $\hat{Y}_{H\&E}$, utilizing discriminators D_i

and $D_{\text{H\&E}}$ correspondingly. This evaluation process is augmented by the application of the stain mask M_i, which enables $D_{\text{H\&E}}$ to specifically target regions activated by the IHC, as illustrated in Fig.3.6. The adversarial loss in this unpaired setting is calculated using the following formulation:

$$\begin{aligned}\mathcal{L}_{\text{adv},i} = \mathcal{L}(D_i(\hat{Y}_i), 1) \\ + \alpha_i \cdot \mathcal{L}(M_i \odot D_{\text{H\&E}}(\hat{Y}_{\text{H\&E}}), 1) \\ + \beta_i \cdot \mathcal{L}(\bar{M}_i \odot D_{\text{H\&E}}(\hat{Y}_{\text{H\&E}}), 1) \quad \forall i \in \{1, \ldots, S\}\end{aligned} \tag{3.11}$$

In the paired setting, the methodology aligns with that of the unpaired setting but includes an explicit emphasis on the direct correspondence between the $H\&E$ stain and image i, as depicted in Fig.3.5. The calculation of adversarial loss in paired configurations accounts for this correspondence and is articulated by the subsequent equation:

$$\begin{aligned}\mathcal{L}_{\text{adv},i} = \mathcal{L}(D_{\text{H\&E}}(\hat{Y}_{\text{H\&E}}), 1) \\ + \alpha_i \cdot [\mathcal{L}(M_i \odot D_{\text{H\&E}}(\hat{Y}_{\text{H\&E}}), 1) + \mathcal{L}(M_i \odot D_i(\hat{Y}_i), 1)] \\ + \beta_i \cdot [\mathcal{L}(\bar{M}_i \odot D_{\text{H\&E}}(\hat{Y}_{\text{H\&E}}), 1) + \mathcal{L}(\bar{M}_i \odot D_i(\hat{Y}_i), 1)] \quad \forall i \in \{1, \ldots, S\}\end{aligned} \tag{3.12}$$

Direct Supervision Loss (only paired setting): In the paired setting of our framework, the synthesis loss is meticulously designed to include a variety of components, notably the supervised loss ($\mathcal{L}_{\text{sup},i}$), alongside the cycle consistency loss ($\mathcal{L}_{\text{cyc},i}$) as referenced in equation (3.10) and the adversarial loss ($\mathcal{L}_{\text{adv},i}$) as detailed in equation (3.12). Crucially, the supervised loss establishes a direct connection between the H&E cycle and the specific stain cycle i. It does this by evaluating the fidelity of the generated stains $\hat{Y}_i$ and H&E images $\hat{Y}_{\text{H\&E}}$ against their actual counterparts (X_i and $X_{\text{H\&E}}$, respectively). Furthermore, it incorporates a common mask (M_i), derived from X_i, to concentrate the loss computation on pertinent areas of the image. This ensures that the generated images maintain both structural and stylistic integrity in relation to the original samples. The formula for the supervised loss is articulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{sup},i} = \mathcal{L}(\hat{Y}_i, X_i) + \mathcal{L}(\hat{Y}_{\text{H\&E}}, X_{\text{H\&E}}) \\ + \alpha_i \cdot [\mathcal{L}(M_i \odot \hat{Y}_i, M_i \odot X_i) + \mathcal{L}(M_i \odot \hat{Y}_{\text{H\&E}}, M_i \odot X_{\text{H\&E}})] \\ + \beta_i \cdot [\mathcal{L}(\bar{M}_i \odot \hat{Y}_i, \bar{M}_i \odot X_i) + \mathcal{L}(\bar{M}_i \odot \hat{Y}_{\text{H\&E}}, \bar{M}_i \odot X_{\text{H\&E}})] \quad \forall i \in \{1, \ldots, S\}\end{aligned} \tag{3.13}$$

Integrating Knowledge During the Discriminator Training Phase: To optimize the training of the D_i discriminator within our model, authentic samples denoted as X_i are utilized to generate corresponding stain masks M_i. These masks facilitate the discriminator's ability to detect subtle variations within the IHC-activated regions via the $\mathcal{L}_{\text{real}_i}$ loss function. Simultaneously, the discriminator is trained to identify synthetic images $\hat{Y}_i$ as non-authentic by employing the $\mathcal{L}_{\text{synthetic}_i}$ loss function. The mathematical expressions for both loss functions are specified as follows:

$$\begin{cases} \mathcal{L}_{\text{real}_i} = \mathcal{L}(D_i(X_i), 1) + \alpha \cdot \mathcal{L}(M_i \odot D_i(X_i), 1) + \beta \cdot \mathcal{L}(\bar{M}_i \odot D_i(X_i), 1) \\ \mathcal{L}_{\text{synthetic}_i} = \mathcal{L}(D_i(\hat{Y}_i), 0) \end{cases} \tag{3.14}$$

Similarly, for the $D_{\text{H\&E}}$ discriminator, the loss functions $\mathcal{L}_{\text{real}_{\text{H\&E}}}$ and $\mathcal{L}_{\text{synthetic}_{\text{H\&E}}}$ are formulated to assess the authenticity of H&E stained images and to identify their synthetic analogues, respectively. The definitions of these loss functions are delineated as follows:

$$\begin{cases} \mathcal{L}_{\text{real}_{\text{H\&E}}} = \mathcal{L}(D_{\text{H\&E}}(X_{\text{H\&E}}), 1) \\ \mathcal{L}_{\text{synthetic}_{\text{H\&E}}} = \mathcal{L}(D_{\text{H\&E}}(\hat{Y}_{\text{H\&E}}), 0) \end{cases} \tag{3.15}$$

3.4.1.2 Enhancing training using regularization

A Universal H&E Representation for Governing Multiple Staining Modalities:

The primary goal of our study is to design a unique H&E encoder and generator, capable of managing various staining modalities. Traditional techniques encounter significant challenges, primarily due to the diverse demands associated with different stains during the training phase. Notably, some stains present inherent complexities that can result in inconsistent learning progress and potential oversight of less prevalent stains.

To mitigate these issues, we have devised a novel regularization strategy that promotes uniform distribution of learning focus across all stains by the components of our H&E encoder. This methodology effectively reduces bias towards any particular stain.

The regularization process is initiated by a systematic selection mechanism, where stains are randomly chosen from a complete set, numbered from 1 to S, to ensure exhaustive inclusion. Subsequent updates are applied to the encoder E_i and generator G_i corresponding to each selected stain i. Concurrent updates are made to the shared H&E components ($E_{\text{H\&E}}$ and $G_{\text{H\&E}}$), guaranteeing fair representation of each stain throughout the training cycles. Upon completing a cycle through all S stains in a randomized order, a mean synthetic loss is calculated across the stains, as defined by the following equation:

$$\mathcal{L}_{\text{H\&E}} = \frac{1}{S} \sum_{i=1}^{S} \mathcal{L}_{\text{IHC},i} \tag{3.16}$$

The calculated loss, $\mathcal{L}_{\text{H\&E}}$, serves to refine both the H&E encoder $E_{\text{H\&E}}$ and generator $G_{\text{H\&E}}$, signifying the end of one training iteration. This approach guarantees equitable attention to each stain, thereby significantly improving the model's adaptability across various staining modalities. Consequently, this strategy leads to a more stable and scalable training process, enhancing the overall efficacy of the model in processing a diverse array of stains.

Stain Synthesis Regularization: We introduce a comprehensive methodology specifically developed for regularization in virtual staining. Our approach involves encapsulating the entire spectrum of considerations pertinent to this process. It is based on integrating essential knowledge derived from IHC-activated regions across the stain mask, represented by M_i. The core of this methodology is the computation of a regularized loss, expressed as

$\mathcal{L}_{\text{reg},i}$. This loss is formulated as a weighted sum incorporating three primary components: the identity loss ($\mathcal{L}_{\text{idt},i}$), the latent loss ($\mathcal{L}_{\text{lat},i}$), and the forward loss ($\mathcal{L}_{\text{fwd},i}$). For each stain index i within the range $\{1, ..., S\}$, the regularized loss is defined as follows:

$$\mathcal{L}_{\text{reg},i} = \lambda_{\text{idt}} \cdot \mathcal{L}_{\text{idt},i} + \lambda_{\text{lat}} \cdot \mathcal{L}_{\text{lat},i} + \lambda_{\text{fwd}} \cdot \mathcal{L}_{\text{fwd},i} \tag{3.17}$$

In this framework, the coefficients $\lambda_{\text{idt}}, \lambda_{\text{lat}}$, and λ_{fwd} are used to denote the respective weights of each loss component in the combined regularized loss formulation.

The identity loss ($\mathcal{L}_{\text{idt}}$) measures the discrepancy between the original and synthesized images within the same domain. It uses the encoder and generator from an auto-encoder configuration to ensure that the encoder captures sufficient features necessary for accurately reproducing the input image. The application of this concept is expanded to incorporate the stain mask M_i as follows:

$$\mathcal{L}_{\text{idt},i} = \begin{cases} \mathcal{L}(G_i(E_i(X_i)), X_i) & \\ \quad + \alpha \cdot \mathcal{L}(M_i \odot G_i(E_i(X_i)), M_i \odot X_i) & \text{for stain } i \text{ image } X_i, \\ \quad + \beta \cdot \mathcal{L}(\bar{M}_i \odot G_i(E_i(X_i)), \bar{M}_i \odot X_i) \qquad \forall i \in \{1, \ldots, S\}, & \\ \mathcal{L}(G_{\text{H\&E}}(E_{\text{H\&E}}(X_{\text{H\&E}})), X_{\text{H\&E}}) & \text{for H\&E image } X_{\text{H\&E}}. \end{cases} \tag{3.18}$$

The latent loss, denoted as $\mathcal{L}_{\text{lat}}$, is designed to reduce disparities within the latent space by capturing the variance between the latent representations and their reconstructed counterparts. This process aligns the embeddings from both the H&E and stain i encoders. It effectively incorporates IHC-activated regions as follows:

$$\mathcal{L}_{\text{lat},i} = \begin{cases} \mathcal{L}(\hat{Z}_i, Z_i) & \\ \quad + \alpha \cdot \mathcal{L}(M_i \odot \hat{Z}_i, M_i \odot Z_i) & \text{for stain } i \text{ embeddings } Z_i \text{ and } \hat{Z}_i, \\ \quad + \beta \cdot \mathcal{L}(\bar{M}_i \odot \hat{Z}_i, \bar{M}_i \odot Z_i) \qquad \forall i \in \{1, \ldots, S\}, & \\ \mathcal{L}(\hat{Z}_{\text{H\&E}}, Z_{\text{H\&E}}) & \text{for H\&E embeddings } Z_{\text{H\&E}} \text{ and } \hat{Z}_{\text{H\&E}}. \end{cases} \tag{3.19}$$

Finally, The forward loss evaluates the divergence between the degraded versions of the original images, represented as x_i and $x_{\text{H\&E}}$, and their corresponding outputs, denoted $\hat{y}_{\text{H\&E}}$ and $\hat{y}_i$. This relationship is specified as follows:

$$\mathcal{L}_{\text{fwd},i} = \begin{cases} \mathcal{L}(\hat{y}_{\text{H\&E}}, x_i) & \\ \quad + \alpha \cdot \mathcal{L}(m_i \odot \hat{y}_{\text{H\&E}}, m_i \odot x_i) & \text{for stain } i, \\ \quad + \beta \cdot \mathcal{L}(\bar{m}_i \odot \hat{y}_{\text{H\&E}}, \bar{m}_i \odot x_i) \qquad \forall i \in \{1, \ldots, S\}, & \\ \mathcal{L}(\hat{y}_i, x_{\text{H\&E}}) & \text{for H\&E.} \end{cases} \tag{3.20}$$

By integrating three distinct loss functions—$\mathcal{L}_{\text{idt},i}$, $\mathcal{L}_{\text{lat},i}$, and $\mathcal{L}_{\text{fwd},i}$—and capitalizing on insights from IHC-activated regions, this regularization framework significantly enhances the capacity for targeted adaptation to task-specific challenges. Specifically engineered

for unpaired scenarios, this methodology markedly advances the nuanced orchestration of associated tasks. In contrast, in paired settings, these strategies yield marginal benefits, as the direct correlations between the H&E and S stains are already comprehensively addressed through supervised loss functions. However, the theoretical potential for incorporating these regularization techniques, $\mathcal{L}_{\text{reg, i}}$, indicates a possibility for their application well beyond the initially envisaged contexts.

3.4.2 Trust in virtual stains through self-inspection–anomaly detection

To streamline the analytical process for pathologists and diminish cognitive strain, the dual confidence maps, C_{lum} and C_{rgb}, are amalgamated into a singular map through the computation of their pixel-wise minimum, resulting in C_{all} (PatchGAN discriminator "C. Li and Wand, 2016; Isola et al., 2016; J. Zhu et al., 2017" explicitly differentiate between synthetic and authentic images). This unified confidence map C_{all} undergoes normalization across a range from 0 (denoting an anomaly) to 1 (signifying authenticity), thereby facilitating the computation of various diagnostic metrics, as depicted in Figure 3.7. Further, the application of a Jet-color map via OpenCV version 4.9.0 (Bradski, 2000b) transforms C_{all} into an 8-bit unsigned integer RGB confidence map, exemplified in Figure 3.8.

This methodology yields comprehensive confidence maps that are capable of identifying a wide spectrum of anomalies, as illustrated in Figures 3.7 and 3.8. Given the uniform application of this discriminator architecture across different stains, the approach is universally applicable to all staining scenarios.

3.5 Cloud-Based Multi-Virtual Staining: Proof-of-Concept

Configuring the hardware and software for complex generative models in pathology labs is both resource-intensive and time-consuming. It requires specific technical skills that may not always be available in the fast-paced settings of these labs. An accessible system that operates through a browser could significantly improve both time efficiency and user comfort. In our study, we aim to offer a comprehensive approach to managing multi-virtual staining. To this end, we use Cytomine (Marée et al., 2016), an open-source platform, as a proof of concept for deploying our multi-virtual staining technique, as discussed in Section 3.1.2. This choice facilitates the integration of advanced DL models into everyday applications, providing clear guidelines for utilizing cloud-based, open-source platforms.

To implement this, we dockerized our multi-virtual staining method and deployed it on the platform. This allows pathologists to execute complex algorithms directly through their web browsers, as shown in Figure 3.12. Such integration streamlines the use of sophisticated DL models in routine pathological analysis, thereby enhancing the accessibility and utility of digital histopathology tools.

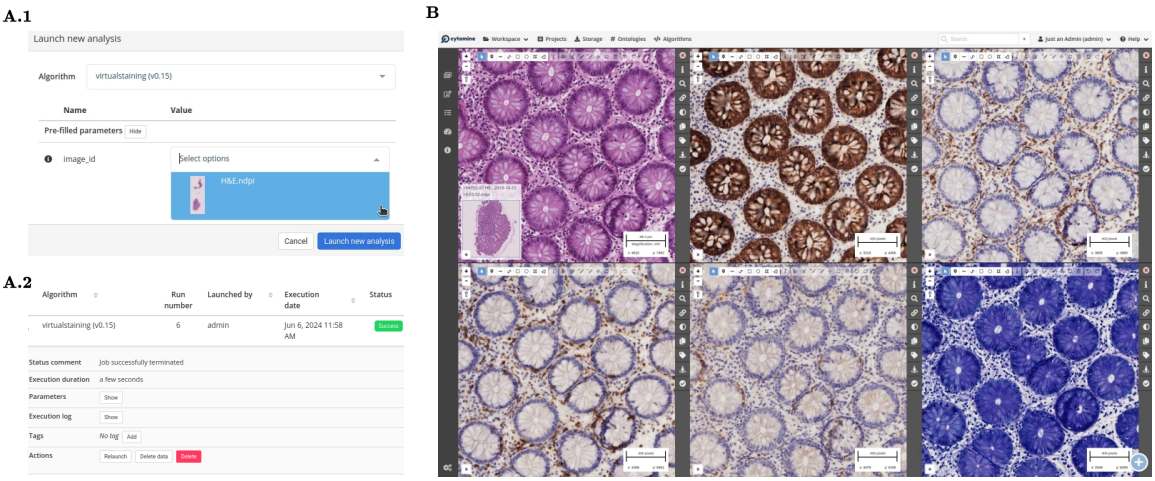


FIGURE 3.12: **Cloud-Based Multi-Virtual Staining on the Cytomine Platform: A Proof of Concept. A.1.** Showcases the user interface for selecting a H&E WSI and setting inference parameters. **A.2.** Depicts the panel that monitors the progress of the multi-virtual staining process, managed by a slurm job. **B.** Displays synchronized views of virtually stained slides next to the original H&E slide (upper left). This setup illustrates our implementation of dockerized multi-virtual staining on the open-source Cytomine platform (Marée et al., 2016). All computations occur on a backend server managed through slurm, requiring the user only to upload the H&E slide and start the algorithm via a web browser. The results are presented in a synchronized view, significantly reducing user effort.

We utilize the open-source platform, Cytomine Community Edition Legacy 3.1.0, to integrate our virtual staining model into a web application. This software employs a containerized architecture through Docker, which simplifies the creation and deployment process across various modules including applications, web UI, databases, nginx proxy, and jobs management. The primary component for deploying deep learning applications is a *software* Docker container.

Our Python-based application, which performs virtual staining, is also Dockerized. Afterward, it is uploaded to the software_router and converted into a Singularity image. This

application encompasses both the virtual staining process and the functionalities to import input WSI and to upload the generated virtual stains to the database. To facilitate these tasks, we leverage the Cytomine Python API, which ensures effective communication between the *software* container and the image database container.

The inputs for our Python-based application are specified in a JSON descriptor, which is uploaded to the *software* container. This descriptor is subsequently transformed into a user-friendly web interface, enabling users to select the input H&E WSI.

Upon launching the virtual staining algorithm, its execution is overseen by a job scheduling system utilizing SLURM. This system initiates the corresponding Singularity image with the specified inputs. Once the process completes, the resulting stains can be directly visualized within the web user interface.

Cytomine enhances the user experience by optimizing the display of multiple instances of aligned WSIs. This feature allows for simultaneous visualization and comparison of different staining results, significantly improving analytical capabilities within a unified interface. Such advancements offer substantial benefits for digital pathology and related research fields.

3.6 Pediatric Crohn's Disease Multi-Virtual Staining Dataset

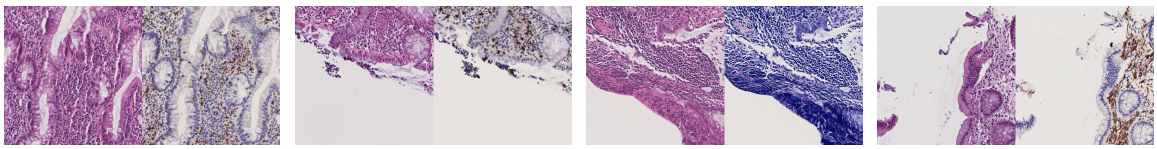
As discussed in Section 3.1.3, a significant hurdle in multi-stain data analysis is the scarcity of publicly available datasets. For example, the dataset from De Haan et al. (2021) remains unpublished (Haan et al., 2021). Moreover, the quality and alignment of paired data are often suboptimal; typically, datasets like AHNIR (Borovec et al., 2020) are assembled from adjacent slides, leading to imperfectly matched samples. Notably, the AHNIR kidney dataset includes only a limited number of slides, with five slides for each stain type: H&E, PAS, PASM, and MAS.

This problem persists across other studies as well. For instance, MVFStain (R. Zhang et al., 2022) uses only a subset of the AHNIR dataset for lung lesions, employing one WSI for training and another for testing. This approach is similarly applied to datasets concerning lung lobes and breast tissues, where two WSIs are used for training and one for testing, ensuring methodological consistency.

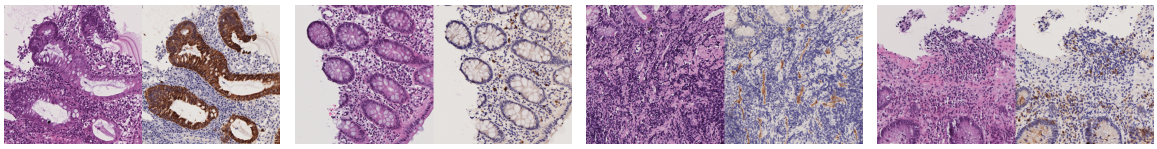
The challenges associated with the public availability and diversity of such data are compounded by the significant pairing issues due to samples derived from adjacent slides. These issues complicate the objective evaluation of computational methods and often necessitate the use of error-prone techniques like elastic registration (e.g., VALIS (Gatenbee et al., 2023)), which are susceptible to variations in tissue characteristics.

To address these limitations, we propose the creation of a new dataset that includes paired H&E to eight different stains, specifically focusing on pediatric Crohn's disease. This dataset will include 30 H&E WSIs and 30 stained WSIs across eight stains (AE1AE3, CD117, CD15, CD163, CD3, CD8, D240, and GIEMSA), totaling 480 WSIs. Each sample will consist of perfectly matched data from identical tissue sections, as illustrated in Figure 3.13. By offering a comprehensive collection of 480 high-quality, diverse WSIs, we aim to set a new benchmark for methodologies in virtual staining, segmentation, detection,

and other computational histopathology applications, thus catalyzing further research in computational pathology.



(A) H&E ↔ CD3 (B) H&E ↔ CD8 (C) H&E ↔ GIEMSA (D) H&E ↔ CD163



(E) H&E ↔ AE1AE3 (F) H&E ↔ CD117 (G) H&E ↔ D2-40 (H) H&E ↔ CD15

FIGURE 3.13: **Illustration of Perfectly Paired Samples in Our Multi-Stain Pediatric Crohn's Disease Dataset.** This figure highlights the meticulous matching of WSIs from identical tissue sections, underscoring the dataset's significance for advancements in computational pathology.

Data: This study utilizes a rigorously curated dataset critical for our research on pediatric Crohn's disease, acquired from Robert Debré Hospital in Paris. The dataset collection was approved by the INSERM ethic committee (IRB3888, ref 21-761) in March 2021, adhering to the Declaration of Helsinki principles. All patients or their legal guardians provided written informed consent.

Population: The dataset includes pediatric patients diagnosed with Crohn's disease based on the ESPGHAN criteria (European Society for Paediatric Gastroenterology Hepatology and Nutrition). These patients were monitored at Robert Debré Hospital, Paris, France for at least one year following an initial biopsy at diagnosis. The cohort spans diagnoses from 1988 to 2019, excluding any whose slides were too degraded, resulting in 59 patients with available slides. This population is predominantly male (69%) with an average age of 11.11 years (standard deviation 3.64).

Dataset Description: The dataset consists of 480 digital slides, evenly split among eight paired combinations of H&E and IHC stains. Each pairing includes 30 matched sets of an H&E slide and its corresponding IHC-stained slide, with the following markers: Anticytokeratin AE1/AE3 (AE1AE3), CD117 (c-Kit), CD15 (Lewis X or SSEA-1), CD163 (macrophage marker), CD3 (T-cell co-receptor), CD8 (T-cell co-receptor), Cluster D2-40 (D240), and Giemsa stain. All slides were scanned at a uniform 40x magnification, using the same scanner, with a resolution of 0.22μm per pixel. For experimental purposes, the slides are randomly divided into two subsets: 20% for testing and 80% for training, taking into account the tissue coverage per slide (ensuring at least 10% tissue presence in tiles).

3.7 Discussion

Our analysis of the current state-of-the-art in computational pathology has identified several critical challenges. These include the inherently opaque nature of deep learning technologies and the pervasive shortage of high-quality, publicly accessible datasets. Such limitations are considerable barriers to integrating advanced computational tools into routine clinical practice effectively.

To overcome these challenges, our study adopts a comprehensive strategy aimed at improving system performance, trustworthiness, scalability, and the quality and quantity of data. We emphasize creating user-friendly systems via secure, cloud-based platforms, essential for seamless clinical integration.

Our developed methodology offers substantial improvements to computational pathology. It enhances scalability by employing regularization and knowledge-guided techniques during the training phase. Trustworthiness is bolstered through the integration of discriminators that assess input quality and provide output confidence scoring. The practical implementation of our model, showcased in an open-source, cloud-based setup for virtual staining, indicates strong potential for real-world application.

By advancing the understanding of virtual staining, we have introduced a new dataset consisting of 480 whole slide images. This release not only establishes a benchmark for quantitative assessment in the field but also facilitates a variety of applications, including segmentation and detection. Offering these resources publicly encourages the scientific community towards conducting more reproducible research.

Looking forward, there are plans to expand this dataset to encompass a broader spectrum of pathological conditions, which will enhance the generalizability of our model. Additionally, expending to diverse interpretable deep learning architectures is anticipated to further refine system performance.

In summary, our research has significantly advanced computational pathology by integrating a unified H&E encoder, tailored loss functions, innovative regularization techniques, and context-driven learning within a cloud-based architecture. These advancements not only meet but surpass the current benchmarks for quality and trustworthiness in stain transformations. This progress paves the way for a more reliable, accessible, and efficient future in computational pathology, ultimately aiming to improve clinical outcomes.

Chapter 4

Conclusion and Future Directions

4.1 Conclusion

In conclusion, this doctoral research illustrates the pivotal dual role of XAI as both a scientific discipline and a regulatory imperative, particularly within the context of evolving AI legislative frameworks such as the GDPR discussed in Chapter 1. This convergence of methodological advancements and compliance imperatives highlights the critical importance of incorporating explainability into AI systems to meet ethical and legal standards, essential for the trustworthiness and societal acceptance of AI technologies.

The thesis further demonstrates the application of XAI through two detailed case studies. The first, discussed in Chapter 2, details the development of an interpretable deep learning framework for cell segmentation in video microscopy. This framework, named PhagoStat, specifically targets the quantification of phagocytosis in unstained, dynamic cells, crucial for advancing our understanding of neurodegenerative diseases such as FTD. PhagoStat offers a scalable and versatile solution capable of real-time analysis, significantly enriching our insight into phagocytic activity. It promotes sustainable practices by tailoring the pipeline to specific needs and reducing technological carbon footprints. Furthermore, transitioning to interpretable models has maintained high performance while reducing model size and conserving computational resources, thus enhancing the system's accessibility across diverse hardware platforms.

The second case study, presented in Chapter 3, explores the application of XAI in generative models for computational pathology. This work aims to enhance the reliability of virtual staining techniques in histopathology. By enriching training models with constraints derived from chemically stained slides and optimizing loss functions, the approach minimizes errors and mitigates false staining. The innovative use of insights from a discriminator, typically discarded post-training, facilitates robust assessments of data quality and the authenticity of synthetic stains, providing reliable confidence scores for the produced stains.

Both applications underscore that XAI not only deepens our understanding of complex models but also significantly enhances performance compared to traditional black-box approaches. For example, the application of XAI in video microscopy has revealed critical insights into the behavior of FTD mutant cells, indicating their increased size and activity relative to controls. These findings underscore the potential of XAI-enhanced pipelines to advance our understanding of both the tools and their application domains, such as neurodegenerative diseases.

Moreover, this work highlights the importance of addressing systemic biases in labeling, as evidenced by biases inherent in established methods like Free-Surfer and FSL (detailed in Section 4.2.1). Recognizing these biases is crucial, as they represent the initial link in a long chain of causality within XAI frameworks, providing users with essential context regarding the models' training and limitations.

Finally, the rapid development of large foundational models and the dynamic nature of legal and regulatory frameworks underscore the need for generic XAI methods that are adaptable across various models and tasks. In response, this thesis proposes a conceptual framework (detailed in Section 4.2.2) aimed at ensuring the effectiveness and ethical compliance of AI systems across diverse applications, upholding the principles of responsible AI throughout their lifecycle.

4.2 Perspectives

Scientific publication

Elements of the Section 4.2.1 are published in:

Valabregue, R., Khemir, I., Auzias, G., Rousseau, F., & **Ounissi, M.** (2024). Unraveling Systematic Biases in Brain Segmentation: Insights from Synthetic Training. In Medical Imaging with Deep Learning.
https://openreview.net/pdf?id=B3xO0c2Q3h

Elements of the Section 4.2.2 are published in:

Racoceanu, D, **Ounissi, M.**, Kergosien Y. L. "Explicabilité en Intelligence Artificielle ; vers une IA Responsable - Instanciation dans le domaine de la santé." (2024) Techniques de l'ingenieur, 29 Feb. 2024. https://doi.org/10.51257/a-v1-h5030.

4.2.1 Systematic Biased Labels: MRI Brain Segmentation

In (Valabregue, Khemir, et al., 2024) we explored the implications of defining "ground truth" labels utilized for the training and evaluation of segmentation models for brain MRI. Access to a large volume of anatomically precise segmentation maps is critical for the development of effective machine learning models. Recent research has introduced the concept of using pseudo ground truths–segmentation maps derived from established techniques like Freesurfer (Fischl, 2012; Henschel et al., 2020; Bontempi et al., 2020; Billot et al., 2023; W. Li, W. Huang, and Zheng, 2024)–for model training. These pseudo ground truths, however, are not devoid of errors, and the influence of such inaccuracies on model performance and generalization has been minimally addressed in existing literature.

Intuitively, systematic errors in training labels can lead supervised learning methods to replicate existing biases. To counteract this, improving the quality of training labels manually has proven beneficial, yet this method scales poorly with increasing data volumes. In our research, we investigate the efficacy of synthetic learning in overcoming these challenges. Specifically, we examine whether generating images from labels can mitigate image/label mismatches during training, thereby reducing potential biases induced by the model.

Special emphasis is placed on the selection of labeling protocols and their influence on biases in label definitions, particularly in anatomical regions with poor contrast. We concentrate our efforts on regions with clear contrast, which allows for direct validation of segmentation accuracy from the data itself. As our work on bias mitigation is ongoing, we present findings specifically related to the Putamen brain region. These results not only demonstrate the practical applications of our research but also underscore the importance of scrutinizing both labels and input data. We believe that careful consideration of these elements is crucial for developing reliable systems in the realm of XAI. This attention to detail ensures that the AI systems we develop are bias-free (over/under quantification), addressing critical challenges in the field.

4.2.1.1 Bias Mitigation: Insights from Synthetic Training

We performed an analysis of two predominant methodologies, FreeS and FSL, which generate notably different outcomes on multiple datasets. As demonstrated in Figure 4.1, the SynthFSL model excels on the MICCAI dataset compared to manual segmentation, achieving a DICE score comparable to AssN—even though AssN was trained on real data. In contrast, utilizing FreeS as ground truth markedly diminishes performance across all models, particularly evident in the HCP dataset where the performance and ranking of methods remain consistent.

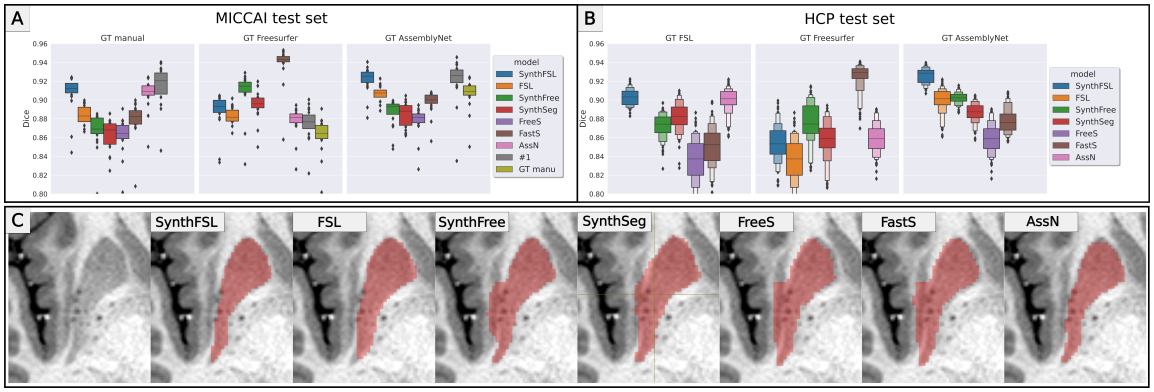


FIGURE 4.1: **Dice scores for the Putamen across various models.** Panel A displays results for 20 subjects from the MICCAI test set, while Panel B shows data for 80 subjects from the HCP test set. In each panel, the segmentation used as Ground Truth varies by column, including manual segmentation, FSL, Freesurfer, and AssemblyNet. Panel C presents the results for an axial slice from a single HCP subject. For more details about the training refer to (Valabregue, Khemir, et al., 2024).

Despite enhanced image contrast in the Putamen, large systematic errors are noted with FreeS, leading to bias reproduction in models trained with these labels. Specifically, predictions from FastS closely mirror those from FreeS, perpetuating the same bias. However, SynthFSL, although trained on synthetic data from FSL labels, yields predictions more aligned with AssN than FSL. This outcome underscores the potential of synthetic models to counteract inductive biases inherent in the input labels.

Nevertheless, the effectiveness of synthetic models is influenced by the label maps' definitions used to generate training data. Predictions from SynthFree and SynthSeg align more closely with FreeS than with SynthFSL, highlighting a dependency that may be specific to the Putamen due to significant shape alterations caused by FreeSurfer's systematic errors.

The challenge of quantifying and characterizing inductive bias in supervised learning is well-documented. Designing unbiased manual annotation datasets is both difficult and resource-intensive. Our findings advocate for the synthetic learning approach as a viable solution to mitigate these biases. However, previous studies, such as those by (Billot et al., 2023) and (Valabregue, Girka, et al., 2023), suggest that synthetic models generally underperform relative to models trained on real data, as evaluated by DICE scores. This discrepancy can be attributed in part to the systematic biases in the ground truth (GT). When using manual GT as a reference, SynthFSL performs comparably to AssN, trained on real data. Conversely, a notable difference in performance between SynthFree and FastFS reflects the indirect impact of systematic bias when FreeSurfer is used as GT.

Our research underscores the importance of refining the anatomical accuracy of labels for synthetic image generation. While our current findings focus on the Putamen structure, further studies are necessary to determine whether these insights extend to other anatomical structures. Additionally, the prevalence of systematic biases—such as those introduced by variations in imaging quality due to different acquisition settings—highlights the broader implications for automated methods. The contemporary movement towards training segmentation models on extensive multicentric datasets, using automated segmentation as GT, must carefully consider the potential for perpetuating initial biases, as demonstrated in our study.

4.2.2 Explainable and Responsible AI road-map

4.2.2.1 Proposal of a conceptual framework

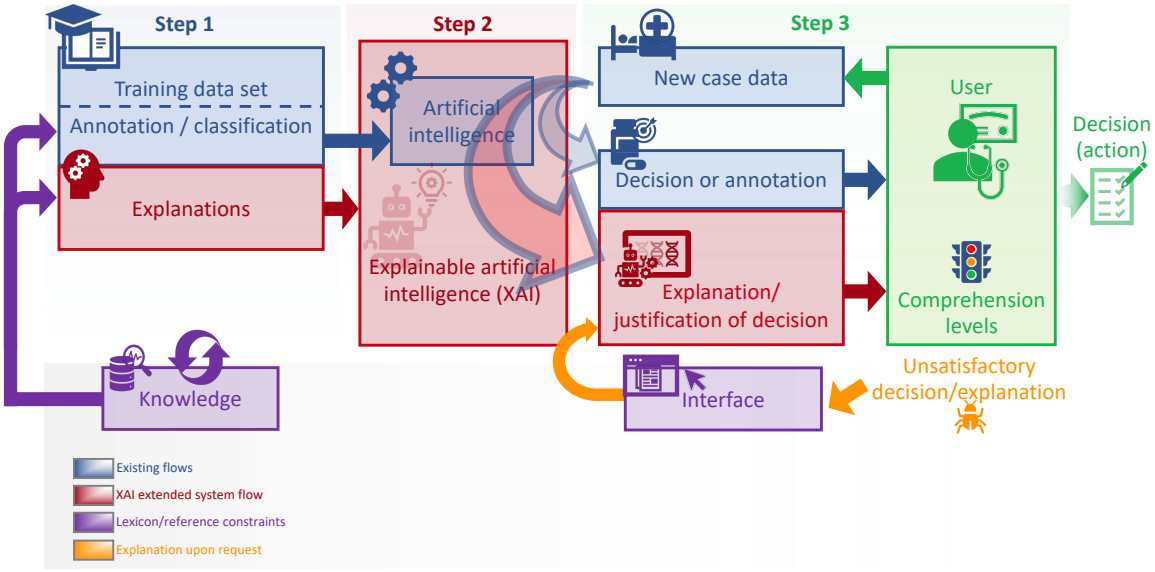


FIGURE 4.2: **Workflow of an AI System Enhanced with XAI Capabilities:** This figure illustrates a three-step process in deploying AI systems with explainable artificial intelligence functionalities. **Step 1** involves the use of a training dataset for model training and explanation generation. **Step 2** shows the application of the trained AI model to new case data, generating decisions or annotations along with explanations or justifications. **Step 3** highlights the role of the user in assessing the AI's output. The user evaluates the decision or annotation based on their comprehension levels, which may lead to accepting the outcome or requesting further explanations if the initial output is deemed unsatisfactory.

In pursuit of a formalism for specifying the explanation functions of intelligent systems, we propose several definitions (Racoceanu, Ounissi, and Kergosien, 2024): we define *explanation* as the message sent by the system that needs to explain itself (the *source*) and received by the requester/user (the *target*), possibly in response to a query (a request for explanation, which may be more or less specified). It is advisable not to restrict the targets to humans; other intelligent systems may also be involved.

We define *understanding* as a particular state change (comprehension levels) in the target, which is considered as a finite automaton. Depending on its effects on the target, an explanation can be more or less satisfactory: a metric to qualify it could rely on the target's

structure and initial state. The effect of an explanation appears to be akin to that of a projection by its property of idempotence: if a complete understanding is achieved after a first explanation, a second explanation no longer alters the state (refer to Figure 4.2).

The search for metrics on the quality of an explanation can draw on interrogations and psychological methods estimating a degree of satisfaction of the target, or quantify other aspects of the explanation: adequacy of the response to the request, the target's level of expertise, the existence of reference vocabularies, completeness of the response (in a sense to be specified).

This approach could be deemed sufficient for the majority of known and documented tasks such as segmentation or classification. Here, "sufficient" implies that the *target* requirements are met, and no further explanation is necessary. However, there are scenarios where an entirely automated response is not permissible by law, especially in sensitive fields like healthcare, or when the system fails to bring the *target* to a satisfactory condition after N attempts. In these cases, explanations alone may not suffice to enhance the *target*'s understanding to an acceptable level.

Under such circumstances, the shortcomings are escalated to *experts*–such as computer scientists, doctors, or lawyers–who are then incorporated into the process. The initial task for these experts is to provide explanations that are in compliance with regulations, such as the "right to explanation." They must also validate responses to ensure they meet the *target*'s satisfaction. This expert involvement not only aids in addressing the immediate issue but also ensures compliance with legal and ethical standards.

Furthermore, this feedback loop is crucial for the system's continuous improvement. It is recorded for traceability and integrated into the system through methods like continual learning and bug fixes. This integration aids in refining the system's ability to handle similar issues in the future, thereby streamlining the experts' roles to primarily validation without the need for direct intervention. This structured approach enhances both the system's efficacy and its reliability in complex decision-making environments (refer to Figure 4.3).

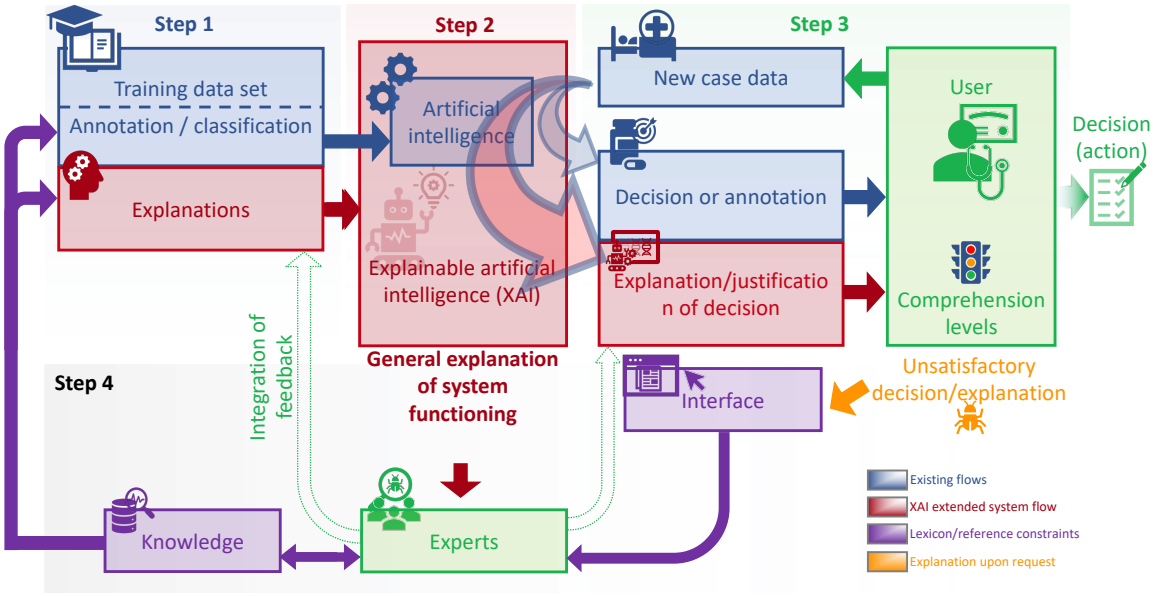


FIGURE 4.3: **Integration of Expert Feedback in an XAI System:** This diagram extends the workflow shown in Figure 4.2 by introducing **Step 4**, which involves expert intervention when AI-generated explanations are deemed unsatisfactory or when handling sensitive domains such as healthcare. Experts provide a higher level of scrutiny and validation, offering a deeper explanation that aligns with regulatory requirements and enhances system trustworthiness. This feedback is integrated back into the system to refine its future responses and ensure compliance with ethical standards.

This framework can be considered as a responsible AI system if it consistently meets all the criteria outlined in Section 1.2.6. It must adapt continually to comply with the dynamic legal and regulatory frameworks that govern privacy, accountability, and fairness. These requirements vary not only between different domains but also within each domain, tailored to specific downstream tasks. Such an approach ensures that the AI system remains effective and ethical across its various applications, thereby upholding the principles of responsible AI throughout its lifecycle.

How the XAI framework can be applied to radiology (use case): Let us analyze what explanation and understanding between humans might entail to then generalize these to other intelligent systems. Consider a radiologist (the source) who must explain their diagnosis of malignant breast tumor to a correspondent (the target) following a mammographic examination. The radiological report might include: "on the frontal view, poorly defined rounded opacity with spiculated margins measuring 15mm in diameter in the upper outer quadrant, without architectural distortions", and later conclude with "probable adenocarcinoma, BIRADS classification 5".

The descriptive part allows an expert to locate the lesion on the images and confirm the diagnosis, possibly to discuss or contest it. A less experienced correspondent might request additional explanations. To the query "where is the lesion located?" the source could respond by annotating the center of a disc approximating the round image. Arrows might indicate the contour, other arrows the spicules. Thus, the abstract concept of "rounded image" is instantiated as a precisely defined disc within the image. The question of the presence of spicules is replaced by the appreciation of the spiculated character of a segment of the contour perfectly identified. We can say that the explanation has completely instantiated

the abstract concepts used in the verbal description, constituting a logical interpretation (which assigns elements of a set called the domain to variable symbols: here, the variable "rounded mass" corresponds to a unique disc in the plane). To the question "why this type of tumor?" the source might respond with bibliographic references ("knowledge") justifying such a deduction from the description elements now understood, which this time, explains a deductive reasoning within a logical formalism.

Bibliography

Abraham, T. M. et al. (2022). "Mode-mapping qOBM microscopy to virtual hematoxylin and eosin (H&E) histology via deep learning". In: *Unconventional Optical Imaging III*. Ed. by M. P. Georges, G. Popescu, and N. Verrier. Vol. 12136. International Society for Optics and Photonics. SPIE, 121360Q. DOI: 10.1117/12.2622160.

Abraham, T. H. (2004). "Nicolas Rashevsky's mathematical biophysics". In: *Journal of the History of Biology* 37.2, pp. 333–385.

Agency, D. A. R. P. (Aug. 2016). "Explainable Artificial Intelligence (XAI)". In.

Allen, G. I. (2020). "Handbook of Graphical Models". In: *Journal of the American Statistical Association* 115.531, pp. 1555–1557. DOI: 10.1080/01621459.2020.1801279.

Anoosheh, A. et al. (2017). "ComboGAN: Unrestrained Scalability for Image Domain Translation". In: *CoRR* abs/1712.06909. arXiv: 1712.06909.

Arai T, e. a. (2006). "TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis". In: *Biochemical and Biophysical Research Communications* 351.3, pp. 602–611. DOI: 10.1016/j.bbrc.2006.10.093.

Arbelle, A., S. Cohen, and T. R. Raviv (2022). "Dual-Task ConvLSTM-UNet for Instance Segmentation of Weakly Annotated Microscopy Videos". In: *IEEE Transactions on Medical Imaging* 41.8, pp. 1948–1960. DOI: 10.1109/TMI.2022.3152927.

Arbelle, A. and T. Riklin Raviv (2019). "Microscopy Cell Segmentation via Convolutional LSTM Networks". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. DOI: 10.1109/ISBI.2019.8759447.

Bai, B. et al. (2023). "Deep learning-enabled virtual histological staining of biological samples". In: *Light: Science & Applications* 12.1, p. 57. DOI: 10.1038/s41377-023-01104-7.

Bai, M. and R. Urtasun (2017). "Deep Watershed Transform for Instance Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. DOI: 10.1109/CVPR.2017.305.

Baldi, P. (2012). "Autoencoders, unsupervised learning, and deep architectures". In: *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings, pp. 37–49.

Bankhead, P., M. Loughrey, and J. e. a. Fernández (2017). "QuPath: Open source software for digital pathology image analysis". In: *Scientific Reports* 7.1. DOI: 10.1038/s41598-017-17204-5.

Barredo Arrieta, A. et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115. DOI: https://doi.org/10.1016/j.inffus.2019.12.012.

Beebee, H., C. Hitchcock, and P. Menzies (Dec. 2009). *The Oxford Handbook of Causation*. English. United Kingdom: Oxford University Press. ISBN: 9780199279739.

Besson, S. et al. (2019). "Bringing Open Data to Whole Slide Imaging". In: *Digital Pathology*. Ed. by C. C. Reyes-Aldasoro et al. Cham: Springer International Publishing, pp. 3–10. ISBN: 978-3-030-23937-4.

Beucher, S. and F. Meyer (Jan. 1993). "Segmentation: The Watershed Transformation. Mathematical Morphology in Image Processing". In: *Optical Engineering* 34, pp. 433–481.

Billot, B. et al. (2023). "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining". In: *Medical Image Analysis* 86, p. 102789. DOI: https://doi.org/10.1016/j.media.2023.102789.

Bontempi, D. et al. (2020). "CEREBRUM: a fast and fully-volumetric Convolutional Encoder-decodeR for weakly-supervised sEgmentation of BRain strUctures from out-of-the-scanner MRI". In: *Medical Image Analysis* 62, p. 101688. DOI: https://doi.org/10.1016/j.media.2020.101688.

Boorboor, S. et al. (2023). "NeuRegenerate: A Framework for Visualizing Neurodegeneration". In: *IEEE Transactions on Visualization and Computer Graphics* 29.3, pp. 1625–1637. DOI: 10.1109/TVCG.2021.3127132.

Borhani, N. et al. (2019). "Digital staining through the application of deep neural networks to multi-modal multi-photon microscopy". In: *Biomed. Opt. Express* 10.3, pp. 1339–1350. DOI: 10.1364/BOE.10.001339.

Borovec, J. et al. (2020). "ANHIR: Automatic Non-Rigid Histological Image Registration Challenge". In: *IEEE Transactions on Medical Imaging* 39.10, pp. 3042–3052. DOI: 10.1109/TMI.2020.2986331.

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152.

Bove, A. et al. (2017). "Local cellular neighborhood controls proliferation in cell competition". In: *Molecular Biology of the Cell* 28. DOI: 10.1091/mbc.E17-06-0368.

Bradski, G. (2000a). "The openCV library." In: *Dr. Dobb's Journal: Software Tools for the Professional Programmer* 25.11, pp. 120–123.

— (2000b). "The opencv library." In: *Dr. Dobb's Journal: Software Tools for the Professional Programmer* 25.11, pp. 120–123.

Breiman, L. (1996). "Bagging predictors". In: *Machine learning* 24, pp. 123–140.

— (2001). "Random forests". In: *Machine learning* 45, pp. 5–32.

Bright F, e. a. (2021). "TDP-43 and Inflammation: Implications for Amyotrophic Lateral Sclerosis and Frontotemporal Dementia". In: *International Journal of Molecular Sciences* 22.15, p. 7781. DOI: 10.3390/ijms22157781.

Bronstein, M. M. et al. (2017). "Geometric deep learning: going beyond euclidean data". In: *IEEE Signal Processing Magazine* 34.4, pp. 18–42.

Buggenthin F. Marr C., S. M. e. a. (2013). "An automatic method for robust and fast cell detection in bright field images from high-throughput microscopy". In: *BMC bioinformatics* 14.1, p. 297. DOI: 10.1186/1471-2105-14-297.

Burlingame, E. A. et al. (2020). "SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning". In: *Scientific Reports* 10.1, p. 17507. DOI: 10.1038/s41598-020-74500-3.

Canny, J. (1986). "A computational approach to edge detection". In: *IEEE Transactions on pattern analysis and machine intelligence* 6, pp. 679–698.

Castellano, G. et al. (2004). "Texture analysis of medical images". In: *Clinical radiology* 59.12, pp. 1061–1069.

Chattopadhyay, A. et al. (2017). "Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks". In: *CoRR* abs/1710.11063. arXiv: 1710.11063.

— (Mar. 2018). "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847. DOI: 10.1109/WACV.2018.00097.

Chen, J., L. Song, M. Wainwright, et al. (2018). "Learning to explain: An information-theoretic perspective on model interpretation". In: *International conference on machine learning*. PMLR, pp. 883–892.

Chen, L.-C. et al. (2018). "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *ECCV*.

Chen, T. et al. (2015). "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4, pp. 1–4.

Choi, Y., M. Choi, et al. (2017). "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation". In: *CoRR* abs/1711.09020. arXiv: 1711.09020.

Choi, Y., Y. Uh, et al. (2019). "StarGAN v2: Diverse Image Synthesis for Multiple Domains". In: *CoRR* abs/1912.01865. arXiv: 1912.01865.

Christoph Sommer, D. W. G. (2013). "Machine learning in cell biology - teaching computers to recognize phenotypes". In: *Journal of Cell Science* 126.24, pp. 5529–5539. DOI: doi.org/10.1242/jcs.123604.

Ciompi, F. et al. (2017). "Artificial intelligence for pathology: challenges and opportunities". In: *Journal of Clinical Pathology*.

Ciresan, D. C., U. Meier, and J. Schmidhuber (2012). "Multi-column Deep Neural Networks for Image Classification". In: *CoRR* abs/1202.2745. arXiv: 1202.2745.

Courtney, P. et al. (2018). "Fully convolutional networks for multiclass segmentation of histopathology handbag imagery". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*.

Daly, L. and G. J. Bourke (2008). *Interpretation and uses of medical statistics*. John Wiley & Sons.

Dawid, A. P. (2010). "Beware of the DAG!" In: *NIPS Causality: Objectives and Assessment*.

Dienes, Z. (2014). "Using Bayes to get the most out of non-significant results". In: *Frontiers in psychology* 5, p. 781.

Directorate-General for Communications Networks, C. and E. C. Technology (European Commission) (Feb. 2020). "WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust". In.

Doshi-Velez, F. and B. Kim (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. DOI: 10.48550/ARXIV.1702.08608.

Echle, A. et al. (2021). "Deep learning in cancer pathology: a new generation of clinical biomarkers". In: *British Journal of Cancer* 124.4, pp. 686–696. DOI: 10.1038/s41416-020-01122-x.

Ershov, D. et al. (2022). "TrackMate 7: integrating state-of-the-art segmentation algorithms into tracking pipelines". In: *Nature Methods* 19.7, pp. 829–832. DOI: 10.1038/s41592-022-01507-1.

Ethics of Scientific Knowledge, W. C. on the and Technology (Feb. 2019). "Preliminary study on the Ethics of Artificial Intelligence". In.

Evangelidis, G. D. and E. Z. Psarakis (2008). "Parametric Image Alignment Using Enhanced Correlation Coefficient Maximization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.10, pp. 1858–1865. DOI: 10.1109/TPAMI.2008.113.

Fischl, B. (2012). "FreeSurfer". In: *Neuroimage* 62.2, pp. 774–781.

Friedman, H. (2007). "Interpretations according to Tarski". In: *This is one of the 2007 Tarski Lectures at Berkeley. The lecture is available at http://www. math. osu. edu/˜ friedman* 8.

Fu, R. et al. (Aug. 2020). "Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs". In: *arXiv:2008.02312 [cs, eess]*.

Gatenbee, C. D. et al. (July 26, 2023). "Virtual alignment of pathology image series for multi-gigapixel whole slide images". In: *Nature Communications* 14.1, p. 4502. DOI: 10.1038/s41467-023-40218-9.

Gentleman, S. (2013). "Review: microglia in protein aggregation disorders: friend or foe?" In: *Neuropathology and Applied Neurobiology* 39.1, pp. 45–50. DOI: 10.1111/nan.12017.

Ghahramani, Z. (2015). "Probabilistic machine learning and artificial intelligence". In: *Nature* 521.7553, pp. 452–459.

Gibney, E. (2022). "Is AI fuelling a reproducibility crisis in science". In: *Nature* 608.7922, pp. 250–1.

Goodfellow, I., J. Pouget-Abadie, et al. (2014). "Generative adversarial nets". In: *Advances in neural information processing systems*.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT press.

Guerrero Peña, F. A. et al. (2020). "J Regularization Improves Imbalanced Multiclass Segmentation". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. DOI: 10.1109/ISBI45749.2020.9098550.

Gundersen, O. E. and S. Kjensmo (2018). "State of the art: Reproducibility in artificial intelligence". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1.

Gunning, D. and D. W. Aha (2019). "DARPA's Explainable Artificial Intelligence (XAI) Program". In: *AI Magazine*.

Haan, K. de et al. (2021). "Deep learning-based transformation of H&E stained tissues into special stains". In: *Nature Communications* 12.1, p. 4884. DOI: 10.1038/s41467-021-25221-2.

Hamming, R. W. (1998). *Digital Filters*. Dover Publications.

Harris, C. R. et al. (Sept. 2020). "Array programming with NumPy". In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

Haukedal H, F. K. (2019). "Implications of Microglia in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia". In: *Journal of Molecular Biology* 431.9, pp. 1818–1829. DOI: 10.1016/j.jmb.2019.02.004.

He, K., G. Gkioxari, et al. (2017). "Mask R-CNN". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.

He, K., X. Zhang, et al. (2015). "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385. arXiv: 1512.03385.

Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology press.

Henschel, L. et al. (2020). "FastSurfer - A fast and accurate deep learning based neuroimaging pipeline". In: *NeuroImage* 219, p. 117012. DOI: https://doi.org/10.1016/j.neuroimage.2020.117012.

Hong, Y. et al. (2021). "Deep learning-based virtual cytokeratin staining of gastric carcinomas to measure tumor–stroma ratio". In: *Scientific Reports* 11.1, p. 19255. DOI: 10.1038/s41598-021-98857-1.

Hosmer Jr, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied logistic regression*. John Wiley & Sons.

How, J. P. (2018). "Ethically Aligned Design [From the Editor]". In: *IEEE Control Systems Magazine* 38.3, pp. 3–4. DOI: 10.1109/MCS.2018.2810458.

Howson, C. and P. Urbach (2006). *Scientific reasoning: the Bayesian approach*. Open Court Publishing.

Huang, X. et al. (2018). "Multimodal Unsupervised Image-to-Image Translation". In: *CoRR* abs/1804.04732. arXiv: 1804.04732.

Huff, D. T., A. J. Weisman, and R. Jeraj (2021). "Interpretation and visualization techniques for deep learning models in medical imaging". In: *Physics in Medicine & Biology* 66.4, 04TR01.

Isensee, F. et al. (Feb. 1, 2021). "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2, pp. 203–211. DOI: 10.1038/s41592-020-01008-z.

Isola, P. et al. (2016). "Image-to-Image Translation with Conditional Adversarial Networks". In: *CoRR* abs/1611.07004. arXiv: 1611.07004.

J., C., S. Cooper, and F. e. a. Heigwer (2017). "Data-analysis strategies for image-based cell profiling". In: *Nature Methods* 14.9, pp. 849–863. DOI: doi.org/10.1038/nmeth.4397.

Jamie Sherman, P. W. (2023). "aicspylibczi v3.1.0: Python module to expose libCZI functionality. github.com/AllenCellModeling/aicspylibczi". In: *GitHub*.

Janda, E., L. Boi, and A. Carta (2018). "Microglial Phagocytosis and Its Regulation: A Therapeutic Target in Parkinson's Disease?" In: *Frontiers in Molecular Neuroscience* 11, p. 144. DOI: 10.3389/fnmol.2018.00144.

Jobin, A., M. Ienca, and E. Vayena (2019). "The global landscape of AI ethics guidelines". In: *Nature machine intelligence* 1.9, pp. 389–399.

Kim, J. et al. (2019). "U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation". In: *CoRR* abs/1907.10830. arXiv: 1907.10830.

Kingma, D. P. and J. Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kleinbaum, D. G. et al. (2002). *Logistic regression*. Springer.

Kosaraju, S. C. et al. (2020). "Deep-Hipo: Multi-scale receptive field deep learning for histopathological image analysis". In: *Methods* 179, pp. 3–13. DOI: https://doi.org/10.1016/j.ymeth.2020.05.012.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc.

Kruschke, J. K. and T. M. Liddell (2018). "The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective". In: *Psychonomic bulletin & review* 25, pp. 178–206.

Kumar, V. et al. (2012). "Radiomics: the process and the challenges". In: *Magnetic resonance imaging* 30.9, pp. 1234–1248.

Lahiani, A. et al. (2021). "Seamless Virtual Whole Slide Image Synthesis and Validation Using Perceptual Embedding Consistency". In: *IEEE Journal of Biomedical and Health Informatics* 25.2, pp. 403–411. DOI: 10.1109/JBHI.2020.2975151.

Lall D, e. a. (2021). "C9orf72 deficiency promotes microglial-mediated synaptic loss in aging and amyloid accumulation". In: *Neuron* 109.14, 2275–2291.e8. DOI: 10.1016/j.neuron.2021.05.020.

Lecun, Y. et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791.

Levy, J. and et al. (2020). "Preliminary Evaluation of the Utility of Deep Generative Histopathology Image Translation at a Mid-sized NCI Cancer Center". In: *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies*. Valletta, Malta: SCITEPRESS, pp. 302–311.

Li, C. and M. Wand (2016). "Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks". In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe et al. Cham: Springer International Publishing, pp. 702–716. ISBN: 978-3-319-46487-9.

Li, Q. and B. Barres (2018). "Microglia and macrophages in brain homeostasis and disease". In: *Nature Reviews Immunology* 18.4, pp. 225–242. DOI: 10.1038/nri.2017.125.

Li, Q. and M. Haney (2020). "The role of glia in protein aggregation". In: *Neurobiology of Disease* 143, p. 105015. DOI: 10.1016/j.nbd.2020.105015.

Li, W., W. Huang, and Y. Zheng (2024). "CorrDiff: Corrective Diffusion Model for Accurate MRI Brain Tumor Segmentation". In: *IEEE Journal of Biomedical and Health Informatics* 28.3, pp. 1587–1598. DOI: 10.1109/JBHI.2024.3353272.

Li, Z. et al. (2021). "A survey of convolutional neural networks: analysis, applications, and prospects". In: *IEEE transactions on neural networks and learning systems* 33.12, pp. 6999–7019.

Liang, P. et al. (2022a). "H-EMD: A Hierarchical Earth Mover's Distance Method for Instance Segmentation". In: *IEEE Transactions on Medical Imaging* 41.10, pp. 2582–2597. DOI: 10.1109/TMI.2022.3169449.

— (2022b). "H-EMD: A Hierarchical Earth Mover's Distance Method for Instance Segmentation". In: *IEEE Transactions on Medical Imaging* 41.10, pp. 2582–2597. DOI: 10.1109/TMI.2022.3169449.

Lin, Y. et al. (2022). "Unpaired Multi-Domain Stain Transfer for Kidney Histopathological Images". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2, pp. 1630–1637. DOI: 10.1609/aaai.v36i2.20054.

Lindeberg, T. (1990). "Scale-space for discrete signals". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.3, pp. 234–254. DOI: 10.1109/34.49051.

— (2012). "Scale Invariant Feature Transform". In: *Scholarpedia* 7, p. 10491. DOI: 10.4249/scholarpedia.10491.

Lindemann, B. et al. (2021). "A survey on long short-term memory networks for time series prediction". In: *Procedia CIRP* 99, pp. 650–655. DOI: 10.1016/j.procir.2021.03.088.

Linkert, M. et al. (May 2010). "Metadata matters: access to image data in the real world". In: *Journal of Cell Biology* 189.5, pp. 777–782. DOI: 10.1083/jcb.201004104. eprint: https://rupress.org/jcb/article-pdf/189/5/777/1477121/jcb_201004104.pdf.

Liu, M. et al. (2019). "Few-Shot Unsupervised Image-to-Image Translation". In: *CoRR* abs/1905.01723. arXiv: 1905.01723.

Liu, Q. et al. (2018). "Regularization techniques for fine-tuning in neural networks". In: *arXiv preprint arXiv:1810.00553*.

Liu, S., C. Zhu, et al. (2022). *BCI: Breast Cancer Immunohistochemical Image Generation through Pyramid Pix2pix*. arXiv: 2204.11425 [eess.IV].

Liu, S., B. Zhang, et al. (2021). "Unpaired Stain Transfer Using Pathology-Consistent Constrained Generative Adversarial Networks". In: *IEEE Transactions on Medical Imaging* 40.8, pp. 1977–1989. DOI: 10.1109/TMI.2021.3069874.

Liu, Z., L. Jin, and J. C. et al (2021). "A survey on applications of deep learning in microscopy image analysis". In: *Computers in Biology and Medicine* 134, p. 104523. DOI: https://doi.org/10.1016/j.compbiomed.2021.104523.

Lui H, e. a. (2016). "Progranulin Deficiency Promotes Circuit-Specific Synaptic Pruning by Microglia via Complement Activation". In: *Cell* 165.4, pp. 921–935. DOI: 10.1016/j.cell.2016.04.001.

Lundberg, S. M. and S.-I. Lee (2017a). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 4765–4774.

— (2017b). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 4765–4774.

Lundberg, S. M., B. Nair, et al. (2018). "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery". In: *Nature Biomedical Engineering* 2.10, p. 749.

Lundberg, S. M., G. Erion, et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1, pp. 2522–5839.

Maaten, L. van der and G. Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605.

Magaki, S. et al. (2019). "An Introduction to the Performance of Immunohistochemistry". In: *Biobanking: Methods and Protocols*. New York, NY: Springer New York, pp. 289–298. ISBN: 978-1-4939-8935-5. DOI: 10.1007/978-1-4939-8935-5_25.

Marée, R. et al. (Jan. 2016). "Collaborative analysis of multi-gigapixel imaging data using Cytomine". In: *Bioinformatics* 32.9, pp. 1395–1401. DOI: 10.1093/bioinformatics/btw013. eprint: https://academic.oup.com/bioinformatics/article-pdf/32/9/1395/49019373/bioinformatics_32_9_1395.pdf.

Maška, M. et al. (July 1, 2023). "The Cell Tracking Challenge: 10 years of objective benchmarking". In: *Nature Methods* 20.7, pp. 1010–1020. DOI: 10.1038/s41592-023-01879-y.

Matula, P. et al. (2015). "Cell Tracking Accuracy Measurement Based on Comparison of Acyclic Oriented Graphs". In: *PLoS ONE* 10.12, e0144959. DOI: 10.1371/journal.pone.0144959.

Mayerhoefer, M. E. et al. (2020). "Introduction to radiomics". In: *Journal of Nuclear Medicine* 61.4, pp. 488–495.

McCulloch, W. S. and W. Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5, pp. 115–133.

Meijering, E., O. Dzyubachyk, and I. Smal (2012). "Chapter nine - Methods for Cell and Particle Tracking". In: *Imaging and Spectroscopic Analysis of Living Cells*. Ed. by P. M. conn. Vol. 504. Methods in Enzymology. Academic Press, pp. 183–200. DOI: https://doi.org/10.1016/B978-0-12-391857-4.00009-4.

Mercan, C. et al. (2020). *Virtual staining for mitosis detection in Breast Histopathology*. arXiv: 2003.07801 [eess.IV].

Metropolis, N. and S. Ulam (1949). "The monte carlo method". In: *Journal of the American statistical association* 44.247, pp. 335–341.

Mhaskar, H., Q. Liao, and T. Poggio (2016). "Learning functions: when is deep better than shallow". In: *arXiv preprint arXiv:1603.00988*.

Minsky, M. and S. A. Papert (2017). *Perceptrons, reissue of the 1988 expanded edition with a new foreword by Léon Bottou: an introduction to computational geometry*. MIT press.

Mitchell, R., E. Frank, and G. Holmes (2020). "GPUTreeShap: Fast Parallel Tree Interpretability". In: *CoRR* abs/2010.13972. arXiv: 2010.13972.

Moen, E. et al. (2019). "Deep learning for cellular image analysis". In: *Nature Methods* 16.12, pp. 1233–1246. DOI: 10.1038/s41592-019-0403-1.

Montgomery, D. C., E. A. Peck, and G. G. Vining (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Neubert, P. and P. Protzel (2014). "Compact Watershed and Preemptive SLIC: On Improving Trade-offs of Superpixel Segmentation Algorithms". In: *ICPR*, pp. 996–1001. DOI: 10.1109/ICPR.2014.181.

Neumann M, e. a. (2006). "Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis". In: *Science* 314.5796, pp. 130–133. DOI: 10.1126/science.1134108.

Noble, W. S. (2006). "What is a support vector machine?" In: *Nature biotechnology* 24.12, pp. 1565–1567.

OECD, L. (May 2019). "Recommendation on the Ethics of Artificial Intelligence". In.

Oktay, O. et al. (2018). "Attention U-Net: Learning Where to Look for the Pancreas". In: *CoRR* abs/1804.03999. arXiv: 1804.03999.

Olazaran, M. (1996). "A Sociological Study of the Official History of the Perceptrons Controversy". In: *Social Studies of Science* 26, pp. 611–659.

Oppenheim, A. V. and R. W. Schafer (1999). *Discrete-Time Signal Processing*. 2nd ed. Prentice Hall.

Oumarou Hama, H. et al. (2022). "Immunohistochemical diagnosis of human infectious diseases: a review". In: *Diagnostic Pathology* 17.1, p. 17. DOI: 10.1186/s13000-022-01197-5.

Paszke, A. et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pp. 8024–8035.

Pearl, J. and D. Mackenzie (2018). *The book of why: the new science of cause and effect*. Basic books.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12, pp. 2825–2830.

Philippe Bracke Anupam Datta, C. J. and S. Sen (Aug. 2019). "Machine learning explainability in finance: an application to default risk analysis". In.

President, E. O. of the, J. M. Holden, and Smith (2016). "Preparing for the future of artificial intelligence". In.

President, U. S. (E. O. of the, J. Holdren, and M. Smith (Oct. 2016). "Preparing for the future of artificial intelligence". In.

PyVips Library (2024). https://www.libvips.org/.

Racoceanu, D., M. Ounissi, and Y. L. Kergosien (2024). *Explicabilité en Intelligence Artificielle ; vers une IA Responsable - Instanciation dans le domaine de la santé*. DOI: 10.51257/a-v1-h5030.

Rashevsky, N. (1948). *Mathematical biophysics*. The University of Chicago Press.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). """ Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Rivenson, Y. et al. (2019). "PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning". In: *Light: Science & Applications* 8.1, p. 23. DOI: 10.1038/s41377-019-0129-y.

Ronneberger, O., P. Fischer, and T. Brox (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Vol. 9351. Springer. DOI: 10.1007/978-3-319-24574-4_28.

Rosenblatt, F. (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*. Tech. rep. Cornell Aeronautical Lab Inc Buffalo NY.

Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5, pp. 206–215.

Saha, M., C. Chakraborty, and D. Racoceanu (2018). "Efficient deep learning model for mitosis detection using breast histopathology images". In: *Computerized Medical Imaging and Graphics* 64, pp. 29–40. DOI: https://doi.org/10.1016/j.compmedimag.2017.12.001.

Scheiblich, H. et al. (2021). "Microglia jointly degrade fibrillar alpha-synuclein cargo by distribution through tunneling nanotubes". In: *Cell* 184.20, 5089–5106.e21. DOI: https://doi.org/10.1016/j.cell.2021.09.007.

Schmidt, U. et al. (2018). "Cell Detection with Star-Convex Polygons". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Vol. 11071. Springer. DOI: 10.1007/978-3-030-00934-2_30.

Schultheiß, S. and D. Lewandowski (2023). "Misplaced trust? The relationship between trust, ability to identify commercially influenced results and search engine preference". In: *Journal of Information Science* 49.3, pp. 609–623.

Selbst, A. and J. Powles (2018). ""Meaningful information" and the right to explanation". In: *conference on fairness, accountability and transparency*. PMLR, pp. 48–48.

Selvaraju, R. R. et al. (2016). "Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization". In: *CoRR* abs/1610.02391. arXiv: 1610.02391.

— (Jan. 2017). "Grad-CAM: Why did you say that?" In: *arXiv:1611.07450 [cs, stat]*.

Sengupta, D., P. Gupta, and A. Biswas (2022). "A survey on mutual information based medical image registration algorithms". In: *Neurocomputing* 486, pp. 174–188. DOI: 10.1016/j.neucom.2021.11.023.

Shahriari, K. and M. Shahriari (2017). "IEEE standard review Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems". In: *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, pp. 197–201. DOI: 10.1109/IHTC.2017.8058187.

Sheller, M. J. et al. (July 2020). "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data". In: *Scientific reports* 10.1, p. 12598. DOI: 10.1038/s41598-020-69250-1.

Sirinukunwattana, K. et al. (2016). "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images". In: *IEEE transactions on medical imaging*.

Spontón, H. and J. Cardelino (2015). "A review of classic edge detectors". In: *Image Processing On Line* 5, pp. 90–123.

Sprott, J. C. (2003). *Chaos and time-series analysis*. Oxford university press.

Srinivasan, G. and G. Shobha (2008). "Statistical texture analysis". In: *Proceedings of world academy of science, engineering and technology*. Vol. 36. December, pp. 1264–1269.

Stringer, C. et al. (2021). "Cellpose: a generalist algorithm for cellular segmentation". In: *Nature Methods*. DOI: 10.1038/s41592-020-01018-x.

Stritt, M., A. K. Stalder, and E. Vezzali (Feb. 2020). "Orbit Image Analysis: An open-source whole slide image analysis tool". In: *PLOS Computational Biology* 16.2, pp. 1–19. DOI: 10.1371/journal.pcbi.1007313.

Strobl, C., J. Malley, and G. Tutz (2009). "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests." In: *Psychological methods* 14.4, p. 323.

Szczerba, L. W. (1977). "Interpretability of elementary theories". In: *Logic, Foundations of Mathematics, and Computability Theory: Part One of the Proceedings of the Fifth International Congress of Logic, Methodology and Philosophy of Science, London, Ontario, Canada-1975*. Springer, pp. 129–145.

Tellez, D. et al. (2018). "Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks". In: *IEEE transactions on medical imaging*.

Trocin, C. et al. (2023). "Responsible AI for digital health: a synthesis and a research agenda". In: *Information Systems Frontiers* 25.6, pp. 2139–2157.

Tsima, K. (2023). "The reproducibility issues that haunt health-care AI". In: *Nature* 613.

Turaga, S. C. et al. (Feb. 2010). "Convolutional networks can learn to generate affinity graphs for image segmentation". English. In: *Neural Computation* 22.2, pp. 511–38. DOI: 10.1162/neco.2009.10-08-881.

Ulicna, K. et al. (2021). "Automated Deep Lineage Tree Analysis Using a Bayesian Single Cell Tracking Approach". In: *Frontiers in Computer Science* 3. DOI: 10.3389/fcomp.2021.734559.

UNESCO (2021). "Recommendation on the Ethics of Artificial Intelligence". In.

Valabregue, R., F. Girka, et al. (2023). *Comprehensive analysis of synthetic learning applied to neonatal brain MRI segmentation*. arXiv: 2309.05306 [stat.ML].

Valabregue, R., I. Khemir, et al. (2024). "Unraveling Systematic Biases in Brain Segmentation: Insights from Synthetic Training". In: *Medical Imaging with Deep Learning*.

van der Velden, B. H. et al. (2022). "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *Medical Image Analysis* 79, p. 102470. DOI: https://doi.org/10.1016/j.media.2022.102470.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Villani, C. et al. (2018). "Donner un sens à l'intelligence artificielle: pour une stratégie nationale et européenne". In.

Wachter, S., B. Mittelstadt, and L. Floridi (2017). "Transparent, explainable, and accountable AI for robotics". In: *Science robotics* 2.6, eaan6080.

Wachter, S., B. Mittelstadt, and C. Russell (2017). "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31, p. 841.

Wachter, S., B. D. Mittelstadt, and C. Russell (2017). "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR". In: *Cybersecurity*.

Walt, S. van der et al. (2014). "scikit-image: Image processing in Python". In: *PeerJ* 2, e453. DOI: 10.7717/peerj.453.

Xiang, T. et al. (2020). "BiO-Net: Learning Recurrent Bi-directional Connections for Encoder-Decoder Architecture". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Vol. 12261. Springer. DOI: 10.1007/978-3-030-59710-8_8.

Xie, W. et al. (2022). "Prostate Cancer Risk Stratification via Nondestructive 3D Pathology with Deep Learning-Assisted Gland Analysis". English. In: *Cancer Research* 82.2, pp. 334–345. DOI: 10.1158/0008-5472.CAN-21-2843.

Xing, F. et al. (2018). "Deep Learning in Microscopy Image Analysis: A Survey". In: *IEEE Transactions on Neural Networks and Learning Systems* 29.10, pp. 4550–4568. DOI: 10.1109/TNNLS.2017.2766168.

Xing, L., L. Cai, et al. (2018). "A multi-scale contrast-based image quality assessment model for multi-exposure image fusion". In: *Signal Processing* 145, pp. 233–240. DOI: https://doi.org/10.1016/j.sigpro.2017.12.013.

Xing, L., H. Zeng, et al. (2017). "Multi-exposure image fusion quality assessment using contrast information". In: *2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 34–38. DOI: 10.1109/ISPACS.2017.8265641.

Xu, N. et al. (2016). "Deep Interactive Object Selection". In: *CoRR* abs/1603.04042. arXiv: 1603.04042.

Yang, X. et al. (Sept. 2022). "Virtual Stain Transfer in Histology via Cascaded Deep Neural Networks". In: *ACS Photonics* 9.9, pp. 3134–3143. DOI: 10.1021/acsphotonics.2c00932.

Yip, S. S. and H. J. Aerts (2016). "Applications and limitations of radiomics". In: *Physics in Medicine & Biology* 61.13, R150.

Zhang, R. et al. (2022). "MVFStain: Multiple virtual functional stain histopathology images generation based on specific domain mapping". In: *Medical Image Analysis* 80, p. 102520. DOI: https://doi.org/10.1016/j.media.2022.102520.

Zhou, B. et al. (2016). "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.

Zhu, J. et al. (2017). "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *CoRR* abs/1703.10593. arXiv: 1703.10593.

Zhu, Y. and E. Meijering (July 2021). "Automatic improvement of deep learning-based cell segmentation in time-lapse microscopy by neural architecture search". In: *Bioinformatics* 37.24, pp. 4844–4850. DOI: 10.1093/bioinformatics/btab556. eprint: https://academic.oup.com/bioinformatics/article-pdf/37/24/4844/50334743/btab556.pdf.

Appendix A

(Appendix A) PhagoStat: Reproduction Details

Scientific publication

Reproduction details also available in the paper:

Ounissi, M., Latouche, M. and Racoceanu, D. PhagoStat a scalable and interpretable end to end framework for efficient quantification of cell phagocytosis in neurodegenerative disease studies. Sci Rep 14, 6482 (2024). https://doi.org/10.1038/s41598-024-56081-7
Github: https://github.com/ounissimehdi/PhagoStat
Dataset: https://zenodo.org/records/10803492

A.1 Data efficient loading and normalization

. Most of the proprietary microscope software (biologists friendly) are working exclusively on Windows. Therefore, data preparation required specific steps: (i) raw data was first transferred from the microscope machine to a Windows machine; (ii) the acquisition was converted into tagged image file format (TIFF) frames; (iii) the frames were arranged to be compatible with the computational pipeline and iv) the resulting frames were transferred to a high performance computing (HPC) cluster for processing. A notable challenge in this process was the lack of transparency we found when running the software's preprocessing (black-box) steps (i.e., normalization, the conversion of the raw data 16bit to TIFF 8bit frames). Besides, the rich and user-friendly visualization interface, such packages turned out to be overly opaque, inefficient, resource-intensive and time-consuming for analyzing big data. To overcome these drawbacks, we have adapted the 'aicspylibczi'Jamie Sherman, 2023 to develop a flexible, robust, and open-source module capable of reading and converting proprietary raw data formats into universal image formats. Our module was tested on converting Carl Zeiss Image (CZI) files to TIFF format, and it is compatible with Windows, macOS, and Unix. Additionally, the module can be easily adapted to take advantage of HPC clusters and parallelization schemes (refer to Fig.2.11.a and Fig.2.11.b).

The 'aicspylibczi'Jamie Sherman, 2023 Python package was used and extended to read the 'CZI' raw data file using delayed reading. This approach allowed us to read cell and

aggregate channels image-wise without loading all the sequences to the RAM. Images were in 16bit representation. Frame reading time can be accelerated by coupling the Python package with multiple CPUs for parallel processing. This involves several scenes being simultaneously read, with each scene being allocated to its own CPU. This deviates from sequential processing, in which scenes are placed in a queue and read sequentially; (i) in the local machine, our package can use multi-CPUs for parallelism, or (ii) in a HPC clusters that use simple Linux utility for resource Management (SLURM), where our package launches an array of jobs (i.e., attributing to each scene a job ID) on the same node or different nodes.

While using the same parallelism scheme, global percentile normalization is used to re-scale the image pixels' intensities of the whole sequence; 0.5% and 99.5% percentiles for aggregates; 0%, 100% for cells. Aggregate and cell images were re-scaled from 16bit to 8bit using 'img_as_ubyte' function from the 'scikit-image'Walt et al., 2014 Python package. Histogram matching with a normal pixel distribution as reference is used on all cell images, and we apply it using the 'match_histograms' function from the 'scikit-image' Python package. Finally, if needed, images were resized from 2048×2048 to 1024×1024, then saved in 'TIF' format using the 'PIL' Python package.

A.1.1 Isolating aggregate and cell signals for precise quantification and segmentation

In typical microscopy practices, biologists rely on default software provided with the microscope, which often merges the signals from cell instances and aggregates. This fusion allows for visual interpretation, although it may not always suit quantitative analyses or automated processing.

Our approach diverges by treating these channels distinctly to facilitate precise quantification. Specifically, the aggregate channel, which we use to identify clusters of particles, is marked by a unique fluorescent tag (for instance, a red chromatin signature). This enables us to isolate these aggregates onto a separate grayscale image layer using fluorescence microscopy.

Simultaneously, the non-fluorescent signal, corresponding to individual cell instances, is captured on a different grayscale layer. This separation is crucial because it allows us to apply specialized image processing techniques to each channel independently, enhancing the accuracy of cell instance segmentation and aggregate quantification.

During data loading in our system, we maintain this separation. Each channel is loaded individually, thereby preserving the integrity of the information they contain. Consequently, this segregation of data not only simplifies the subsequent image analysis but also ensures that any computational models or algorithms applied later can be fine-tuned to the characteristics of each channel without cross-contamination of signals.

A.1.2 8-bit conversion for performance, transparency, and storage

We evaluated the impact of image bit-depth on the performance of convolutional neural networks in segmentation tasks. The decision to employ an 8-bit conversion of images was informed by a comprehensive ablation study. This study involved training two identical

Unet(XAI) networks: one with 16-bit image inputs and the other with 8-bit image inputs. The findings indicated a marginal enhancement in segmentation performance for the 8-bit model, with mean test losses registering at 0.0802 compared to 0.0827 for the 16-bit counterpart. These results suggested that the lower bit-depth conversion did not hinder, and may in fact have slightly improved, model efficiency for our microglial-dataset.

We used the Fluo-N2DL-HeLa dataset, publicly available from the CTC. Notably, the images in this dataset are inherently 16-bit, and correspondingly, the annotations were also done in 16-bit. Our approach involved training two UNet(XAI) models, one with 16-bit images and the other with 8-bit images. Both models were trained using identical random seeds. We adopted a specific training methodology: we utilized the first sequence of the dataset for training and validation, while the second sequence was reserved for testing. The performance of each model was quantitatively assessed. For the 16 bit model: test loss = 0.1464 and Dice = 0.9642, in contrast with the 8 bit model: test loss = 0.1514 and Dice = 0.9621

We analyzed the loss of information when images are converted from 16-bit to 8-bit format. The process began with re-scaling the intensity of 16-bit images to utilize the full 16-bit range. Following this, we converted these images into 8-bit format. Besides, we divide 16-bit images by 2^{16} and 8-bit images by 2^8 for a direct comparison (images between 0 and 1). Then, we measured the Mean Squared Error (MSE), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) between the re-scaled 16-bit and the converted 8-bit images. This methodology provided a thorough assessment of how the image quality and information fidelity are impacted by the conversion process. The analysis yielded the following results: $MSE = 1.3698e-06 \pm 5.5903e-08$; $PSNR = 58.6368 \pm 0.1747$; $SSIM 0.9996 \pm 2.2954e-05$.

The results across the test loss and Dice show negligible loss (0.005 difference in the test loss and 0,0021 difference in the Dice score), also, across MSE, PSNR and SSIM metrics show negligible loss, thereby providing evidence that the conversion between 16-bit and 8-bit imaging does not significantly impact the model training (at least on the microglial-dataset and the Fluo-N2DL-HeLa dataset).

Furthermore, in collaboration with domain experts in biology and neuroscience, we recognized the necessity for a transparent analytical pipeline. Our collaborators expressed the need for visibility into the intermediate processing stages of CZI file handling to establish trust and ensure traceability of results. To address this request, we incorporated intermediate outputs in universally accessible formats such as GIF, TIFF, and PNG, thus enhancing the interpretability of our pipeline.

Storage optimization also played a pivotal role in our methodology. The original microscope images were captured at a resolution of 2048x2048 pixels in 16-bit format. Given the substantial data storage requirements, particularly when retaining all intermediate outputs, we implemented a strategy to resize images to 1024x1024 pixels in 8-bit format. This approach reduced the storage per frame from approximately 4.2MB to 0.6MB, achieving a nearly seven-fold decrease in data size. This significant reduction facilitated easier data sharing and handling within the research community, enabling peers to download and utilize our dataset more efficiently.

In summary, our methodological adaptations, particularly the conversion to 8-bit image processing, have resulted in an efficient, transparent, and storage-optimized pipeline without sacrificing the integrity and performance of our deep learning models.

A.1.3 Quantitative performance evaluation of the readout module

The proprietary Carl Zeiss Microscopy ZEN light v3.3.89.00 software was used as a reference for data loading and saving. This proprietary software was evaluated using a Windows 11 machine with 8 cores i7 9700K CPU, 16GB RAM, Nvidia RTX2080 GPU and Samsung 970 PRO SSD; all drivers were up-to-date. CUDA acceleration was enabled from the software configuration panel; all parameters were left at their default values, and no tasks ran in the background before/during the benchmark.

Our approach uses only open-source Python packages, as described and cited before. Ubuntu 20.04LTS was used with: 8 cores, i7 9700K CPU, 16GB RAM, HDD or SSD. For single-CPU tests, the hardware used was limited to 1 CPU using 'taskset'. The test was monitored, and during the test, RAM usage did not exceed 1GB (while using an HDD or SSD). For the multi-CPU test, SLURM was used to process 20 job arrays. Each one uses 1 core Xeon Gold 6126 CPU and 1GB RAM (while using HDD or SSD storage node).

All data transfers (single raw data file or frames) were performed using a 1GB/s Ethernet port with 'FileZilla' v3.46.3. SFTP transfer protocol was used while directly connected to the internal institute network (no VPN used); maximum simultaneous transfers were set to 10 files (FileZilla's upper bond).

A.1.4 Technical aspects of our data normalization strategy for large-scale video-microscopy datasets

We initially considered the straightforward approach of loading complete CZI files into a numpy array for processing. This method is feasible for sequences with limited data or shorter duration, given a conventional computational setup with 16/32GB of RAM. However, this became impractical with our datasets, where individual CZI files were approximately 76GB each, comprising 20 sequences captured over 7 hours with a two-minute frame interval across two imaging channels. Standard computing resources were quickly overwhelmed by these files, as evidenced by system performance when RAM capacity was exceeded (refer to Fig. 2.11.c), leading to the use of SWAP space which is much slower than RAM.

In the realm of HPC environments, even though they offer a more robust infrastructure, the extensive size of our datasets still resulted in considerable processing delays and resource contention. This was due to the immense memory requirements (80GB of RAM per CZI file) which induced periods of computational idleness within the HPC's dynamic job scheduling system.

To circumvent these bottlenecks, we adopted a delayed data reading methodology. This technique does not enhance the speed of normalization computation directly but rather optimizes memory consumption. By strategically fetching only the necessary data segment for processing—such as a single frame from a sequence—we were able to initiate 20 Slurm

jobs (one job per sequence) in parallel, each consuming only 1GB of RAM as opposed to the full 80GB that the entire file would require. This also enabled selective access to data, permitting us to isolate specific frames from the imaging channel for detailed morphological analysis.

Technical differentiation between "global" and "local" normalization is crucial in our work. "Global" normalization refers to the standardization of intensities across the entirety of each sequence, whereas "local" normalization pertains to adjustments made on a per-frame basis within a sequence. For global normalization of the aggregate channel, we utilize the first frame's 0.5% and 99.5% intensity percentiles as a baseline (reducing outlier intensities and improving contrast), since all aggregates are present at the beginning of the experiments (non phagocyted yet). Subsequent frames are normalized against this reference, ensuring consistent visualization of morphological features. For the cell channel we use intensities (0% to 100% percentiles) from the first frame to adjust each frame thereafter. Following background noise reduction, we observed a Gaussian distribution in pixel intensities, prompting the adoption of a normal distribution model for local histogram normalization. This method effectively enhances contrast and feature prominence without altering the intrinsic cellular characteristics.

For the aggregate channel, only the percentile values of the first frame are loaded for global normalization, with the rest of the sequence can be processed in parallel. Similarly, for the cell channel, we load the first frame's percentile values and the parameters for the Gaussian distribution, with the rest of the sequence can be processed in parallel.

To conduct an ablation study on the impact of histogram normalization, two UNet models were trained using the same seed dataset for training and validation. When evaluated on the test set, the model utilizing histogram normalization demonstrated up to a 10% improvement in the Dice coefficient, indicating its effectiveness.

In summary, our method of delayed reading, coupled with parallel processing, has been meticulously designed to tackle the challenges presented by extensive datasets and the practical limitations of available computational resources. This detailed explanation should provide a comprehensive understanding of our data normalization methods and the technical reasoning behind our approach.

A.2 Frame registration and correction

We applied SIFT Lindeberg, 2012 algorithm to two frames affected by the shift problem. First, we identified the main points of interest and cross-referenced them with the next frame. Then, the outliers were discarded to estimate the transformation matrix, thereby canceling the shift. However, according to our performance evaluation, SIFT was sufficient to correct the shift (see Fig. **??**.a also in Table.2.3). Finally, it obtained an average error of $0.0153 \pm 0.0609px$ along the x-axis, while $0.0228 \pm 0.1221px$ along the y-axis. Thus, SIFT proved to be directionally biased by the shift. SIFT was tested using the 'SIFT_create' function from the 'OpenCV'Bradski, 2000a library to compute key points and their source and target image descriptors. The Euclidean distance (default sift error=0.7) matches points between the two key-points descriptors. The random sample consensus (RANSAC) algorithm eliminates

outliers (the 'RANSACRegressor' function from the 'scikit-learn' library). The matched points are used to find the transformation matrix.

Our recursive scheme (in our CECC module) contemplates the intricacies and challenges of this registration task. Initially, we employed a relatively large Gaussian kernel of size 513x513 ('getGuassianKernel' function from OpenCV), where $\sigma = 0.3 \cdot ((kernel\_size - 1) \cdot 0.5 - 1) + 0.8)$, which corresponds to a σ value of approximately 77.3. This choice was made to retain the essential details of the aggregates while diminishing noise. Using this, we computed an initial transformation matrix TM_0 to adjust for the offset between successive frames. Subsequently, we transitioned to a smaller Gaussian kernel, specifically 257x257, equivalent to a σ of roughly 38.9. This finer kernel resolution introduced more granularity in the aggregate details. With the previously estimated TM_0 as a starting point, we initialized a second registration process, leading to TM_1. This new transformation matrix proved to be more adept at countering the acquisition shift. We iteratively proceeded with this method, progressively reducing the kernel size during each step until we reached a point where no kernel was necessary. The final transformation matrix, TM_N, was determined using ECCM directly on the untouched frames, taking TM_{N-1} as its initialization.

To implement the CECC approach, we used the ECC implemented in the 'OpenCV' v4.5.1 Python library named 'findTransformECC'. Each cascade used a different Gaussian kernel, with 1000 max iterations or 10^{-4} error as a termination criterion for finding the corresponding transformation matrix. The last cascade computed the effective transformation matrix.

When the transformation matrix is estimated, the 'warpAffine' function (from the 'OpenCV' library) is used to register the image by the computed transformation matrix.

In order to validate our registration approach, 1000 'x' and 'y' shifts were randomly generated and then saved between -400 pixels and 400 pixels (x and y shifts are independent). For each test, we loaded the reference image (containing aggregates) and the same random shifts, in the exact same order. We created a shifted version of the reference image using 'warpAffine', with the loaded shifted 'x' and 'y'. Both images (reference and shifted) are 2048×2048 gray-scale. Each test was submitted as a job via SLURM to a computational cluster. For CECC, we used 4 cores Xeon Gold 6126 CPU, 1GB RAM and for SIFT 4 cores Xeon Gold 6126 CPU, 2GB RAM (1GB RAM for SIFT is not sufficient). The execution time is computed and reported for each registered image.

For blurry frame detection module, we computed the Laplacian of two images: *image(t)* and *image(t+k)*, where k is the step between two images (i.e., k=1 means comparing two consecutive images). Then, the module evaluate the variance of the resulting images. Images with no blur give high variance values, and images with blur give low variance values. This mechanism effectively detects sudden drops in Laplacian variance values (using the relative difference compared to a given threshold), thus detecting blurry and unusable frames (i.e. dropped from the stack).

In order to compute the Laplacian image, we used the 'Laplacian' function from the 'OpenCV' Python library in 64 float representation. Variance is then computed on the resulting image. Every two consecutive frames, the relative difference is computed, and the

blurriness is detected if:

$$|1 - \frac{\sigma^2(\nabla_5^2 f_{t+1})}{\sigma^2(\nabla_5^2 f_t)}| > \varepsilon_{blur},\ where\ \varepsilon_{blur} \in [0, 1] \tag{A.1}$$

with σ^2 the statistical variance and ∇_5^2 the five point operator Lindeberg, 1990. If the blur is detected (i.e., $\varepsilon_{blur} = 0.01$), a loop is launched to check for the disappearance of the fuzziness in the next B frames (i.e., $B = 14$).

When faced with long episodes of blur (many consecutive fuzzy frames), a bigger B value is recommended. However, one usually look for low ε_{blur} values, corresponding to a higher quality standards. This module record and save all the shift correction parameters as the rejected blurry frames.

A.3 Aggregate segmentation and quantification

. After aggregate image normalization and data check, we used a fixed 0.5 threshold to separate the aggregates from the background. Next, we labeled the segmented aggregates to extract features (i.e., count, area and centroid) using the 'label' and 'regionprops' functions from the 'scikit-image' library. To consider that a given labeled aggregate is phagocytosed by a cell, we checked every two consecutive frames if the following conditions are met: the change in the size of the labeled aggregate (decrease by half) and its centroid movement ($0.7\mu m \approx 7\ pixels$). Finally, all aggregates' features for each time point are reported/saved.

A.4 Scene instance-level cell segmentation and tracking

A.4.1 DL and IDL approaches

U-Nets used four depth levels. In the down-sampling pass, for U-Net and Attention-U-Net, each depth level had a duplication of the following sequence: 2D convolution layers (Conv2D) with 3x3 filters, 2D batch-normalization and leakyReLU, then, 2-factor max-pooling. BiO-Net used a duplication of the following sequence: Conv2D, 2D batch-normalization and ReLU. This sequence is followed by Conv2D, ReLU, 2D batch-normalization and 2-factor max-pooling. The results of each depth level are connected to the symmetrical depth of the decoder as 'skip' connections.

The midsection (bottleneck) for U-Net and Attention-U-Net was composed of a duplication of the sequence: Conv2D, 2D batch-normalization, leakyReLU. The BiO-Net bottleneck was composed of Conv2D, ReLU, 2D batch-normalization, Conv2D, ReLU, 2D batch-normalization, 2D transposed convolution, ReLU, 2D batch-normalization.

In the up-sampling pass, each depth level used up-sampling with a scale factor of two, and then the skip connection is concatenated differently for each model. In U-Net, it is directly concatenated along the first dimension with two times: Conv2D, 2D batch-normalization and leakyReLU. The Attention-U-Net passed the up-sampled signal through: Conv2D, batch-normalization, and leakyReLU. Then, the attention module (see details Oktay et al., 2018) processes the resulting signal and the skip connection. This result is concatenated with the skip connection along the first dimension and passed through the sequence:

Conv2D, 2D batch-normalization and leakyReLU. The BiO-Net used the same U-Net decoder module, by only replacing leakyReLU with ReLU. In addition, batch-normalization comes after the ReLU activation function.

For all U-Nets, the output was a single channel image (after Conv2D followed by a sigmoid function). We used the same 2D convolution layers for the encoder and decoder. DL U-Nets contains a (64, 128, 256, 512) sequence of layers for each depth level with a midsection of 1024 layers. IDL U-Nets involved (24, 48, 96, 192) series of layers for each depth level and a midsection of 384 layers. For the BiO-Net, the default 1 iteration and a multiplier of 1.0 are used.

LSTM modules (described in Fig.**??**.a) were connected to the frozen U-Nets (forward-pass only). The highest encoder depth convolution results (64x1024x1024) were concatenated with the prediction image (1x1024x1024) and passed to the $LSTM_0$ when the given frame is the first one in the chosen time-window (successions of frames), otherwise to $LSTM_i$.

The TTCM (presented in Fig.**??**.c) concatenated the probability maps from U-Nets. These results were then normalized by the number of the time points. Seeds were finally extracted using a high thresholding (i.e., 0.9), corresponding to selecting the pixels presented in most of the frames of the time-window.

The visual explanation module is connected to the XAI U-Nets. Each depth level (encoder and decoder) output was extracted before the mean activation heat map was computed along axis 1. The resulting image was scaled to match the input image dimensions (1024x1024) using the 'resize' function from the 'PIL' Python library.

A.4.2 DL training phase

Building on the established conventions for U-Nets in semantic segmentation, our model introduces a critical enhancement with the alpha factor (α_i) for calculating the global loss. This factor is dynamically computed for each training image, allowing our binary cross-entropy loss function to adapt to the unique ratio of background to foreground pixels in each image's ground truth data. For a given image i, if the cellular density is low, the alpha factor increases the weight of the cell pixel class within the loss function. Thus, the global loss for image i is defined by incorporating the alpha factor, α_i, to ensure that the loss is representative of the actual class imbalance on a per-image basis. This approach shares conceptual similarities with methodologies such as NeuRegenerate's density multiplierBoorboor et al., 2023, which adapts model behavior to address the tile-stitching artifacts. In NeuRegenerate's case, this adaptation is based on the overlap between synthetic and real inputs in a 3D volumetric context, particularly when computing the reconstruction loss within a generative adversarial network setting. Our alpha factor, however, is specifically tailored for 2D image segmentation, enhancing sensitivity to the nuances of each training image, presenting a substantial improvement in how class imbalances are addressed in the model. Where the global loss is formulated as:

$$loss_{global} = -\frac{1}{I}\sum_{i=1}^{I}[gt_i \cdot log(pred_i) + (1 - gt_i) \cdot \alpha_i \cdot log(1 - pred_i)] \tag{A.2}$$

With α_i is the ratio between the number of background and foreground pixels from the ground truth for the i-th image, I represents the number of training images, gt_i the ground truth binary mask, $pred_i$ the model prediction. In order to optimize the speed, the 'PyTorch' library is used to flatten the masks. Then the loss function is computed directly on the GPU, and reducing the delay between the forward and the backward passes.

In a first phase of the DL training, all U-Nets were trained using 5-fold cross-validation and testing while using: (i) $loss_{global}$ for retro-propagation (see equation A.2); (ii) Adam optimizer; (iii) 10^{-4} learning rate and (iv) batch size of one. After twenty epochs, the best model was saved for each validation fold based on its $loss_{global}$ score on the validation set, then tested on the test set. In order to take into account the cell border and to reduce the training time, border masks were automatically generated for the dataset in the following manner: cell binary mask was dilated using the 'binary_dilation' function from the 'scipy' library for two iterations, the pixels of the original mask was subtracted (keeping only the borders after dilation), then the border mask was dilated for 4 successive iterations (see Fig.**??**.a). We defined the border loss as:

$$loss_{border} = \frac{1}{I}\sum_{i=1}^{I}\frac{|pred_i \cap gt\_border_i|}{|gt\_border_i|} \tag{A.3}$$

Let I represent the number of training images. For each i^{th} image, gt_border_i is the set of pixels constituting the automatically generated ground-truth border, while $pred_i$ is the set of pixels where the model predicts a border. If the model's prediction, $pred_i$, does not intersect with any of the true border pixels from gt_border_i, the intersection is empty and thus $loss_{border} = 0$. Conversely, if every pixel in gt_border_i is also in $pred_i$, indicating a total overlap, then $loss_{border} = 1$.

In a second phase of the DL training, the parameters of the U-Nets were frozen, inhibiting any back-propagation. Subsequently, the U-Nets were linked to LSTM modules that functioned with a two-time point window at a time ($LSTM_0$ and $LSTM_1$), permitting back-propagation to modify only the LSTM parameters (refer to Fig.**??**.a). In our training, the loss function was a combination of equations A.2 and A.3:

$$total\_loss = \omega \cdot loss_{global} + (1-\omega)\cdot loss_{border}, \quad \omega \in [0, 0.5] \tag{A.4}$$

The coefficient ω is pivotal for controlling the weightage given to the global versus the border loss. From preliminary experimentation, $\omega = 0.4$ was discerned to be a balanced choice, thereby augmenting cell separation. For instance, a lower ω value of 0.1 improved precision but slightly detracted from recall. It is noteworthy that values of ω exceeding 0.5 jeopardized cell separation, eliciting declines in both precision and recall metrics. As a consequence, the scope of ω was confined to the interval [0, 0.5]. This value of ω not only emphasizes cell borders—a crucial factor for our cell detection quality—but also ensures the retention of important global features of the image, such as demarcating the foreground from the background.

The LSTM modules were trained using the 5-fold cross-validated U-Net frozen models, $total\_loss$ with $\omega = 0.4$ (see equation A.4), Adam optimizer, 10^{-4} learning rate and a batch

size of one. After twenty epochs, the best model was saved for each validation fold based on its *total_loss* score on the validation set and then tested on the test set.

A.4.3 IDL training phase

U-Nets were trained using 5-fold cross-validation and testing, $loss_{global}$ for retro-propagation (see equation A.2), Adam optimizer, 10^{-4} learning rate, batch size of one. After twenty epochs, the best model was saved for each validation fold based on its $loss_{global}$ score on the validation set, then tested on the test set.

A.4.4 DL/IDL inference phase

For DL we used UNets+LSTM and for IDL we used UNets+TTCM. These modules combination produced time-series-based probability maps (high-values: cells, low-values: background and cell borders). Then, cell seeds (centroid coordinates) were extracted after 0.9 thresholding. Watershed method combined the probability map as a distance map, cell centroids as seeds and the binary mask (U-Nets predictions after 0.5 thresholding) as a foreground delimiter (see Fig.**??**.b, Fig.**??**.c). Moreover, the execution time evaluation (during inference) presented in Fig.**??**.d was performed using the following hardware 8 cores i7 9700K CPU, 16GB RAM, Nvidia RTX2080 GPU and Samsung 970 PRO SSD.

A.4.5 Data input size to all models

Each frame was resized to 1024 x 1024 pixels before being input into the model. This decision was taken to strike a careful balance between maintaining high resolution for effective model performance, storage scalability and ensuring computational efficiency. The original resolution of 2048 x 2048 pixels was reduced to fit most hardware capabilities while still preserving sufficient detail for the model's tasks. Indeed, we did not utilize a tiling strategy for the frames; each was processed in its entirety at the reduced size (1024 x 1024 pixels). This approach eliminates concerns about tiling overlap and its potential implications.

A.4.6 Point2Cell annotation tool

Point2Cell integrates a pair of UNet models for distinct purposes. The first UNet model predicts binary cell masks, while the second focuses on cell density estimation. Cell density maps are generated using the Distance-map libraryXu et al., 2016, which employs a geodesic distance measure from a Python library to create 2D distributions around cell centroid coordinates, or seeds. In this system, the Distance-map library uses the Euclidean distance metric, adjusted by a linear alpha parameter, and applies it to cell centroids. These distributions are refined using binary masks from the first UNet model, yielding accurate cell density maps. Both UNet models are trained from scratch, with a batch size of one. Optimization of parameters is done using the RMSprop optimizer, at a steady learning rate of 10$^-$4. The training employs early stopping at 10 epochs and is capped at 200 epochs. Point2Cell also incorporates a user-interactive cell seeding feature for image annotation, where users manually identify each cell in an image. This feature includes options like undo, reset, and save, enhancing annotation efficiency. This manual seeding is essential for

accurate annotations. After manual seeding, Point2Cell uses its trained models to generate pseudo-cell masks and density maps. These, along with the user-provided cell seeds, are processed through a watershed algorithm Beucher and Meyer, 1993; Neubert and Protzel, 2014; M. Bai and Urtasun, 2017. This method effectively isolates and labels individual cells. In tests, Point2Cell showed greater annotation accuracy than polygon-based annotation. Using 10 images from the HeLa cells dataset from CTC, it achieved a Dice score of 94.97% in just 14.6 seconds. In comparison, the polygon-based 'labelme' tool scored 91.3% but took much longer, at 96.1 seconds. Point2Cell's efficiency is highlighted by its speed, being about six times faster than 'labelme', and its superior precision, with a 3.64% higher Dice score. Point2Cell's source code is publicly available on GitHub. It streamlines cell annotation, requiring only single-click input from the user for highly precise, pixel-level cell annotations.

A.4.7 Cell tracking

The Bayesian Tracker (btrack)Ulicna et al., 2021; Bove et al., 2017 Python library was used to track cells over time. It used the centroid and area to form cell tracks. Only the tracks with at least 100 min long were kept. Speed was computed at each time point (mean cell displacement divided by time unit), quantifying speed over time, and then, mean speed over a whole sequence was computed and reported. A similar approach was used to compute total displacement over time and for the whole sequence.

A.5 AttUNet(XAI) and UNet(XAI) as pre-trained models

The training of our models on the full dataset for 20 epochs took approximately 30 hours for the Att-Unet(XAI) and 20 hours for the Unet(XAI) using the Nvidia GPU: Tesla V100-SXM2-32GB. We consider this to be quite efficient given the high input resolution (1024 x 1024 pixels) and the complexity of the task the model is designed to perform. We trained the two models from scratch using the entire dataset so that the community will have the possibility to use it as is on similar data or fine-tune it for similar tasks. We used 3 FTD + 3 WT experiments (around 120 sequences, 22,412 images, and 22,412 masks) for training, 1 FTD + 1 WT experiment (around 40 sequences, 7,323 images, and 7,323 masks) for validation, and 1 FTD + 1 WT experiment (around 40 sequences, 6,761 images, and 6,761 masks).

A.6 Microglia primary culture

Microglia primary cultures were performed using newborn brains of controls (C57BL6/J), of FTD-mutant animals (line C9orf72-/- or GrnR493X/R493X). Newborn mice brains (less than two days old) are collected by dissection of the skull. Brains are recovered in a 50mL Falcon and mechanically dissociated by gentle pipetting into 5mL of Hank's Balanced Salt Solution (HBSS Thermo Fisher Scientific 14025050). After dissociation, the resulting cell suspension is then centrifuged at 1200rpm for 10 minutes at 4°C. The pellet is re-suspended with culture medium containing DMEM (Thermo Fisher Scientific 31885023), supplemented

with 10% de-complemented calf fetal serum free of endotoxins (HI FBS Thermo Fisher Scientific 10082147), 1% Penicillin + Streptomycin (Thermo Fisher Scientific 15070063). The cell suspension is cultured in flasks (75 mm2) previously coated with Poly-L-Lysine (SIGMA P4832) for 30 minutes at 37°C (5% CO2) then washed three times with 1X Phosphate Buffered Saline. The culture flasks are incubated at 37°C (5% CO2). Fifteen days later, microglia are ready for harvest. Microglia are obtained by light shaking and recovery of the culture medium in a 50mL Falcon. After centrifugation, cells are re-suspended in fresh culture medium and plated.

A.7 Phagocytosis assay

Aggregates of recombinant human full length TAR DNA-binding protein 43 (TDP-43, Abcam ab156345) were conjugated to Alexa Fluor 555 NHS Ester (ester succinimidyl, Thermo Fisher Scientific A20009) at equimolar concentration and deposited on a 35 mm glass-bottom dish (Ibidi, 81218-200) for 2 hours at 37°C, 5% CO2. The dish was then washed 3 times with 1X phosphate buffered saline (PBS) and 12.5 x 10^5 freshly harvested primary mouse microglia (WT, *Grn* KO or *C9orf72* KO) were seeded on top of the fluorescent aggregates in DMEM (Thermo Fisher Scientific 31885023), supplemented with 1% N2 supplement (Thermo Fisher Scientific 17502048) and 1% Penicillin + Streptomycin (Thermo Fisher Scientific 15070063). Within 30 minutes after seeding the culture dish was placed in a Zeiss Axio Observer 7 video-microscope at 37°C, 5% CO2 and video were acquired at 63X for 7h (2048 × 2048 images with $0.103\mu m$ x $0.103\mu m$ per pixel). For the sake of simplicity, we summarize the steps of data preparation as follows: (i) fluorescent aggregates were deposited onto a glass bottom culture dish and incubated for two hours; (ii) the dishes were washed three times after incubation; (iii) freshly harvested primary mouse microglia wild type (WT) or FTD-mutants were implanted on top of the fluorescent aggregates; and (iv) the culture dishes were placed in a video microscope 30 minutes following seeding, and a video was recorded accordingly.

A.8 FTD-mutants versus WT cells

The results presented in Fig.**??**.g, Fig.**??**.h, Fig.**??**.i, Fig.**??**.j and Fig.2.15 were computed in the following manner: (i) we computed the mean curves of all scenes, where we had 20 scenes maximum per acquisition, and each acquisition is (n=1) and (ii) we computed the mean values from the curves between 0 and 200 min.

A.9 Data collection

Imaging of cells/aggregates in 2D+time was performed on a Zeiss Axio Observer 7 video-microscope at 37°C, 5% CO2 and videos were acquired at 63X for 7h (2048 × 2048 images with $0.103\mu m$ x $0.103\mu m$ per pixel). We used the ZEN Microscope Software v2.6.76.0. We conducted a total of 10 experiments, with 5 using wild-type (WT) cells and 5 using FTD mutant cells. In each of these experiments, we included the cells from 6 pups. Thus, we

analyzed cells from 30 pups for WT and another 30 pups for FTD mutant cells, totaling 60 pups across all experiments. Regarding the details provided in the Methods section, the term "20 scenes per acquisition" corresponds to the capture of 20 distinct imaging sequences in each experiment. Each sequence represented 7 hours of continuous imaging using phase-contrast video microscopy, resulting in a collection of 200 unique sequences across all experiments (calculated as 10 experiments multiplied by 20 sequences per experiment). It is not a multiplicative factor of the number of pups but rather the number of sequences per experiment.

A.9.1 Laboratory animals

Mus musculus, C57BL6J, newborn mice were euthanized by decapitation as recommended for rodents up to 10 days of age. They were sacrificed to generate the microglial primary culture, parents were 4 to 8 months old. Mice were kept on a 12h light/dark cycle with food and water available ad libitum. Temperature between 19 and 24°C and humidity between 45% and 65%. To do microglial primary cultures, postnatal day one mice pups of both sexes are used and cells from all animals dissected on the same day are only pooled by genotype. As the same occurs for all genotypes it does not impair our differential analysis.

A.9.2 Compliance with essential ARRIVE guidelines

Study design:

a) Control group: Wild type (WT) (C57BL/6JRj), FTD-mutant (line C9orf72-/- or GrnR493X/R493X)

b) Experimental unit: Litter (each experiment was performed with cells extracted from one litter of pups per genotype)

Sample size:

a) Six pups per experiment, resulting in a total of 60 pups for all experiments conducted in this study.

b) Sample sizes of n=5 for WT and n=5 for FTD mutants are typical for in vivo studies.

Inclusion and exclusion criteria:

a) Experimental units with abnormally low production of microglial cells (less than 10^6 microglial cells per animal) were excluded.

b) No data had to be excluded as these samples were not used in the study.

c) The criteria have been thoroughly applied.

Randomization is not applicable in this study. For details on blinding, outcome measures, statistical methods, experimental animals, and experimental procedures, please refer to the methods section. For information on the results, please refer to the results section.

Appendix B

(Appendix B) Virtual-Staining: Validation and Reproduction Details

Submitted scientific publication and patents

Reproduction details also available in the paper:

Ounissi, M., Sarbout, I., Hugot, J. P., Martinez-Vinson, C., Berrebi, D., & Racoceanu, D. (2024). Scalable, Trustworthy Generative Model for Virtual Multi-Staining from H&E Whole Slide Images. arXiv preprint.
https://arxiv.org/abs/2407.00098.

B.1 Validation protocol for virtual staining

B.1.1 Quantitative evaluation

To address the inherent limitations of patch-level evaluation in virtual staining, such as restricted contextual information and potential inconsistencies across different tissue regions, we developed an adapted validation protocol. Traditional metrics often fail to capture the nuanced discrepancies that can occur across various regions of a tissue slide, leading to an incomplete assessment of staining quality. Our protocol, by contrast, incorporates both PSNR and SSIM to comprehensively assess the quality of WSIs. These metrics are crucial for evaluating the fidelity and structural integrity of virtually stained images. Furthermore, MSE metric is specifically employed to provide a quantitative assessment at the tissue pixel-level, significantly enhancing the precision in evaluating staining accuracy.

The use of a paired dataset, where each virtual stain is directly compared to a chemically stained ground truth counterpart (GT stain WSI), is pivotal. This pairing ensures that each evaluation metric not only measures the error or similarity in isolation but does so in a context that reflects true biological and clinical scenarios, ensuring the relevance and applicability of the findings.

The refined validation protocol involves several steps. Initially, an H&E stained WSI is processed to extract the foreground, effectively distinguishing the tissue from the background. Subsequent virtual staining algorithms synthesize the stain, producing a stain WSI that is then compared against the ground truth stain WSI obtained from chemical staining. This comparison is essential for assessing the virtual staining's performance across entire

slides and at the pixel level as illustrated in Figure.B.1. Through these metrics, our protocol addresses critical gaps in existing evaluation methods and sets a clear validation of virtual staining technologies in pathology.

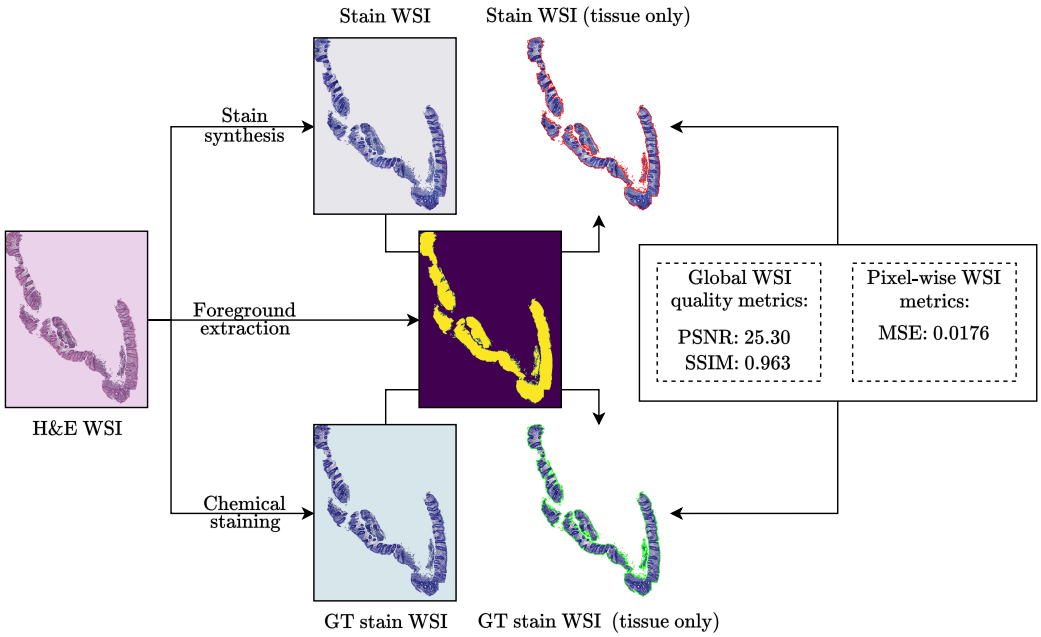


FIGURE B.1: **Evaluation protocol for virtual staining performance.** Workflow diagram illustrating the validation process for virtual staining techniques. The process begins with an H&E stained whole slide image (H&E WSI), from which the foreground is extracted. This image undergoes virtual staining to produce the Stain WSI, which is then compared to the chemically stained ground truth WSI (GT stain WSI). The evaluation metrics include PSNR and SSIM for assessing overall image quality, and MSE for pixel-wise accuracy, indicating the effectiveness of the staining simulation.

B.1.2 Qualitative evaluation

In our study, we recognize the importance of qualitative evaluation alongside quantitative metrics, particularly from a pathological perspective. Despite utilizing a paired dataset, qualitative assessment remains crucial for verifying the applicability and accuracy of our virtual staining techniques from a clinical standpoint.

In Figure.B.2, we conducted a poll involving 26 images stained with AE1AE3, where a pathologist was shown the original H&E image alongside virtual staining results. These included images processed through real chemical staining (ground truth) and those generated via our paired and unpaired DL models. Pathologist was instructed to rate the images on a scale from 1 (worst) to 5 (best) and provide feedback.

The outcomes of our study were somewhat counter-intuitive. In the assessment of 26 AE1AE3-stained images, the ground truth images, which involved actual chemical staining, generally scored lower than those from both the paired and unpaired settings. Specifically, the ground truth images received an average score of 2.69 ± 1.46. In contrast, images from the paired setting, where virtual staining was trained on paired data, scored slightly higher at 3.11 ± 1.63. Most notably, the unpaired setting, involving virtual staining trained without paired data, performed the best with an average score of 3.42 ± 1.65. This suggests an unexpected performance trend where the virtually generated stains were preferred over the

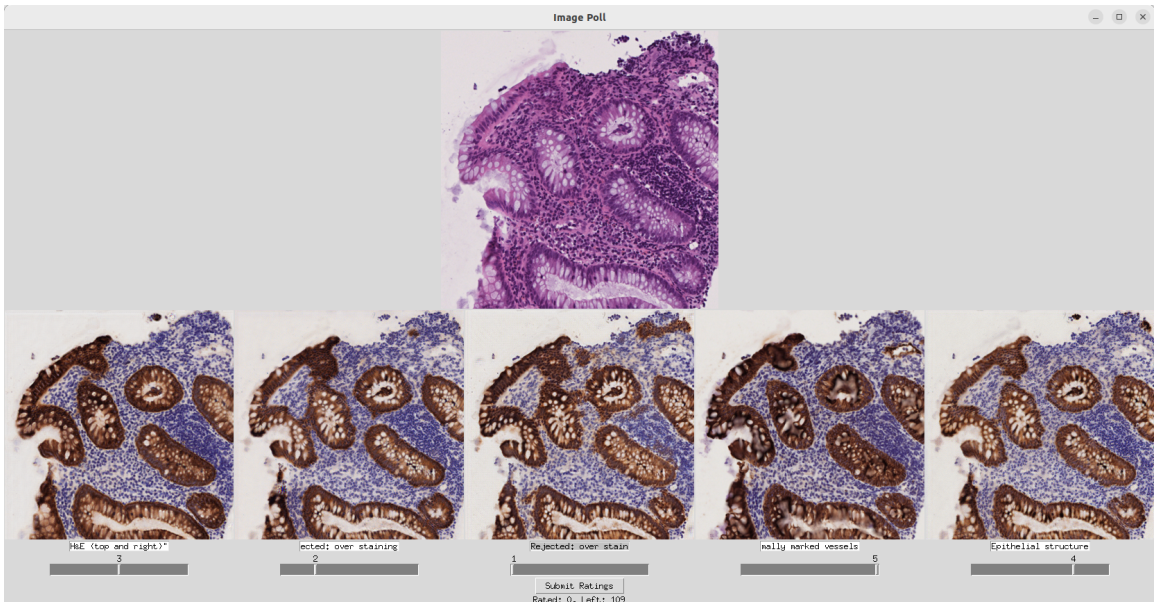


FIGURE B.2: **Software for poll results and feedback collection of pathologist ratings on staining quality.** We show the original H&E image at the top, followed by a sets of virtual stains in different conditions including the ground truth randomly showed. Pathologist was asked to rate each image based on the clarity and preservation of morphological details 1 "worst" 5 "best" with a feedback.

actual chemical stains, indicating a discrepancy in quality perception between the traditional and computational methods.

Upon analyzing the pathologists' feedback, a critical observation was made, as illustrated in Figure.B.3. It appears that a water-like blur inherent in the chemical staining process tended to obscure the morphological details of the tissue. This issue was less pronounced in the images from the paired and unpaired settings.

Notably, the unpaired model displayed superior preservation of morphological features. This is likely because, during training, the model does not directly correlate the H&E images with specific stains, allowing it to learn where to place stains effectively without replicating the blurring seen in the ground truth. Conversely, the paired model, learning from

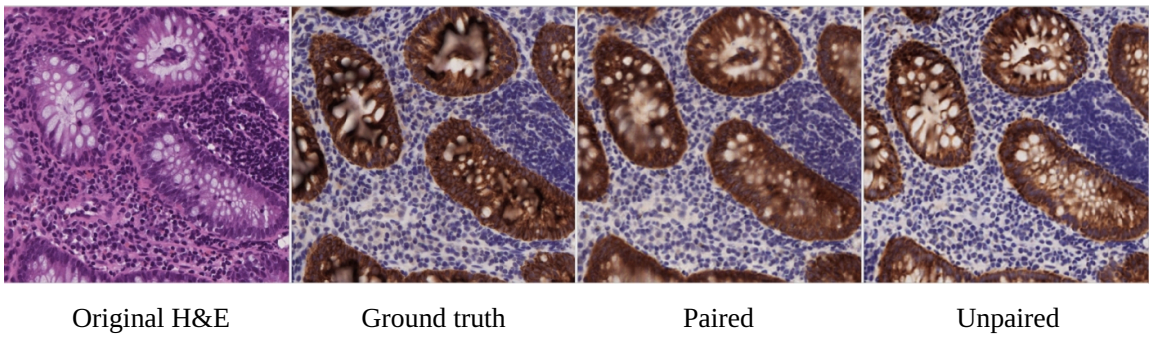


FIGURE B.3: **Morphological detail comparison in H&E stained images.** This figure shows a closer view of the morphological features in the original H&E stain (left) versus the ground truth, paired, and unpaired virtual stains. The comparison highlights the impact of water-like blur in chemical stains and its reduction in virtual stains, aiding in the qualitative assessment by pathologists.

the blurred ground truth images, tends to reproduce similar artifacts, thus inheriting and replicating these biases. These findings underscore that unpaired training can provide more generalized and unbiased results. While objective metrics might suggest lower performance compared to the paired settings, the qualitative benefits from a pathological perspective are worth noting. The unpaired setting yields visual results that surpass the ground truth (when there is a blur effect), offering enhanced clarity and detail that are crucial for accurate medical diagnosis.

B.2 Reproducibility: experimental configurations

To ensure reproducibility, it is important to note that all experiments conducted in this study utilized the same architecture for the encoder, decoder, and discriminator. The number of parameters was aligned with those specified in Anoosheh et al., 2017 and implemented using the PyTorch library (version 2.2.0 with CUDA v12.1 and cuDNN v8.902) Paszke et al., 2019. All training sessions were performed using 2048x2048 tiles resized to 512x512 tiles (no overlap) from the Crohn's dataset, as discussed in Section 3.6, in either paired or unpaired settings. The models employed an Adam optimizer with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and a batch size of 6. We used only random flip and random rotation (data augmentation strategies). Each training epoch contains 728 iterations. Each training was conducted on a single NVIDIA A100 80GB GPU. The experimental setup is outlined below.

B.2.1 Enhanced performance and efficiency in multi-virtual staining using unified H&E encoder

In Table 3.1, we trained two different approaches—our unified method and CycleGAN (refer to Figure 3.10)—. For CycleGAN, a separate model was trained for H&E to each of the different stains, with a total of eight stains. This involved 16 encoders, 16 decoders, and 16 discriminators. Each model underwent 75 epochs at a fixed learning rate of 2×10^{-4}, followed by 75 decay epochs with a linearly reducing rate, totaling 150 epochs per stain (1200 epochs overall). The loss weights were set to $\lambda_{\text{cyc}} = 10$ and $\lambda_{\text{adv}} = 1$, with no regularization as $\alpha = 0$ and $\beta = 0$.

In contrast, our approach involves simultaneous training for H&E to the eight different stains, using a total of 9 encoders, 9 decoders, and 9 discriminators. The training consists of 500 epochs at a fixed learning rate of 2×10^{-4}, followed by 500 decay epochs with a linearly reducing rate. The loss weights and regularization settings are identical to those used in the CycleGAN models.

The values presented in Table 3.1 represent the mean tile-wise (no overlap) MSE for each stain tested on the Crohn's dataset (refer to Section 3.6). These MSE values are computed for both approaches – our unified method and CycleGAN –, being reported in Table 3.1.

B.2.2 Impact of incorporating IHC loss functions and H&E regularization on stain synthesis quality

In Table 3.5, the training involved 9 encoders, 9 decoders, and 9 discriminators. The model underwent 500 epochs at a fixed learning rate of 2×10^{-4}, followed by 500 decay epochs with a linearly reducing rate, summing up to a total of 1000 epochs (paired and unpaired). The impact of incorporating different loss functions and regularization was studied, specifically:

- $\mathcal{L}_{\text{H\&E}}$ (✓): This regularization was applied at the end of each iteration, where the cycle consistency losses $\mathcal{L}_{\text{cyc},i}$ from the 8 components of the Crohn dataset were summed and averaged. The loss weights were set as $\lambda_{\text{cyc}} = 10$ and $\lambda_{\text{adv}} = 1$, with $\alpha = 0$ and $\beta = 0$.
- $\mathcal{L}_{\text{IHC}}$ (✓): For the IHC-specific loss, $\lambda_{\text{cyc}} = 10$ and $\lambda_{\text{adv}} = 1$ were maintained, and values of α and β were computed as detailed in Section 3.4.1.1.
- Combined $\mathcal{L}_{\text{H\&E}}$ (✓) and $\mathcal{L}_{\text{IHC}}$ (✓): Both H&E regularization and IHC loss were applied similarly as described above, with α and β values computed according to the method outlined in Section 3.4.1.1.

The values presented in Table 3.5 represent the MSE, PSNR abd SSIM computed at the WSI level. These metrics are calculated for WSIs reconstructed with 0% overlap and represent the mean values across all eight different stains of the Crohn dataset. Further details on the validation protocol are provided in Section B.1.

B.2.3 Comparison of our model's performance across different magnifications

In Table 3.2, we evaluated the performance under both paired and unpaired settings using specific magnifications. This involved 9 encoders, 9 decoders, and 9 discriminators. The model underwent 500 epochs at a fixed learning rate of 2×10^{-4}, followed by 500 decay epochs with a linearly reducing rate, summing up to a total of 1000 epochs (paired and unpaired). The magnifications tested were:

- x10 with an original tile size of 2048x2048 pixels, which corresponds to approximately $450.56 \times 450.56 \mu m$,
- x20 with an original tile size of 1024x1024 pixels, approximately $225.28 \times 225.28 \mu m$,
- x40 with an original tile size of 512x512 pixels, approximately $112.64 \times 112.64 \mu m$.

All images are resized to 512x512 for training, following the configuration detailed in Section B.2.2. This configuration employs combined loss functions $\mathcal{L}_{\text{H\&E}}$ and $\mathcal{L}_{\text{IHC}}$, with parameters $\lambda_{\text{cyc}} = 10$ and $\lambda_{\text{adv}} = 1$. The performance metrics, listed in Table 3.2, include MSE, PSNR, SSIM. These metrics are computed on WSIs reconstructed with 0% overlap (mean values across all eight different stains of the Crohn dataset). It is important to note that training at a magnification of x40 and testing at x10 requires resizing the synthetic x40 WSI to match the size of the x10 slide. After resizing, metrics are calculated to compare

the ground truth slide at x10 with the resized slide. This procedure is applicable to other magnifications as well. Further details on the validation protocol between two WSIs are provided in Section B.1.

B.2.4 Effects of various regularization techniques on unpaired virtual staining performance

In Table 3.6, we evaluated the performance under an unpaired setting using x10 magnification (original tile size of 2048x2048 pixels, corresponding to approximately $450.56 \times 450.56 \mu m$), resized to 512x512. The first row details our approach using the configuration described in Section B.2.1. The second row uses the same configuration, combining IHC loss functions with H&E regularization, as referenced in Section B.2.2. For subsequent rows, whenever a specific stain regularization is applied, the parameters $\lambda_{\text{cyc}} = 10$, $\lambda_{\text{adv}} = 1$, and values for α and β are computed according to the method outlined in Section 3.4.1.1. Additionally, $\mathcal{L}_{\text{idt}} = 1$, $\mathcal{L}_{\text{lat}} = 1$, or $\mathcal{L}_{\text{fwd}} = 1$ may be applied, with detailed descriptions of each stain regularization found in Section 3.4.1.2.

B.2.5 Hamming window-based approach for clean tile-stitching

To address the inevitable stitching artifacts encountered during the reconstruction of synthetic WSIs, we applied a tailored image processing approach. Central to our methodology was the use of a two-dimensional (2D) Hamming window Hamming, 1998; Oppenheim and Schafer, 1999, designed to smooth the transitions between adjacent image patches and mitigate edge effects.

The Hamming window, traditionally used in signal processing Hamming, 1998; Oppenheim and Schafer, 1999 to taper the signal edges, was adapted to two dimensions to suit the image patches. Each patch, representing a portion of the larger image, was processed through this window to ensure a gradual transition at the borders. With an overlap > 0, this was achieved by computing the outer product of a one-dimensional Hamming window with itself, thus creating a symmetrical 2D window $w(x, y)$ for a patch of size $M \times M$ is defined as:

$$w(x, y) = 0.54 - 0.46 \cos\left(\frac{2\pi x}{M-1}\right) \cdot \left(0.54 - 0.46 \cos\left(\frac{2\pi y}{M-1}\right)\right) \tag{B.1}$$

where x, y range from 0 to $M-1$. This results in a 2D Hamming window which reduces the pixel values towards the edges of each patch. This window was then applied across the three color channels of the image. Each image patch (across all RGB cahnnels) was element-wise multiplied by this matrix, reducing the intensity at the peripheries and thereby softening the boundaries between stitched patches. This operation is described by the following equation:

$$P_{\text{weighted}}(x, y) = P(x, y) \cdot w(x, y) \tag{B.2}$$

Where $P(x, y)$ is the original pixel value at coordinates (x, y) within the patch for a given color channel, and $w(x, y)$ is the value from the 2D Hamming window at these coordinates.

Post application of the Hamming window, the weighted patches were summed to form the complete WSI. In regions where patches overlapped, pixel values from multiple patches were combined. To ensure uniformity, the accumulated weights of the patches were recorded and used to normalize the pixel values in these overlapping areas. This normalization process was crucial for maintaining consistent intensity across the WSI, preventing visual discontinuities that could hinder the quality of the synthetically stained WSIs. This methodology can be applied to any tile-based virtual staining approach to reconstruct a clean WSI output.

The final processed image was saved in pyramidal TIFF format, suitable for high-quality WSI. The processing pipeline was implemented using Python, utilizing libraries such as NumPy (Harris et al., 2020) v1.26.3 for numerical operations and PyVIPS (*PyVips Library* 2024) v2.2.2 for image handling, ensuring efficient memory usage and scalability.

B.3 Multi-Virtual Staining Production Phase

1. User Input: Prompt the user to select desired IHC stains from the available options.
   (a) If the user provides an invalid selection, display an error message and prompt again.
2. Data Storage: Store the user's selections.
   (a) Ensure that the storage mechanism confirms successful storage. If storage fails, try again or provide an error notification.
3. Image Load: Load the H&E whole slide digital image labeled "K" and the H&E discriminator.
   (a) If the image or discriminator fails to load, log the error, notify the user, and prompt for a different slide or retry.
4. Data Quality Check for the H&E "K" whole slide:
   (a) Implement validation checks to ensure that calculated reliability scores are within expected ranges. If not, log discrepancies.
   (b) Predefined Threshold: If more than 5% of the image scores below a trust level of 0.9, the stain may not meet the reliability standard.
   (c) Global Confidence Level: A heatmap may be produced if the average or median reliability score of the entire image falls below 0.85.
   (d) Local Variance: Areas of the image with significant score variations may suggest model inconsistencies and should be flagged.
   (e) ROI relevance and robustness: Regions like those near tumors, which are critical, may have a stringent reliability threshold set, such as 95%.
   (f) Past Error Rate: If similar tissue regions have had unreliable predictions in the past, current predictions with analogous scores are treated as unreliable.
5. Check Results:

   (a) If the check is successful, release memory by deleting the H&E discriminator.
   (b) If not:
      i. Determine if the H&E "K" slide is normalized. If yes, present a heatmap/XAI to the user. Show the user where the problem is originating from spatially in the image using a color code (i.e., red=problem, green=OK). Give the user the option to ignore the issue and move to step 6, end the process, move to step 1 with H&E whole slide digital slide "K1".
      ii. If not, proceed with normalization, ensuring that any issues during this process are caught and handled. Then move to step 4.

6. Data Transformation: Convert the H&E image into an embedding.
   (a) Confirm successful transformation. If there's an issue, log it and notify the user.
7. Embedding Storage: Save the H&E embedding.
   (a) Ensure successful storage. If storage fails, try again or provide an error notification.
8. Memory Management: Delete the H&E encoder to free up memory.
   (a) Confirm successful deletion. If an issue arises, log it.
9. Decoder Load: Using the user's selection, load the appropriate IHC stain decoders.
   (a) If any decoder fails to load, log the error, notify the user, and attempt to reload.
10. Stain Generation: Produce the IHC stains from the saved embedding.
   (a) Monitor for any anomalies or errors. If encountered, log them and notify the user.
11. Decoder Deletion: Delete all activated IHC stain decoders to free up memory.
   (a) Confirm successful deletion and check memory status.
12. Discriminator Load: Load discriminators for the chosen IHC stains.
   (a) If a discriminator fails to load, log the error, notify the user, and attempt to reload.
13. Quality Check for Stains: Perform a data quality check for each stain.
   (a) Ensure that each check completes successfully. If an issue is detected, log it.
14. Model Deletion: Delete all active models to free up memory.
   (a) Confirm successful deletion and check memory status.
15. User Feedback: Display the original H&E image, the selected IHC stains, and their XAI visual heatmaps to the user.

(a) Ensure that all displays load properly. If there's a display issue, log it and attempt to reload the visuals.

16. End or Restart: Offer the user the choice to end the program or restart with the H&E whole slide digital image "K1".

    (a) If the user chooses to restart and there's an issue loading "K1", notify the user and provide options.